

Karl Aberer Andreas Flache
Wander Jager Ling Liu Jie Tang
Christophe Guéret (Eds.)

LNCS 7710

Social Informatics

4th International Conference, SocInfo 2012
Lausanne, Switzerland, December 2012
Proceedings

Soc
Info 2012

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Karl Aberer Andreas Flache
Wander Jager Ling Liu Jie Tang
Christophe Guéret (Eds.)

Social Informatics

4th International Conference, SocInfo 2012
Lausanne, Switzerland, December 5-7, 2012
Proceedings



Springer

Volume Editors

Karl Aberer

Distributed Information Systems Laboratory, Lausanne, Switzerland

E-mail: karl.aberer@epfl.ch

Andreas Flache

University of Groningen, Department of Sociology, Groningen, The Netherlands

E-mail: a.flache@rug.nl

Wander Jager

Groningen Center for Social Complexity Studies, Groningen, The Netherlands

E-mail: w.jager@rug.nl

Ling Liu

Georgia Institute of Technology, Atlanta, GA, USA

E-mail: lingliu@cc.gatech.edu

Jie Tang

Tsinghua University, Beijing, China

E-mail: jietang@tsinghua.edu.cn

Christophe Guéret

Data Archiving and Networked Services, Den Haag, The Netherlands

E-mail: christophe.gueret@dans.knaw.nl

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-35385-7

e-ISBN 978-3-642-35386-4

DOI 10.1007/978-3-642-35386-4

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: Applied for

CR Subject Classification (1998): C.2, H.5, H.4, H.3, I.2.6, J.4

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

© Springer-Verlag Berlin Heidelberg 2012

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

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

This volume contains the papers presented at SocInfo 2012, the Fourth International Conference on Social Informatics, held on December 5–7, 2012 in Lausanne.

SocInfo 2012 provided an interdisciplinary venue for researchers from computer science, informatics, social sciences and management sciences to exchange ideas, opinions and original research work. After a year of hard work on the part of authors, reviewers and conference organizers, we were delighted to share with you a strong technical program at the conference.

There were 69 submissions to the research track of SocInfo 2012, of which 21 full length presentation papers and 18 short presentation papers were accepted. Each submission was reviewed by at least 1, and on average 2.9, program committee members. The acceptance decision was carefully made based on both reviews and online PC discussions. Our sincere thanks to all our colleagues who volunteered to serve as program committee members and reviewers on this year’s program committee. We want to especially acknowledge the hard work of the PC Co-chair, Jie Tang, who took on the relentless task of driving the SocInfo 2012 review process to a conclusion.

We are also extremely grateful to General Co-chairs Karl Aberer and Andreas Flache for their leadership, and to Surender Yerva from EPFL, who provided outstanding assistance and help to the program co-chairs with respect to the EasyChair system.

We trust that you enjoyed this year’s technical program and the unique opportunity to exchange ideas and research results with researchers in computer science, informatics, social sciences and management sciences.

October 2012

Wander Jager
Ling Liu
Jie Tang
Andreas Flache
Karl Aberer
Christophe Guéret

Organization

Program Committee

| | |
|----------------------------|--|
| Fred Amblard | CNRS IRIT - Université des Sciences Sociales Toulouse 1, France |
| Stuart Anderson | University of Edinburgh, UK |
| Emma Angus | University of Wolverhampton, UK |
| Sitaram Asur | HP Labs Palo Alto, USA |
| George Barnett | University of California, USA |
| Michael Baron | Baron Consulting, Australia |
| Nadia Bennani | Lab. LIRIS INSA de Lyon, France |
| Abraham Bernstein | University of Zurich, Switzerland |
| Jan Blom | Nokia Research, Switzerland |
| Francesco Bolici | Cassino University, Italy |
| Ulrik Brandes | University of Konstanz, Germany |
| Sonja Buchegger | KTH, Sweden |
| Klemens Böhm | Universität Karlsruhe (TH), Germany |
| Amit Chopra | University of Trento, Italy |
| Chris Cornelis | University of Granada, Spain |
| Anwitaman Datta | NTU Singapore, Singapore |
| Martine De Cock | Ghent University, Belgium |
| Yves-Alexandre de Montjoye | Massachusetts Institute of Technology, USA |
| Guillaume Deffuant | Cemagref, France |
| Chris Dellarocas | Boston University, USA |
| Jana Diesner | University of Illinois at Urbana-Champaign, USA |
| Marios Dikaiakos | University of Cyprus, Cyprus |
| Ying Ding | Indiana University, USA |
| Xiaoyong Du | Renmin University of China, China |
| Bruce Edmonds | Manchester Metropolitan University Business School, UK |
| Andreas Ernst | University of Kassel, Germany |
| Alois Ferscha | University of Linz, Austria |
| Richard Forno | UMBC, USA |
| Heljä Franssila | University of Tampere, Finland |
| Timothy French | Bedfordshire University, UK |
| Wai-Tat Fu | University of Illinois, USA |
| Gerhard Fuchs | University of Stuttgart, Germany |
| Armando Geller | Scensei, USA |
| Peter Groenewegen | VU University Amsterdam, The Netherlands |

| | |
|---------------------|--|
| Ido Guy | IBM Research, Israel |
| Chris Hinnant | U.S. Government Accountability Office, USA |
| Wenlian Hsu | Academia Sinica, China |
| Yuh-Jong Hu | National Chengchi University, China |
| Baden Hughes | The University of Melbourne, Australia |
| Stephan Humer | Berlin University of the Arts, Germany |
| Marco Janssen | Arizona State University, USA |
| Mark Jelasity | University of Szeged, Hungary |
| James Joshi | University of Pittsburgh, USA |
| Radu Jurca | Google Inc., Switzerland |
| Jean-Daniel Kant | Pierre and Marie Curie - Paris VI, France |
| Przemysław Kazienko | Wroclaw University of Technology, Poland |
| Andreas Koch | University of Salzburg, Austria |
| Walter Lamendola | University of Denver, USA |
| Georgios Lappas | Technological Educational Institute (T.E.I.) of Western Macedonia, Greece |
| Bangyong Liang | NEC Laboratories, China |
| Ee-Peng Lim | Singapore Management University, Singapore |
| Yefeng Liu | Waseda University, Japan |
| Paul Lukowicz | University of Passau, Germany |
| Christopher Mascaro | Drexel University, USA |
| Winter Mason | Stevens Institute of Technology, USA |
| Paolo Massa | Bruno Kessler Foundation, Italy |
| David Millard | University of Southampton, UK |
| Bamshad Mobasher | DePaul University, USA |
| Tony Moore | Deloitte Development LLC, USA |
| Mikołaj Morzy | Poznan University of Technology, Poland |
| Tsuyoshi Murata | Tokyo Institute of Technology, Japan |
| Keiichi Nakata | University of Reading, UK |
| Taewoo Nam | University at Albany, State University of New York, USA |
| Wolfgang Nejdl | L3S and University of Hanover, Germany |
| Bruce Neubauer | Albany State University, USA |
| See-Kiong Ng | Institute for Infocomm Research, Singapore |
| Carlos Nunes Silva | University of Lisbon, Portugal |
| Huseyin Oktay | University of Massachusetts, Amherst, USA |
| Anne-Marie Oostveen | Oxford Internet Institute, University of Oxford, UK |
| Mario Paolucci | Institute of Cognitive Sciences and Technologies, Italy |
| Elin Rønby Pedersen | Google, San Francisco, USA |
| Gregor Petrič | University of Ljubljana, Slovenia |
| Daniele Quercia | Universtiy College London, UK |
| Alice Robbin | Indiana University, USA |

| | |
|--------------------|--|
| Richard Rogers | University of Amsterdam, The Netherlands |
| Sini Ruohomaa | University of Helsinki, Finland |
| Geoffery Seaver | National Defense University, USA |
| Kalpana Shankar | Indiana University-Bloomington, USA |
| Xiaolin Shi | Microsoft Corporation, USA |
| Vaclav Snasel | VSB-Technical University of Ostrava, Czech Republic |
| Thanakorn Sornkaew | Ramkhamheang University, Thailand |
| Flaminio Squazzoni | Università di Brescia, Italy |
| Thorsten Strufe | TU Darmstadt, Germany |
| Aixin Sun | Nanyang Technological University, Singapore |
| Neel Sundaresan | eBay Research Labs, USA |
| Maurizio Teli | Fondazione Ahref, Italy |
| Klaus G. Troitzsch | University of Koblenz-Landau, Germany |
| Julita Vassileva | University of Saskatchewan, Canada |
| Miguel Vicente | Universidad de Valladolid, Spain |
| Mark Weal | University of Southampton, UK |
| Roger Whitaker | Cardiff University, UK |
| Nanda Wijermans | University of Utrecht, The Netherlands |
| Jing Zhang | University of Minnesota, USA |
| Weining Zhang | University of Texas at San Antonio, USA |
| Honglei Zhuang | Tsinghua University, China |
| Thomas Ågotnes | University of Bergen, Norway |

Additional Reviewers

| | |
|----------------------|---------------------------|
| Bodriagov, Oleksandr | McGee, Jeff |
| Hu, Meiqun | Niu, Wei |
| Kim, Jinseok | Pan, Sinno J. |
| Li, Chenliang | Rietveld, Laurens |
| Liang, Yuan | Rodriguez Cano, Guillermo |
| McDonald, Nora | |

Table of Contents

| | |
|--|-----|
| A System for Web Widget Discovery Using Semantic Distance between User Intent and Social Tags | 1 |
| <i>Zhenzhen Zhao, Xiaodi Huang, and Noël Crespi</i> | |
| An Automated Multiscale Map of Conversations: Mothers and Matters | 15 |
| <i>Ansuya Ahluwalia, Allen Huang, Roja Bandari, and Vwani Roychowdhury</i> | |
| How Influential Are You: Detecting Influential Bloggers in a Blogging Community | 29 |
| <i>Imrul Kayes, Xiaoning Qian, John Skvoretz, and Adriana Iamnitchi</i> | |
| A Simulation Model Using Transaction Cost Economics to Analyze the Impact of Social Media on Online Shopping | 43 |
| <i>Apratim Mukherjee, Shrabastee Banerjee, and Somprakash Bandyopadhyay</i> | |
| Predicting Group Evolution in the Social Network | 54 |
| <i>Piotr Bródka, Przemysław Kazienko, and Bartosz Kołoszczyk</i> | |
| Interpolating between Random Walks and Shortest Paths: A Path Functional Approach | 68 |
| <i>François Bavaud and Guillaume Guex</i> | |
| Dynamic Targeting in an Online Social Medium | 82 |
| <i>Peter Laflin, Alexander V. Mantzaris, Fiona Ainley, Amanda Otley, Peter Grindrod, and Desmond J. Higham</i> | |
| Connecting with Active People Matters: The Influence of an Online Community on Physical Activity Behavior | 96 |
| <i>Maartje Groenewegen, Dimo Stoyanov, Dirk Deichmann, and Aart van Halteren</i> | |
| Detecting Overlapping Communities in Location-Based Social Networks | 110 |
| <i>Zhu Wang, Daqing Zhang, Dingqi Yang, Zhiyong Yu, and Xingshe Zhou</i> | |
| <i>CrowdLang</i> : A Programming Language for the Systematic Exploration of Human Computation Systems | 124 |
| <i>Patrick Minder and Abraham Bernstein</i> | |

| | |
|--|-----|
| Experiments in Cross-Lingual Sentiment Analysis in Discussion Forums | 138 |
| <i>Hatem Ghorbel</i> | |
| Quality Assessment of User Comments on Mobile Platforms Considering Channel of Activation and Platform Design | 152 |
| <i>Christopher Fröch and Martin Schumann</i> | |
| A Method Based on Congestion Game Theory for Determining Electoral Tendencies | 162 |
| <i>Guillermo De Ita, Luis Altamirano, Aurelio López-López, and Yolanda Moyao</i> | |
| A Model to Represent Human Social Relationships in Social Network Graphs | 174 |
| <i>Marco Conti, Andrea Passarella, and Fabio Pezzoni</i> | |
| C4PS - Helping Facebookers Manage Their Privacy Settings | 188 |
| <i>Thomas Paul, Martin Stopczynski, Daniel Puscher, Melanie Volkamer, and Thorsten Strufe</i> | |
| Dynamic “Participative Rules” in Serious Games, New Ways for Evaluation? | 202 |
| <i>Jean-Pierre Cahier, Nour El Mawas, and Aurélien Béné</i> | |
| Mobile Phones, Family and Personal Relationships: The Case of Indonesian Micro-entrepreneurs | 216 |
| <i>Misita Anwar and Graeme Johanson</i> | |
| An Analysis of Topical Proximity in the Twitter Social Graph | 232 |
| <i>Markus Schaal, John O’Donovan, and Barry Smyth</i> | |
| A Foresight Support System to Manage Knowledge on Information Society Evolution | 246 |
| <i>Andrzej M.J. Skulimowski</i> | |
| How Many Answers Are Enough? Optimal Number of Answers for Q&A Sites | 260 |
| <i>Pnina Fichman</i> | |
| Analysis and Support of Lifestyle via Emotions Using Social Media | 275 |
| <i>Ward van Breda, Jan Treur, and Arlette van Wissen</i> | |
| A Computational Analysis of Joint Decision Making Processes | 292 |
| <i>Rob Duell and Jan Treur</i> | |
| Collaboratively Constructing a VDL-Based Icon System for Knowledge Tagging | 309 |
| <i>Xiaoyue Ma and Jean-Pierre Cahier</i> | |

| | |
|---|-----|
| A Multi-dimensional and Event-Based Model for Trust Computation in the Social Web | 323 |
| <i>Barbara Carminati, Elena Ferrari, and Marco Viviani</i> | |
| On Recommending Hashtags in Twitter Networks | 337 |
| <i>Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu</i> | |
| A Framework for the Design and Synthesis of Coordinated Social Systems | 351 |
| <i>Wynn Stirling, Christophe Giraud-Carrier, and Teppo Felin</i> | |
| Swayed by Friends or by the Crowd? | 365 |
| <i>Zeinab Abbassi, Christina Aperjis, and Bernardo A. Huberman</i> | |
| Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 U.S. Republican Presidential Primaries | 379 |
| <i>Lu Chen, Wenbo Wang, and Amit P. Sheth</i> | |
| Web Page Recommendation Based on Semantic Web Usage Mining | 393 |
| <i>Soheila Abrishami, Mahmoud Naghibzadeh, and Mehrdad Jalali</i> | |
| Scalable Analysis for Large Social Networks: The Data-Aware Mean-Field Approach | 406 |
| <i>Julie M. Birkholz, Rena Bakhshi, Ravindra Harige, Maarten van Steen, and Peter Groenewegen</i> | |
| A Survey of Recommender Systems in Twitter | 420 |
| <i>Su Mon Kywe, Ee-Peng Lim, and Feida Zhu</i> | |
| A Multi-view Content-Based User Recommendation Scheme for Following Users in Twitter | 434 |
| <i>Milen Chechev and Petko Georgiev</i> | |
| Spam Fighting in Social Tagging Systems | 448 |
| <i>Sasan Yazdani, Ivan Ivanov, Morteza AnaLoui, Reza Berangi, and Touradj Ebrahimi</i> | |
| The Multidimensional Study of Viral Campaigns as Branching Processes | 462 |
| <i>Jarosław Jankowski, Radosław Michalski, and Przemysław Kazienko</i> | |
| Models of Social Groups in Blogosphere Based on Information about Comment Addressees and Sentiments | 475 |
| <i>Bogdan Gliwa, Jarosław Koźlak, Anna Zygmunt, and Krzysztof Cetnarowicz</i> | |
| Dark Retweets: Investigating Non-conventional Retweeting Patterns | 489 |
| <i>Norhidayah Azman, David E. Millard, and Mark J. Weal</i> | |

| | |
|---|------------|
| Studying Paths of Participation in Viral Diffusion Process | 503 |
| <i>Jarostaw Jankowski, Sylwia Ciuberek, Anita Zbieg, and Radoslaw Michalski</i> | |
| Paradox of Proximity – Trust and Provenance within the Context of Social Networks and Policy | 517 |
| <i>Somya Joshi, Timo Wandhöfer, Vasilis Koulolias, Catherine Van Eeckhaute, Beccy Allen, and Steve Taylor</i> | |
| Namelings: Discover Given Name Relatedness Based on Data from the SocialWeb | 531 |
| <i>Folke Mitzlaff and Gerd Stumme</i> | |
| SocialTrends: A Web Application for Monitoring and Visualizing Users in Social Media | 535 |
| <i>Maurizio Tesconi, Davide Gazzé, and Angelica Lo Duca</i> | |
| Demonstration of Dynamic Targeting in an Online Social Medium | 539 |
| <i>Peter Laflin, Fiona Ainley, Amanda Otley, Alexander V. Mantzaris, and Desmond J. Higham</i> | |
| Navigating between Chaos and Bureaucracy: Backgrounding Trust in Open-Content Communities | 543 |
| <i>Paul B. de Laat</i> | |
| Author Index | 559 |

A System for Web Widget Discovery Using Semantic Distance between User Intent and Social Tags

Zhenzhen Zhao¹, Xiaodi Huang², and Noël Crespi³

¹ Institut Mines-Télécom, Télécom SudParis
91000 Evry, France
Zhenzhen.zhao@it-sudparis.eu

² Charles Sturt University
Albury, NSW 2640, Australia
xhuang@csu.edu.au

³ Institut Mines-Télécom, Télécom SudParis
91000 Evry, France
noel.crespi@institut-telecom.fr

Abstract. Social interaction leverages collective intelligence through user-generated content, social networking, and social annotation. Users are enabled to enrich knowledge representation by rating, commenting, and tagging. The existing systems for service discovery make use of semantic relation among social tags, but ignore the relation between a user information need for services and tags. This paper first provides an overview of how social tagging is applied to discover contents/services. An enhanced web widget discovery model that aims to discover services mostly relevant to users is then proposed. The model includes an algorithm that quantifies the accurate relation between user intent for a service and the tags of a widget, as well as three different widget discovery schemes. Using the online service of Widgetbox.com, we experimentally demonstrate the accuracy and efficiency of our system.

Keywords: content discovery, folksonomy, service discovery, social tagging, algorithm, widget.

1 Introduction

Social web extends the concept of collective intelligence. Such intelligence is hidden in the Web 2.0. The intelligence is distributed over user activities, such as user-generated contents in YouTube, Flickr, Wikipedia, and Blogs for socializing and knowledge sharing; user-enhanced social relationships through social networking such as Facebook, Myspace, and LinkedIn; and user-enriched knowledge representation through social annotation like social tagging, rating, and commenting.

Nowadays, a huge number of web services keep appearing. This makes it more and more difficult to discover services and resources. Traditional methods for web discovery use the WSDL and UDDI [1]. However, this technique has difficulties in

achieving the precision rate of searching. For improving the accurate rate, the semantic search was introduced, which uses the similarity and relations of queries and resources. Furthermore, advanced languages such as OWL-S [2], WSDL-S [3], and WSMO [4], have been developed. The semantic web ontology has two contradicting features. First, current ontology language models perform well for particular service models in particular situations. But the number of the services based on the semantic web and the ontology language are limited. Second, semantic web ontologies are consistent, but also relatively static and inflexible [5]. Their consistence is because they are often created by a small number of experts.

Social tagging can be seen as a complementing approach to ontology building, termed as Folksonomy [6]. Compared with the traditional meta-data organization, folksonomy enriches meta-data resources collaboratively by all web users in lowering barriers to cooperation [7].

This paper attempts to answer the question as to how to support the discovery of web content/service using social intelligence. In particular, how social tagging is used in discovering services in the Internet? By investigating the current web, the service discovery is based mainly on the keyword-matching algorithms, which accept users' input keywords to look for elements that would contain information of the input words. Social tagging can help in improving the accuracy of retrieved results. How to relate hidden, implicit tag information to user intention becomes the key issue.

In particular, we quantify the semantic relation between an input keyword that indicates the user intent on services and tags that are associated with services in a multifaceted way. Social tagging has its own problems as uncontrolled vocabulary and non-hierarchical structure [7]. Previous research has addressed some issues on the ambiguity of tags. However, the ambiguities of user information need and how to build a relation between the ambiguous user intent and ambiguous tags have been ignored. A user who is looking for a service issues a query of a keyword to search for a service, for example. In some cases, the user cannot describe exactly what services she wants due to her ambiguous information need. On the other hand, the keyword cannot accurately describe the service she wants due to the ambiguous meaning of the keyword. In order to accurately retrieve services, we make use of an n-m multiple relations among an input keyword, its synonyms, and the tags, instead of a 1-m relation between the keyword and tags.

There are abundant researches on web content and service discovery using folksonomy. This paper reviews the state of the art of research work in social tagging. Two comparison tables are presented to compare the different approaches of tag relationship discovery and content/service discovery.

As the important contribution of this paper, an enhanced mathematical model of web widget discovery is proposed, together with an implemented system. In our model, the relation between user intent and tags is measured, and such a relation is then used to discover widgets. To evaluate its performance, we implement the model in our

system. The results demonstrate the efficiency of our proposed model by comparing to the current algorithm used in an online service of Widgetbox.com.

The remainder of the paper is organized as follows. Section 2 reviews the literatures on content/service discovery through social tagging. Section 3 presents the proposed enhanced widget discovery model. System design and prototype are described in Section 4 and the relevant results are discussed in Section 5. Finally, we conclude this paper with the future work in Section 6.

2 Related Work

In this section we present the relevant approaches and systems on discoveries of tag relations, and services.

2.1 Tag Relationship Discovery

Many web services such as Del.icio.us and Flickr.com allow users to tag their desire keywords to an element in the web site. As the service grows bigger, the number of users increases and the number of tags in the system also increases. This raises a question as to how tags are related to each other. The relations maybe exist in terms of synonyms (Chukmol et al. [8]), or through the resource they are notated with (Wu et al. [11] & Dubinko et al. [10]), or even through a word ontology (Li et al. [12], Zhou et al. [7]).

Many researchers have investigated and attempted to implement a number of methods for discovering tag relations. Most of the studies tend to use the information from the existing services like Flickr (Dubinko et al. [10]), and Del.icio.us (Zhou et al. [7]). This could be because implementing existing information is better than creating new one, and using tag relations is more efficient using the large scale of information data (Li et al. [12]). Many researches have provided a great perspective on revealing the possibility of discovering tag relations using different kinds of algorithms, models, and methods. This is done from many different fields of studies such as the semantic network, and information retrieval.

Other research can be classified in terms of ideas, different types of implementation methods, and how each of them looks at the problem differently. For example, some papers present tag ontology (Li et al. [12]), others focus on tag clustering (Wu et al. [11]), and the rests consider both (Zhou et al. [7]). In addition, several works are concerned about the evolution of tag relations over a time window (Dubinko et al. [10]).

Our work here is different from existing works in that we examine the relation between user information intent and tags. We argue that user intent [20] should be accurately described in the first instance, and then we are able to retrieve the services that mostly satisfy user requirements.

Table 1. Comparison on tag relationship discovery

| Author(s) | Paper | Goal | Methods |
|----------------|---|--|--|
| Li et al. | Towards Effective Browsing Large Scale Social Annotation | <ul style="list-style-type: none"> - Tag semantic - Hierarchy creation | <ul style="list-style-type: none"> - Tag concept similarity using the term frequency and inverse document frequency in Information Retrieval - Find father-tag by the coverage rule, and sub-tag by intersect rule |
| Zhou et al. | An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotation | <ul style="list-style-type: none"> - Hierarchical cluster | <ul style="list-style-type: none"> - KL-Divergence for finding tag relationship creating cluster of tags - DA Algorithm to create the hierarchical structure |
| Dubinko et al. | Visualizing Tags over Time | <ul style="list-style-type: none"> - Tag relation evolution | <ul style="list-style-type: none"> - Finding the tag relation, using term frequency and inverse document frequency, in accordance to defined time frame |

2.2 Content/Service Discovery

There are few research carried on in discovering resources using social tagging. Aurnhammer et al. [9] use users' resource preferences to recommend more resources, while Chukmol et al. [8] implement a web service, WordNet, to find word synonym and resource containing the synonyms of tags. In their paper, Ding et al., 2010 [13], introduce their own technique of QEBT and QPBT for service discovery.

Table 2. Comparison on content/service discovery

| Author(s) | Paper | Goal | Methods |
|-------------------|---|---|--|
| Aurnhammer et al. | Augmenting Navigation for Collaborative Tagging with Emergent Semantics | <ul style="list-style-type: none"> - Navigation Map - Combine Image Properties and User's Queries | <ul style="list-style-type: none"> - Oriented Gaussian Derivative and Euclidean Distance for image distances - Uses nearest neighbour classifier to find the nearest related image |
| Chukmol et al. | Enhancing Web Service Discovery by using Collaborative Tagging System | <ul style="list-style-type: none"> - Service discovery through notated tags | <ul style="list-style-type: none"> - Word's synonym comparing using WordNet finding word's synonym |
| Ding et al. | A Web Service Discovery Method Based on Tag | <ul style="list-style-type: none"> - Discovering service using user's query | <ul style="list-style-type: none"> - QEBT and QPBT algorithms |

2.3 Systems

Several researchers have built up systems to investigate how social tagging can improve the performance of the systems. Bouillet et al. [14] develop a system on automated web service composition using social tagging. They later expand this method to

facilitate the design and development of composable services. They also propose a novel approach for service design and composition by meeting faceted, tag-based functional requirements provided by end-users. Using examples from a case study in the financial services domain, they demonstrate the performance of their approach for services that can be composed into myriad workflows based on end-user goals. The authors [5] use tag-based descriptions to describe individual services.

Liu et al. [15] conduct research on automated service composition. The authors introduce a user-oriented approach, which aims to simplify service composition. They leverage the plentiful information residing in service tags, from both service descriptions (such as WSDL) and the annotations tagged by users. Based on Web browsers, they develop a user-friendly prototype so that the users are enabled to accomplish service composition in an interactive way. Later in their work [16], they propose an approach to composing data driven mashups, based on tag-based semantics. Mashup developers including end-users can easily search for desired services with tags, and combine several services by means of data flows. Being equipped with the graphical composition user interfaces in their system, developers are allowed to iteratively modify, adjust, and refine their mashups.

Gomadani et al. [17] presents a faceted approach that searches and ranks Web APIs by taking into consideration the attributes or facets of APIs found in their HTML descriptions. In their paper, the concept of “Facet tag vector” is introduced to define the union of tags that have been assigned to the APIs by users, according to the categories grouped under the facets. The authors evaluate classification, search accuracy, and ranking effectiveness using available APIs. In order to provide more meaningful search results to users, Arabshian [18] presents a framework that performs context-aware search for tagged data by using a tag ontology that includes context information, as well as tagged keywords.

To our best knowledge, no system in discovering widgets through social tagging, however, has been reported. Our system is the first attempt in discovering widgets by using social tagging.

3 Our Algorithm and Model for Web Widget Discovery

In this section, we define the methodology that is used to implement in our experiment in the Widget domain. A widget is a light-weight application or a component of an interface, which enables a user to perform a function or access a service. Widget-Box.com, a widget provider, which allows users to share, tag, and rate their created or preferred widgets.

3.1 Tag Discovery by Measuring Semantic Distances

The user information need for wedges is called an event in this paper, which is characterized by a user input keyword. Normally, the user intention for the requirements of wedges cannot be accurately described. The implicit information from the number of synonyms of the user input keyword can remedy this. These synonyms describe the user information need from multi-faceted aspects. However, each

synonym of the user input keyword may be associated with a number of widgets, which each associated widget is also assigned with various numbers of tags. In other words, the synonym of an input keyword and the tags are in an n-m relation via a number of widgets. Different widgets are regarded as different dimensions that measure the semantic similarities between the synonyms and tags. In order to quantify such a diverse relation between a user information need and tags, we make use of the Kullback-Leibler (KL) Divergence metric. As such, we can discover the mostly relevant tags to the user input. The algorithm is given below:

Input: an event

Output: the top 10 tags associated with the event.

Accept the keyword input of an event $e \in \mathcal{E}$ where \mathcal{E} is a universal set of events.

Find the synonyms of the event keyword to form a set of S of event e

for each $s_i \in S$

Retrieve all widgets that contain tag e_i

Store the retrieved widgets into a set W

end

for each $w_i \in W$

Retrieve all tags associated with w_i

Reduce the number of the tags by removing stop words

Store the rest of tags into a set T

end

// Calculate the semantic distance of the relation between each synonym and each tag

for each $s_i \in S$

for each $t_j \in T$

$$d(s_i, t_j) = \sum_{k=1}^{|W|} \left(p(w_k | s_i) \times \log \frac{p(w_k | s_i)}{p(w_k | t_j)} \right) \quad (1)$$

end

end

//calculate the average distance between event s and each tag

$$DA(e, t_j) = \sum_{i=1}^{|S|} \sum_{j=1}^{|T|} p(s_i | t_j) \times d(s_i, t_j) \quad (2)$$

Extract the nine tags with the highest DA scores.

As an example, we assume that a user wants to look for widget on travel. She may input the keyword is ‘Travel’. From WordNet, Miller [22], the algorithm receives a set of synonym words of ‘travel’ such as ‘travelling’, ‘change of location’, ‘locomotion’, ‘go’, ‘move’, ‘locomote’, ‘journey’, ‘trip’, ‘jaunt’, and ‘move around’. These words are stored in a set of S . These words are used for retrieving the widgets, the tags of which are also retrieved. The tags with suffix of ‘-ing’, ‘-s’, and ‘-ed’ are considered to be the same tag. The basic idea of Eq.(1) in the algorithm is that the semantic distance between a synonym of an input keyword and a tag is measured by the distributions of their associated widgets. The smaller the distance is, the closer their relation is. The DA value in Eq.(2) quantifies the average degree of a relation

between a tag and an event. In other words, the value implies the closeness between user intent for a wedge (an event) and each tag. The user intent is represented as an event, which is described by an input keyword, as well as its synonyms, rather than just the keyword.

Only the top ten tags associated with an event are selected for experiments. The reasons for this are as follows.

- Tackle the problem of overflowing tags for widgets. A number of widgets are attached with too many tags. The use all of the tags in the set T may result in retrieving the widgets that do not have the greatest relevance to the event. As an example, Fig. 1 illustrates a widget with 15 tags that results from issuing an event of “traveling”. Note that we use all 15 tags for this example. It is obvious that the retrieved result is quite different from the user input of an event. This is because the tags have the diverse meanings.
- Reduce the computation time for retrieving widgets. By reducing the size of the set to only ten tags from more than ten thousands will speed up the algorithm.



Fig. 1. An example of a widget with overflowing tags

3.2 Widget Discovery by Ranking

As a list of top tags has been extracted, the next step is to use them to discover widgets. In this process, three schemes are considered. Three schemes assign a different, respect value to a tag for ranking. The three schemes are described as follows:

1. Assign the same value to each top tag, say 1.
2. Assign its calculated DA score to each top tag. This score has been calculated by the proposed algorithm.

3. Assign its ranks to each top tag, i.e., the value depends on its ranking in the list. For instance, if the rank of a tag is 1, its value is 11; the rank 2, the value 10, and so on until the rank is 11 (value = 1).

The three schemes follow the same procedure. One of three proposed schemes could be selected as the main one, or all of them would be combined together. This selection depends on experimental evaluation result. The steps of the procedure are as follows:

1. As the value of each tag is available to the system depending on each schemes is used, this first steps is to go through each widget and determined their total tags value. Note that at this stage some widget might not have any value at all, which is considered as being irrelevant to the solution.
2. The system rearranges the list of widgets in descending order starting from the highest value of the total sum to the smallest one.
3. The threshold value is set as 1000, which is used to determine whether a widget is selected in the final list or not. This threshold value is currently set as 1000 to get rid of the widgets that are considered as unpopular, and may not be useful to the social. This could also means that they would not be useful for users.
4. Starting from the top of the list, the system extracts the widgets that have the installation number higher than the threshold number. This step continues until the final list has ten widgets, or all of the widget in the list is empty.
5. The final list of widgets is the combination of all three lists of widgets. To create a final combination of the widgets, the system multiples the widget values as the widget installation number, of which will give out the final widget result. This widget result is used to rank the widgets in order to get the widgets that are mostly useful.

After this procedure, there are in total 30 widgets in the list at the end, whereas the list is in the order of the total tag summation. This list is used in the future at the stage of final widget discovery. Fig. 2 illustrates the method flow.

4 System Design

This section presents the developed system and architecture. This system is implemented as a web application. The system makes use of the widget services from WidgetBox.com. The main objective of the system is automatic in that it can discover services that would match with a user's requirement and tolerance the social adaptability.

4.1 System Flow

There are two main flows in this system: back-end side and front-end side. On the back-end side, the system uses scraped widgets and their information from WidgetBox.com, while the front-end calculates the input information and creates a final widget discovery. Fig. 3 illustrates the flow diagram of the back-end system. Note that this process flow has to be repeated each time when a user enters a keyword. The input of this flow is the keyword entered by the user. In the case of this research, the keyword is 'travelling'. The output of this process is a list of widgets that is the most relevant to the input keyword.

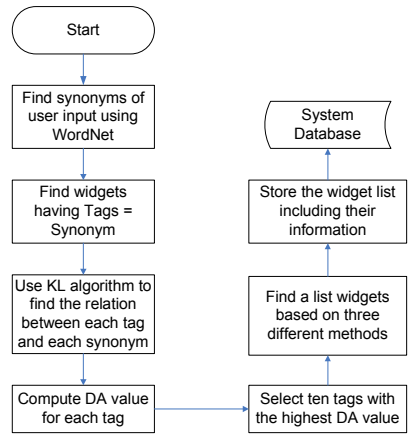
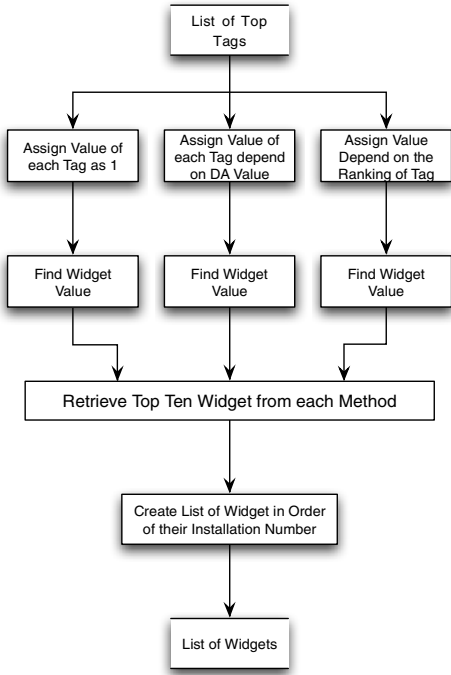


Fig. 2. Calculation schemes of the widget values **Fig. 3.** Flow diagram of the back-end system

4.2 System Interface

Fig. 4 is one of screenshots of the system. The four inputs are as follows:

Event title: the name of an event that a user selects;

Event Type: This is provided by the system currently.

Place: this is the location of a service. As mentioned before, it can be either abstract or specific locations

Party involved: Name of person involving in the event.

After all required information are filled and submitted, the system will generate a list of discovered widgets by running our algorithm and schemes.

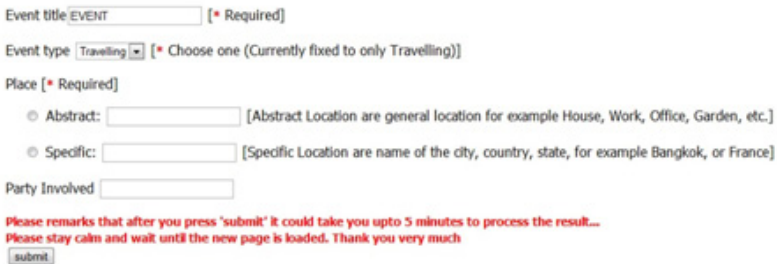


Fig. 4. A screenshot of the system

5 Experimental Results and Discussion

In this section, in terms of retrieved results, we compare our system with the Widget-Box.com system, which is based on the keyword matching discovery.

5.1 Datasets

In this research all of the data is retrieved from WidgetBox.com. The data and information are taken from that website using the scraping technique. The data are retrieved and stored in the database for computing the information and the future references. Table 3 lists the number of data that are used in this research. Note that the number of widgets in this system counts the only widgets that have relations with the input keyword of ‘Travelling’.

Table 3. Dataset information

| Total # Widgets in Widget-Box.com | # Widgets in Database | # Tags | Average # Tags per Widgets |
|-----------------------------------|-----------------------|---------|----------------------------|
| 234, 944 | 2, 924 | 29, 748 | 10.2 |

5.2 Tag Discovery Results

After following the methodology presented in Section 3, the final result of the tag discovery is a final list of ten tags that have the highest relation values with the input term. From the experiments, the top ten tags can best make use of tags information. More tags make no much difference because they include redundant information. Table 4 displays the result, which is ordered from the most related tag to the least one. Again, the data is generated as a result of the enquiry keyword of ‘travelling’.

Table 4. Top 10 tags retrieved

| Tag | DA Value |
|---------------|-----------|
| Blog | 0.020015 |
| Hotel | 0.020577 |
| Culture | 0.020877 |
| Vacation | 0.021136 |
| Photography | 0.0216759 |
| Entertainment | 0.0225166 |
| Life | 0.0248366 |
| Food | 0.0249321 |
| Art | 0.0254775 |
| Photo | 0.0254775 |

The list in Table 4 shows that term ‘blog’ has the strongest relationship with the event ‘travelling’. It has to be pointed out that the synonym may exist in the list. As might be noticed, the tag ‘Photo’ and ‘Photography’ have the similar meaning and the content is generally the equal. However, the system cannot separate them from each

other, on one hand to reduce the affect of ambiguous from the system of trying to detect the word with similar meaning from each. On the other hand, since the objective is to have the system running dynamically it would be more realistic and stick to the tag retrieved dynamically.

5.3 Widget Discovery Results

Table 5 gives 10 widgets that have been retrieved using the tag retrieved and through the widget discovery method. The underlined words in the table are the top tags retrieved.

Table 5. Top 10 widgets retrieved

| Widget Name | Tag | Installation | Rating (Out of 5) |
|---------------------------------------|---|--------------|-------------------|
| Been-Seen: Travel By Design | <u>travel</u> , travel blog, <u>hotel</u> , world, <u>travelling</u> , <u>travels</u> , travel tips, travel photos, <u>blog</u> , <u>blogging</u> , blogosphere, <u>culture</u> , design, <u>entertainment</u> , film, widget, <u>art</u> , blogosphere beenseen, travel by design, writing | 7,693 | 3.5 |
| USA Smarts | learning, geography, usa, quiz, us quiz, <u>blog</u> , community, <u>culture</u> , education, <u>entertainment</u> , marketing, reference, social networks, <u>travel</u> , web20 | 1,676 | 4 |
| French Word-A-Day | france, french, language, paris, provence, europe, <u>culture</u> , life in france, european, <u>travel</u> , wordaday, french words, pronunciation, books, <u>food</u> , interests, <u>photography</u> , pictures, writing, widget | 4216 | 4 |
| Forbes.com: Lifestyle | <u>life</u> , <u>travel</u> , <u>art</u> , beauty, celebrities, diet, fashion, fitness, <u>food</u> , home, real estate, shopping, style, sports, trends, women, forbes, interests, info | 1,591 | 5 |
| Britannica Blog | britannica, ideas, <u>blogging</u> , books, <u>culture</u> , <u>entertainment</u> , events, film, internet, leadership, reference, religion, science, social, <u>travel</u> | 1,286 | 1.5 |
| Trip Countdown | college, student, <u>travel</u> , widget, organize, plan, <u>vacation</u> , countdown, clock, uk, us, sta travel, interests, info | 20,712 | 4 |
| Live TV/Radio | live tv, radio, radio stations, worldwide, <u>entertainment</u> , humor, music, online, politics, religion, rss, video, videos, <u>travel</u> , technology, sports, social networks | 13,448 | 3.5 |
| The Bargainist Deals, Sales & Coupons | shopping, deals, bargains, coupons, discount, fashion, <u>food</u> , gadgets, internet, movies, <u>travel</u> , tech, sports, software | 13,337 | 3.5 |
| Trippermap - mapping Flickr | mapstraffic, <u>photo</u> , map, flickr, <u>travel</u> , photos, journey, world, google, earth, maps | 2,206 | 4.5 |
| Been-Seen: Travel By Design | <u>travel</u> , travel blog, <u>hotel</u> , world, <u>travelling</u> , <u>travels</u> , travel tips, travel photos, <u>blog</u> , <u>blogging</u> , blogosphere, <u>culture</u> , design, <u>entertainment</u> , film, widget, <u>art</u> , blogosphere beenseen, travel by design, writing | 7,693 | 3.5 |

5.4 Comparison

In the current WidgetBox.com system, the widget discovery is based on keyword matching. If a user inputs ‘travelling’, for example, it would find only the widgets that have the tag travelling. Table 6 reports the comparisons of the information from both systems based on the input of “Travelling”.

By comparing the data in the table, our algorithm clearly achieves a better performance. In particular, there are in total 78,416 installations (the number of users) for the proposed algorithm, while there are 8,055 installations in the keyword matching algorithm. Further, the installation average is 7,841.6 installations of the new algorithm, which is almost 10 times that of the keyword matching algorithm. This result clearly reflects the popularity of the widgets in the list. In other words, this reflects that the discovered widgets by the new algorithm capture a way better social popularity.

Table 6. Comparisons between key word matching and our algorithm

| Algorithm | Total Number of Installation | Average Number of Installation | Total Number of People Rating | Average Rating per Widget (Out of 5) |
|----------------------------|------------------------------|--------------------------------|-------------------------------|--------------------------------------|
| Keyword Matching Algorithm | 8,055 | 805.5 | 7 | 1.45 |
| Our Algorithm | 78,416 | 7,841.6 | 121 | 3.75 |

Moreover, the number of user ratings on the widgets retrieved by the new algorithm is 121, which is much higher than 7 by the keyword algorithm. The higher number indicates that the number of users participating in rating the widgets is higher. In other words, the widgets discovered by the proposed algorithm are more popular among users than those by the keyword algorithm.

In Table 6, the average rating of the tag relation algorithm is 3.75 out of 5, which is more than 3 times higher than that of the keyword algorithm. This validates that not only there are more participants, but also users are more satisfied with the widgets retrieved by the tag relation algorithm.

From the above comparison, it could be concluded that the retrieved widgets using the tag relation algorithm is better than those using the keyword algorithm.

6 Conclusion and Future Work

This paper has presented an overview on the use of social tagging in discovering contents and services. A new system that retrieves and ranks widgets from the Widget domain has been described. The proposed algorithm implemented in the system is able to rank the most relevant tags to a user query, and then to retrieve the best widgets. Together with the algorithm, a metric has been presented that quantifies the relation between user intent and the tags associated with widgets. By comparing with the keyword matching algorithm, our system has demonstrated its accuracy and efficiency. For the future work, we plan to test the quality of tags associated with widgets in order to make better recommendation to users.

Acknowledgment. Our sincere thanks to Mr. Anupong Muttaraid for his efforts on the system implementation.

References

1. Newcomer, E.: *Understanding Web Services: XML, WSDL, SOAP, and UDDI*. Addison Wesley, Boston (2002)
2. Martin, D., et al.: *OWL-S: Semantic Markup for Web Services*. W3C member submission, <http://www.w3.org/Submission/OWL-S/>
3. Akkiraju, R., et al.: *Web Service Semantics - WSDL-S*, W3C Member Submission, <http://www.w3.org/Submission/WSDL-S/>
4. Bruijn, J.D., et al.: *Web Service Modeling Ontology (WSMO)*. W3C Member Submission, <http://www.w3.org/Submission/WSMO/>
5. Bouillet, E., Feblowitz, M., Feng, H., Liu, Z., Ranganathan, A., Riabov, A.: *A Folksonomy-Based Model of Web Services for Discovery and Automatic Composition*. In: Proc. IEEE International Conference on Services Computing (SCC 2008), pp. 389–396. IEEE Press (July 2008), doi:10.1109/SCC.2008.77
6. Wal, T.V.: *Folksonomy*. Online Information. London, UK (2005)
7. Zhou, M., Bao, S., Wu, X., Yu, Y.: *An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations*. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) *ASWC 2007 and ISWC 2007*. LNCS, vol. 4825, pp. 680–693. Springer, Heidelberg (2007)
8. Chukmol, U., Benharkat, A.-N., Amghar, Y.: *“Enhancing Web Service Discovery by using Collaborative Tagging System*. In: Proc. 4th International Conference on Next Generation Web Services Practices (NWESP 2008), pp. 54–59. IEEE Press (October 2008), doi:10.1109/NWESP.2008.29
9. Aurnhammer, M., Hanappe, P., Steels, L.: *Augmenting Navigation for Collaborative Tagging with Emergent Semantics*. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) *ISWC 2006*. LNCS, vol. 4273, pp. 58–71. Springer, Heidelberg (2006)
10. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: *Visualizing tags over time*. *ACM Transactions on the Web (TWEB)* 1(2) (August 2007)
11. Wu, X., Zhang, L., Yu, Y.: *Exploring social annotations for the semantic web*. In: Proc. 15th International Conference on World Wide Web (WWW 2006). ACM Press (2006), doi:10.1145/1135777.1135839
12. Li, R., Bao, S., Yu, Y., Fei, B., Su, Z.: *Towards effective browsing of large scale social annotations*. In: Proc. 16th International Conference on World Wide Web (WWW 2007), ACM Press (2007), doi:10.1145/1242572.1242700
13. Ding, Z., Lei, D., Yan, J., Bin, Z., Lun, A.: *A Web Service Discovery Method Based on Tag*. In: Proc. International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2010), pp. 404–408. IEEE Press (February 2010), doi:10.1109/CISIS.2010.18
14. Bouillet, E., Feblowitz, M., Liu, Z., Ranganathan, A., Riabov, A.: *A Faceted Requirements-Driven Approach to Service Design and Composition*. In: Proc. IEEE International Conference on Web Services (ICWS 2008), pp. 369–376. IEEE Press (September 2008), doi:10.1109/ICWS.2008.117

15. Liu, X., Huang, G., Mei, H.: A User-Oriented Approach to Automated Service Composition. In: Proc. IEEE International Conference on Web Services (ICWS 2008), pp. 773–776. IEEE Press (September 2008), doi:10.1109/ICWS.2008.139
16. Liu, X., Zhao, Q., Huang, G., Mei, H., Teng, T.: Composing Data-Driven Service Mashups with Tag-Based Semantic Annotations. In: Proc. IEEE International Conference on Web Services (ICWS 2011), pp. 243–250. IEEE Press (2011), doi:10.1109/ICWS.2011.31
17. Gomadam, K., Ranabahu, A., Nagarajan, M., Sheth, A.P., Verma, K.: A Faceted Classification Based Approach to Search and RankWeb APIs. In: Proc. IEEE International Conference on Web Services (ICWS 2008), pp. 177–184. IEEE Press (September 2008), doi:10.1109/ICWS.2008.105
18. Arabshian, K.: A Framework for Personalized Context-Aware Search of Ontology-Based Tagged Data. In: Proc. IEEE International Conference on Services Computing (SCC 2010), pp. 649–650. IEEE Press (July 2010), doi:10.1109/SCC.2010.73
19. Miller, G.A.: WordNet: a lexical database for English. *Communications of the ACM* 38(1) (November 1995)
20. Case, D.O.: Looking for information: A survey of research on information seeking, needs, and behavior, 350 p. Academic Press, Amsterdam (2012)

An Automated Multiscale Map of Conversations: Mothers and Matters

Ansuya Ahluwalia¹, Allen Huang², Roja Bandari², and Vwani Roychowdhury²

¹ PEC University of Technology, Chandigarh, India
ansuyaahluwalia@outlook.com

² University of California, Los Angeles, USA
huang.allen@gmail.com, roja@ucla.edu, vwani@ee.ucla.edu

Abstract. By augmenting conventional techniques of topic modeling with unigram analysis and community detection, we establish an automated method that generates a comprehensive and meaningful summary of forum conversations over time that also sheds light on patterns of user behavior. We combine these methods to obtain a multiscale representation of what topics are being discussed, what the users are saying about each topic, how the conversation is evolving over time, and how friendships relate to content. As an example of our methodology, we examine discussion boards on Cafemom—an online hub for women to share their experiences and discuss their views on issues pertinent to child rearing. We apply the method with a focus on the issue of vaccination— a subject matter which has become controversial in recent years. We demonstrate how our methodology provides valuable insights into the evolution of conversations and highlights similarities in attitudes of socially connected users.

Keywords: Topic Modeling, Community Detection, Content Analysis, Vaccination, Forums.

1 Introduction

Fostering spaces for discussion and exchange of ideas is one of the central functions of the web. Discussion forums, social network messages, youtube comments, and social news services are examples of these spaces and the study of characteristics and dynamics of these environments has fueled an increasing amount of research in recent years.

While there is a common mental and emotional layer driving users to interact with content in various ways (user-content relations), online spaces also foster connection and interpersonal relationships (user-user relations). These two processes work together to create a dynamic environment of conversations where different topics become prominent at different times, are talked about in different ways, and where user friendships may resonate with emergence of some themes in the topic domain. Therefore, a fundamental need in the study of such spaces is to have at our disposal a fully automated method that provides a comprehensive summary of the dynamics of conversations without being cumbersome.

In this paper we achieve this goal using a multi-layered approach, considering both the temporal variations in content as well as friendship connections in the network. Since these dynamics can be observed at different granularities, we will also use a multi-scale approach utilizing different tools at different scales to reach a meaningful picture of these dynamics.

1.1 Related Work

A number of studies relevant to the current work center on mining and tracking opinions in product review websites. The goal of this family of literature is to extract summaries of online reviews, track user sentiment, or compare products (some examples are [18] [13] and [9]). In contrast with the data in the current paper, product reviews are often more structured, and there are known features of a product (such as the resolution of a camera) which users express positive or negative sentiment with respect to, so extracting and tracking feature-based opinion and sentiment is the focus of this family of studies.

Tracking topics, detecting events, and creating summaries of news content is the subject of another set of studies (e.g. [1]). News datasets are often curated and tagged, and are usually created by experts. In contrast, the current paper takes user-generated content in a public forum. Consequently, the data is extremely noisy and users are quite loosely self-organized around certain topics. Therefore the task of indexing and creating a granular summary of content and users becomes more challenging.

Finally, a number of papers aim at tracking changes in content across time by finding topics at consecutive time slots in the data and mapping them together [14] [20] [2] [5]. In the current paper, we instead detect topics over the whole corpus and use these topics to separate all posted content into topic categories. We then dial in to consecutive time-slots and get a more fine-grained perspective using unigram analysis in each of the topically separated categories of content. Although there are recent papers that propose more sophisticated topic evolution methods (e.g. incorporating temporal evolution in the definition of a topic [16]), in the end the current paper produces a simpler and more comprehensible summary and thus we believe is more readily usable. Our method demonstrates that simple tools used in proper succession can create a comprehensive multi-scale overview of a large forum with noisy data and that one can index this data at a highly granular level, indexing temporally, socially and content-wise.

1.2 Overview and Approach

We propose an automated methodology that provides a granular representation of content over time and reveals patterns of user behavior. The steps of this process are demonstrated in Figure 1. Using Latent Dirichlet Allocation topic modeling [6] on the text of forum posts, we generate a set of distinct themes prevalent in user discussions. These topics establish an initial framework by which we can classify conversations. Within the context of these topics we observe how conversations evolve over time. We find that subsequent sub-topic modeling of each time

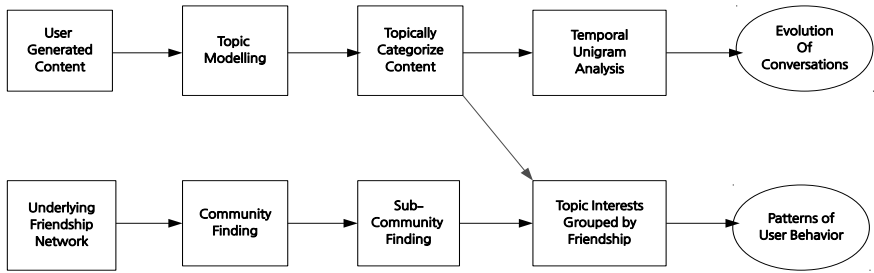


Fig. 1. Flowchart for general methodology used. Figure shows the interplay of content-driven and socially driven approaches employed to study the evolution of conversations and the patterns of user behavior.

segment produces an insufficient characterization of conversations. However, unigram analysis of the segments used in conjunction with topic modeling provides the depth and granularity needed to extract meaningful information. This method provides the right amount of detail without becoming too convoluted.

Concurrently, we perform community detection on the friendship network of users and find that there are clear ties between friendship communities and topics, implying that user connections are highly related to common topics of interest. Finally, we find similarities between topics in terms of user communities that participate in them and we find that some topics are highly correlated.

We implemented this methodology on an online platform called *Cafemom*, a forum for mothers to connect and discuss their views on a variety of issues. In this paper we focus our attention on conversations around vaccination and immunization. Vaccination has become an increasingly prominent topic in the public sphere and speculation about its adverse effects and concerns about safety have been on the rise [4, 21, 23]. These concerns range from short-term vaccine side effects to more serious ones such as the much discussed link between immunizations and autism [17, 12]. Consequently, public health officials are worried about public opinion leading to a drop in vaccination rates, exposing the population to dangerous epidemics.

Applying topic modeling to this dataset, we found areas such as *Religion*, *Autism*, *Government*, *Birth*, and *Food* with different levels of prominence at different times. Further unigram analysis within each topic created the next level of granularity; for example within the topic of *Government* the method was capable of capturing external events such as the 2008 presidential election as well as the 2011 tsunami in Japan and the resulting nuclear crisis. We then used the *Cafemom* friendship network to detect communities and sub-communities and found that a heatmap of communities and topics (Fig. 4) shows strong correlations among the two. Furthermore, a comparison among topics showed that some topics have positive or negative correlations based on the user communities active in them. For example, *Birth* and *Religion* correlate, whereas *Birth* and *Autism* are inversely correlated (more details in Section 5.2).

1.3 Outline

The rest of this paper is organized as follows: in Section 2, we describe the dataset, in Section 3 we discuss topic modeling and illustrate the temporal variation of topics. In Section 4, we compare sub-topic detection with unigram analysis in progressive time windows and show that simple unigram analysis provides more meaningful results at this granularity. In Section 5, we find communities in the friendship network and show that there is a high correlation among communities and topics and that some topics are highly correlated based on the communities of users who generate them. Section 6 discusses the findings and concludes the paper.

2 Data Characteristics

The dataset for this paper is obtained from forum posts in cafemom.com, a US-based online space where mothers discuss and exchange ideas on a variety of issues. Cafemom’s discussion boards are divided into groups (which are in turn divided into forums containing threads of individual posts), and while some portions are open to the public, a majority of the groups are private. Therefore, to access the complete data we create a membership profile and crawl all data from the discussion groups that appear in a keyword search for the relevant issue, i.e vaccination. We obtain a corpus consisting of 139,457 threads spanning 18 discussion groups with a total of 1,700,086 posts from 27,790 users over a span of around 5 years –Feb 6th 2007 to Apr 24th 2012. During this time, there were a total of 18,498,306 thread views (by users and non-users) [1].

3 Topic Generation

We employ Latent Dirichlet Allocation (LDA), an unsupervised method of topic discovery [6], to generate topics for this dataset. These topics help categorize the threads of the forum into distinct themes, and are the basis by which we study the evolution of user interests and concerns over time. In LDA, each document (in this paper we use threads as documents) is comprised of a mixture of topics and each word in a document can be ascribed to one of the topics generated. Listed in Table 1 are the top words for each of the ten sets of topics in the corpus. Note that topic names are assigned by the authors for the purpose of understandability and they are based on the inspection of the words in each topic [2]. In the next section we will describe the levels of prominence of each topic over time [15, 22].

¹ More details about data characteristics and activity on the site are available on http://rostatm.ee.ucla.edu/mediawiki/index.php/An_Automated_Multiscale_Map_of_Conversations:_Mothers_and_Matters

² We also perform sub-topic modeling on these topics to get a deeper insight on how these topics branch down to more granular sub-topics. Results for topic modeling and sub-topic modeling are available on

http://rostatm.ee.ucla.edu/mediawiki/index.php/An_Automated_Multiscale_Map_of_Conversations:_Mothers_and_Matters

Table 1. Top 19 words generated for 10 distinct topics found using LDA Topic Modeling

| Topic # | Top Words | Topic |
|---------|--|----------------------------|
| 0 | people god post group life make person agree read things time point understand women good wrong word feel thing | Religion and Ethics |
| 1 | love girl watch dog fun good show hair year pretty day life funny thought mom great big movie f**king | Love and Fun |
| 2 | time day son back things kids night put room good sleep thing bed home house times work car feel | Day to Day |
| 3 | vaccine vaccines children health autism flu disease research mercury study medical vaccination vaccinate risk cancer parents shots immune vaccinated | Vaccination |
| 4 | child kids children parents people life husband family time make feel things mom mother care school kid good thing | Family |
| 5 | baby group months birth doctor babies time give mother child born hospital moms mom good weeks pregnant feel breastfeeding | Birth and Babies |
| 6 | food eat water make milk good eating diet foods oil organic drink free buy made add weight stuff natural | Food |
| 7 | money home work free pay people make business job time company insurance month team working paid join check family | Money and Work |
| 8 | state government people law country public states news case obama american court america world rights health military police president | Government |
| 9 | autism school son child kids children year good autistic special admin teacher things pdd group great daughter asperger spectrum | Autism |

3.1 Topic Characteristics

LDA topic modeling using Mallet [19] produces a set of proportions associated with each topic for every document (i.e each thread) [6]. In other words, for each thread, we have a list of all ten topics in Table 1 along with the proportion or strength of each topic in that thread. Using these values, we categorized each thread under a topic in the following manner: In a thread, if the topic with the highest proportion has a proportion greater than 0.3, then the thread is categorized under that topic. The threshold is chosen as 0.3 because all such threads were found to have relatively low proportions for the other topics associated with that thread. 65.63% of the total threads fall under this criteria and are clearly associated most with one topic, and thus are used for further analyses.³ By only considering the threads that have a high topic proportion, we can map each thread to exactly one of the 10 topics.

Figure 2 shows the histograms of number of users and number of posts per topic. It can be observed that topics such as *Love and Fun* and *Family* in general

³ In the rest of the paper, we use only those 91,528 threads containing 1,339,250 posts that have a topic proportion greater than 0.3 for further findings.

have a greater volume of posts consistent with our intuition about these topics. On the other hand, a larger number of users post in topics such as *Autism*. Examining the growth of conversations over time as shown in Fig. 3, we find that topics such as Autism and Vaccination started receiving more attention from early 2007 lasting till 2009. From 2010 onwards, the activity levels declined and remained relatively constant. For *Autism*, there are peaks from around July to October 2007 and peaks around early 2008 for *Vaccination*.

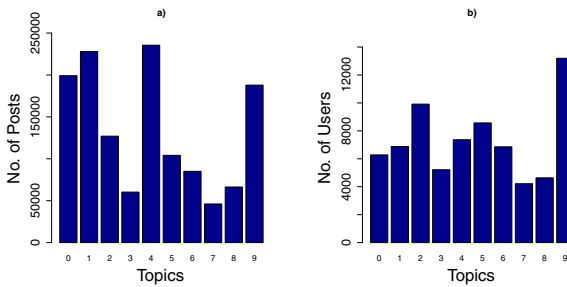


Fig. 2. a) Number of posts in Cafemom about each topic found in Table 1. b) Number of unique Cafemom users who posted on each topic.

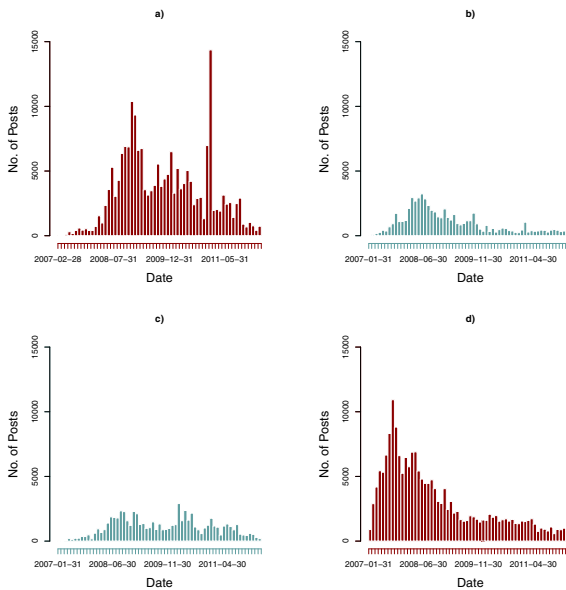


Fig. 3. Select topics exhibiting variable and stable posting activity. a)Religion. b)Vaccination. c)Government. d)Autism.

4 Evolution of Conversations

Unigram analysis can be used to create meaningful representations of the flow of information over time, especially in discovering the effect of external events on forum conversations. In the following sections we show how unigram analysis of posts categorized under each topic provides a more comprehensive depiction of user conversations than sub-topic modeling over time.

4.1 Unigram Processing

Tables 2 and 4 demonstrate the results of unigram analysis on topics of *Government*, and *Money* respectively. As described previously, forum posts are categorized under different topics and divided into 6-month time slots beginning from Feb 6th 2007. For all posts within a time slot, we perform tokenization using appropriate regular expressions, filter out the stop words, and create a bag of words. The term weight $w(t, d)$ for each unigram (or term) t in a time slot d is defined as

$$w(t, d) = \frac{tf(t, d)}{\max_t tf(t, d)} - \frac{tf(t)}{\max_t tf(t)} . \quad (1)$$

where $tf(t, d)$ is the term frequency of term t in time slot d , $\max_t tf(t, d)$ is the maximum term frequency of all terms in time slot d , $tf(t)$ is the term frequency of term t in all time slots, and $\max_t tf(t)$ is the maximum term frequency of all terms in all time slots. We then sort the unigrams in the order of decreasing term weight, filter out words that contribute as noise and select the top 20 unigrams for each time slot.

Looking at the results across the entire time span for two of the topics in Tables 2 and 4, we see many references to major external entities and events. Beginning in the August of 2008 for *Government* (Table 2), names of political candidates appear, capturing the Presidential Elections of 2008. Then in the first half of 2009, the discussion shifts to the topic of swine flu epidemic and the health issues relevant to the pandemic at that time. In Table 4, terms related to numerous major corporations and organizations such as Verizon, Walmart, and the Food and Drug Administration (FDA) are cited. Furthermore, concerns among moms about finance reflect the economic downturn when words such as poor and bankruptcy gain strength around the end of 2011, lasting till early 2012.

One can see that this simple yet fully automated approach provides a picture of the prominence of issues during different time periods while also establishing the context and showing how different topics are talked about.⁴

⁴ Top 20 unigrams for all 10 topics are available on http://rostan.ee.ucla.edu/mediawiki/index.php/An_Automated_Multiscale_Map_of_Conversations:_Mothers_and_Matters

Table 2. Timeline of unigrams of content categorized under topic- 'Government'

| | | | | | | | | | | |
|---|--|--|---|--|---|--|--|---|---|---|
| Feb '07- Aug '07 | Aug '07- Feb '08 | Feb '08- Aug '08 | Aug '08- Feb '09 | Feb '09- Aug '09 | Aug '09- Feb '10 | Feb '10- Aug '10 | Aug '10- Feb '11 | Feb '11- Aug '11 | Aug '11- Feb '12 | Feb '12- May '12 |
| Autism, Bill, Brigid, Children, Education, Vaccines, Immunization, Conference, California, Mercury, Alert | Autism, School, Press, State, Health, Medical, Vaccines, Services, Special, Education, Law | Autism, Drugs, Medical, Vaccine, Savage, Health, Marijuana, Hemp, Legal, California, FDA | Obama, McCain, Palin, Bush, President, Act, Vaccine, Iraq, Health, Vote, Tax, Campaign, FDA | Autism, Swine, Flu, Mexico, Health, Rights, States, North, Public, Illegal, Virus, Gun, Military, Ticker | CPS, Health, Swine, Flu, Emergenc, H1N1, Al-tamira, School, Pan-demic, Prison, Haiti, Nascar, Pot | Israel, Oil, System, Land, Ein-stein, Fetus, Ronald Reagan, Immigration, Palestinians, Abor-tion | CPS, School, CMSD, Political, Slavery, County, Smoke, CCD-CFS, Black, Court, Book, South, Separation | Gun, Home, Women, Japan, Abor-tion, Radiation, Death, Police, Sex, Nuclear, Reactor, Scienc-tology, Water, Jail | Police, Ticket, Speed, Reli-gious, Limit, Student, Traffic, Sticker, Afraid, File | Exemption, Reli-gious, Immu-niza-tion, School, Gov, Santorum, State, Law, Hos-pital, Zimmer-man, Board, Medical |

Table 3. Comparison of sub-topics with unigrams for the time period Feb 2011 to Aug 2011 for content categorized under topic- Government. Sub-topic modeling exhibits lesser granularity and clarity of important information aspects due to formation of overlapping topic clusters.

| Sub-Topic # | Top 5 words | Feb '11- Aug '11 |
|-------------|--|--|
| 0 | case home child gun court | Gun, Home, Women, Japan, Abortion, Radiation, Death, Police, Sex, Nuclear, Reactor, Scientology, Water, Jail |
| 1 | news found time water stu-dents | |
| 2 | state public government sys-tem school | |
| 3 | people states slavery food south | |
| 4 | people country time things war | |

Table 4. Unigrams timeline for content categorized under topic- 'Money and Work'

| | | | | | | | | | | |
|---|---|---|---|---|---|---|--|---|---|--|
| Feb '07- Aug '07 | Aug '07- Feb '08 | Feb '08- Aug '08 | Aug '08- Feb '09 | Feb '09- Aug '09 | Aug '09- Feb '10 | Feb '10- Aug '10 | Aug '10- Feb '11 | Feb '11- Aug '11 | Aug '11- Feb '12 | Feb '12- May '12 |
| SSI, Autism, Income, Med-icaid, Dis-ability, Quality, Insurance, Applied, Family, Job | Walmart, Autism, SSI, United, Tupper-ware, Family, Ac-count, Med-icaid, Mid-dot, Ther-apy, PCP, Medical | Business, GBG, Ameri-plan, Prosper-ity, Prod-uct, Down-line, Train-ing, Pros-Well-ness, Oppor-tunity, Vita-mins | Tally, Free, Secret, Parties, Work, Can-dies, Ebay, Risk, Inven-tory, Vac-ci, Trial, Well-ness, Con-sultant, Woo-mer, United, Shopper | Pay, Income, Food, Job, Free, Check, Insurance, Avon, Ac-count, Tax | Insurance, Pay, Med-icaid, Welfare, Health, Bill, Private, Ser-vices, Credit, Taxes, Food, Money, Afford, Cover, Tip, EIC | Pay, Job, School, Prop-erty, Neces-sity, Gro-cery, Ar-bonne, Tip, Prod-ucts, House, Unem-employment, Food | Baskets, Gift, Moms, Sales, Money, Inter-net, Free, Buy, Gold, Cards, Train-ing, Holiday, Debt | Tax, Money, House, Food, Job, Tip, Stamps, Kids, Welfare, Credit, Service, Loans, Car | Tax, School, Poor, Job, Wealth, Email, Coun-try, Wal-mart, Mil-i-tary, Rich, Face-book, Mort-gage | Insurance, Health-care, College, Train-ing, Cust-omers, Travel, Costs, Edu-cation, Com-pany, Verizon, Bankruptcy |

4.2 Sub-topic Modeling vs Unigram Analysis

Here we demonstrate the advantages of using unigrams over sub-topics to study the evolution of user generated content. We perform further topic modeling on content categorized under each topic at every time period and find several drawbacks. Table 3 compares these two methods during the time period –February to August 2011– for the topic of *Government*. The table lists the top 5 words in each of the five sub-topics (amounting to a total of 25 words). We can see that these 25 words not only have a great deal of overlap, but also bear no value in providing a concrete sense of what is being discussed. In contrast, the top 20 unigrams provide a much more detailed and diverse account of discussions during that period. Therefore, if we wish to create an efficient summary of the forums with as little human involvement as possible, the unigram approach is superior. We immediately see that within the topic of *Government*, the users were discussing issues of sex, abortion, and Japan’s nuclear crisis (a significant external event that happened during that time period).

Producing more topics (e.g 10 sub-topics instead of 5) in each time window and considering more words in each sub-topic (e.g top 20 words instead of top 5) will produce more reasonable results for sub-topic modeling method. However, this would require the study of an order of magnitude greater number of words (e.g 200 words) per time slot in order to extract any meaningful results. In contrast, considering even the top 10 words of the unigram analysis provides a picture that correlate well with external events.

Our methodology provides a simple, yet descriptive view of matters important to users. There are two main inferences drawn through these granular findings: (1) Study of temporal trends of references to external entities and the study of their recurrence and prominence, highlight the importance of latent administrative and governing bodies. (2) The interplay and intersection of topics of interest such as health, education, finance, politics, and law as evident from Table(s) 2 and 4 are indicators of the complexity with which certain topics behave on discussion forums.

5 Friendship Network Communities

In addition to the online discussion boards, Cafemom has an underlying friendship network. Out of 27,790 users, 16,731 (60% of all users) have friends on the site, which forms the underlying friendship network in our dataset. We use a greedy agglomerative community detection approach to cluster users in our network dataset. [8, 10].

The method used for community detection (described in [8]) optimizes the *modularity* –a measure of the distinctness of communities– across the entire network. The vertices (users) are clustered dendrogrammatically, with each vertex initially categorized as its own community. The communities are then iteratively joined until modularity is maximized [8]. The algorithm is fast, using data

structures suited specifically for sparse networks, making it an efficient clustering algorithm for a friendship network of this size. We wish to see how the community structure relates to user content in the context of topic modeling. The method of partitioning networks into sub-networks and the identification of themes based on unique characteristics have also been employed in other fields such those of neural networks [7].

5.1 Network Characteristics

After performing community detection on 16,731 users, we eliminate users belonging to communities having sizes less than 100, leaving us with 15,332 users. Community detection on this set of users produces 88 communities with a modularity of 0.5. We perform sub-community detection on the 5 biggest communities to break them down into smaller communities having sizes less than 1000 to make all communities comparable in size. The top 5 biggest communities have sizes 4030, 3508, 2572, 2546 and 1314 respectively. Further community detections on these 5 large communities yield modularities 0.53, 0.47, 0.65, 0.39 and 0.77 respectively. Our aim was to break down larger communities into smaller chunks in order to find more meaningful groups.⁵

5.2 Communities and Topics

We choose communities with sizes greater than 100 for topic tagging. There are 33 such communities comprising of 11,365 users. To tag communities based on the topic most discussed by that group of users, we calculate a weight for each topic belonging to a community. Every community has users who post in different topics. We assign each topic a count 1 if a user from that community posted for that topic and 0 if not. Topic weight $w(t, g)$ is defined as

$$w(t, g) = \frac{tc(t, g)}{\max_t tc(t, g)} - \frac{tc(t)}{\max_t tc(t)} . \quad (2)$$

where $tc(t, g)$ is the topic count for topic t in community g , $\max_t tc(t, g)$ is the maximum topic count of all topics in community g , $tc(t)$ is the topic count for topic t in all communities, and $\max_t tc(t)$ is the maximum topic count of all topics in all communities.

Comparing the topic scores within the community, we are able to identify the most popular topics for that community. Comparison of the topic scores among different communities (communities and sub-communities) provides a clear picture of the topic prominence for each community. Figure 4 is a heatmap generated for these 33 communities and shows which topics are more prevalent in a community. We find 15 communities that discuss *Autism* more than any other topic. Similarly all sub-communities for community 1 (10 communities) discuss

⁵ Graphs for the entire friendship network and the 5 biggest communities is available on <http://roostam.ee.ucla.edu/mediawiki/index.php/>

[An Automated Multiscale Map of Conversations: Mothers and Matters](#)

Birth and Babies more than anything else. From these findings we can speculate that friends on Cafemom share strong similarities around topics of interests. In fact, related work on user similarity suggest that these characteristics also affect user evaluations of each other [3].

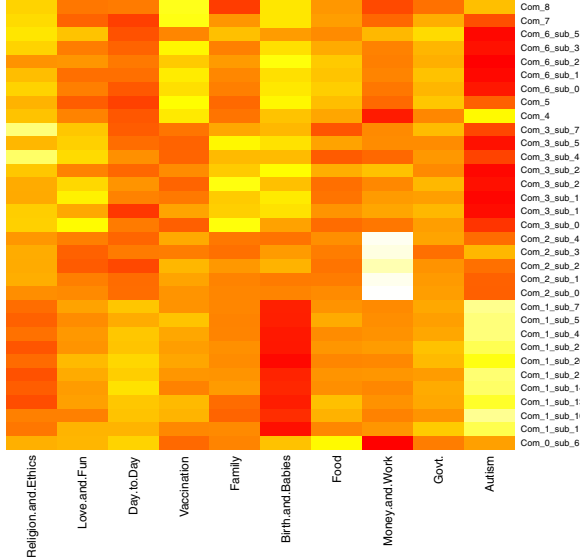


Fig. 4. Topics of interest among different friendship communities detected by a fast-greedy community detection algorithm based on modularity. Heatmap shows how certain communities are topic centric. Communities 6 and 3 along with its sub-communities show high affiliation for topic- Autism whereas community 1 and its sub-communities show high affiliation for topic- Birth and Babies.

Finally, we investigate the correlation between topics based on preference among different communities to discuss them. We calculate the pairwise correlation for all topics as follows. For each topic, we take a vector of its weights $w(t, g)$ over all communities. We compute the correlation matrix for these topics as $cor(u, v)$ where u and v are topic weight vectors. Table 5 shows the computed correlation matrix. Most notably, communities that post most often in *Birth and Babies* also post more in *Religion and Ethics*, with a 0.91 correlation index, and post the least in *Autism* (-0.22). The strong correlation between these two topics as dictated by user behavior reflects our intuition about the birthing process. Many issues in the birthing process have ethical or religious implications, ranging from issues of a natural birth to abortion (Refer top words for these topics stated in Table 4). It also stands to reason that women who are concerned about issues of birth have just or have yet to give birth and since autism is diagnosed usually only after 2 years of age [11], the topic will be of less importance to women in their gestation period or moms with new born babies. Similarly, communities with a high affiliation with *Autism* also post more frequently in *Day*

Table 5. Correlation matrix for topics based on topic weights per community i.e. community/sub-community. Matrix shows communities that are affiliated highly with one topic, also correlate with other topics. This correlation can be verified by examining the heatmap in Fig. 4. For example communities that post most in topic- Birth and Babies, also post highly in topic- Religion and Ethics and much less in topic- Autism.

| Topics | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0 | 1 | 0.75 | 0.38 | 0.73 | 0.81 | 0.91 | 0.78 | 0.19 | 0.88 | -0.22 |
| 1 | 0.75 | 1 | 0.7 | 0.3 | 0.94 | 0.65 | 0.6 | -0.12 | 0.78 | 0.01 |
| 2 | 0.38 | 0.7 | 1 | 0.17 | 0.58 | 0.29 | 0.52 | -0.45 | 0.48 | 0.6 |
| 3 | 0.73 | 0.3 | 0.17 | 1 | 0.4 | 0.79 | 0.81 | -0.06 | 0.71 | -0.17 |
| 4 | 0.81 | 0.94 | 0.58 | 0.4 | 1 | 0.76 | 0.66 | 0.01 | 0.83 | -0.11 |
| 5 | 0.91 | 0.65 | 0.29 | 0.79 | 0.76 | 1 | 0.85 | 0.11 | 0.81 | -0.36 |
| 6 | 0.78 | 0.6 | 0.52 | 0.81 | 0.66 | 0.85 | 1 | -0.19 | 0.82 | -0.05 |
| 7 | 0.19 | -0.12 | -0.45 | -0.06 | 0.01 | 0.11 | -0.19 | 1 | 0.01 | -0.41 |
| 8 | 0.88 | 0.78 | 0.48 | 0.71 | 0.83 | 0.81 | 0.82 | 0.01 | 1 | -0.15 |
| 9 | -0.22 | 0.01 | 0.6 | -0.17 | -0.11 | -0.36 | -0.05 | -0.41 | -0.15 | 1 |

to *Day*(0.6). Intuitively, parents with autistic children will have more questions and discussions concerning the daily happenings and challenges of caring for an autistic child. These findings give a qualitative evaluation of interests of friendship communities as well a quantitative evaluation of topic relationships based on user inclinations. Correlation of topics helps reveal patterns of user behavior and commonality of conversation interests shared among friends.

6 Concluding Remarks

In order to obtain a better understanding of user content and interactions on online forums, we propose an automated methodology that generates a comprehensive and multi-layered depiction of how forum conversations evolve over time and how friendships within a network highlight particular patterns in user conversations. By integrating unigram analysis and topic modeling temporally, we achieve a degree of detail and granularity of user content that efficiently captures external events such as the 2008 presidential election, the 2011 tsunami and nuclear disaster in Japan as well as references to major corporations and organizations such as Verizon and the Food Drug Administration. The level of specificity enables us to track how conversations progress over time. Furthermore, analysis of topic correlations based on friendship communities reveals how user-user interactions reflect inclinations of interest. We identify a strong positive correlation between topics of *Birth and Babies* and *Religion and Ethics* as well as between *Autism* and *Day to Day*. Correspondingly, we also see a strong negative correlation between *Birth and Babies* and *Autism*. By employing a methodology

that takes into account both the content-driven and socially driven aspects of forum conversations, we are able to efficiently generate a detailed summary of the dynamics of conversations as well as the similarities in interest among socially connected users. These results are exciting and present a path for future work where some of the issues in the current paper can be improved—for example choosing the number of topics was somewhat arbitrary. While we don't expect major shifts in the results, nevertheless the individual steps taken can be made more rigorous.

References

- [1] Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A.J., Teo, C.H.: Unified analysis of streaming news. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 267–276. ACM, New York (2011)
- [2] AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of the Eighth IEEE International Conference on Data Mining, ICDM 2008, pp. 3–12. IEEE Computer Society, Washington, DC (2008)
- [3] Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Effects of user similarity in social media. In: Proceedings of the fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 703–712. ACM, New York (2012)
- [4] Betsch, C., Renkewitz, F., Betsch, T., Ulshofer, C.: The influence of vaccine-critical websites on perceiving vaccination risks. *J. Health Psychol.* 15(3), 446–455 (2010)
- [5] Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 113–120. ACM, New York (2006)
- [6] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (2003)
- [7] Boscolo, R., Rezaei, B.A., Boykin, P.O., Roychowdhury, V.P.: Functionality Encoded In Topology? Discovering Macroscopic Regulatory Modules from Large-Scale Protein-DNA Interaction Networks. eprint arXiv:q-bio/0501039 (January 2005)
- [8] Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Physical Review E* 70(6), 066111 (2004)
- [9] Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web, WWW 2003, pp. 519–528. ACM, New York (2003)
- [10] Fortunato, S.: Community detection in graphs. *Physics Reports* 486(3-5), 75–174 (2010)
- [11] Fountain, C., King, M.D., Bearman, P.S.: Age of diagnosis for autism: individual and community factors across 10 birth cohorts. *J. Epidemiol. Community Health* 65(6), 503–510 (2011)
- [12] Freed, G.L., Clark, S.J., Butchart, A.T., Singer, D.C., Davis, M.M.: Parental vaccine safety concerns in 2009. *Pediatrics* 125(4), 654–659 (2010)
- [13] Ghani, R., Probst, K., Liu, Y., Krema, M., Fano, A.: Text mining for product attribute extraction. *SIGKDD Explor. Newsl.* 8(1), 41–48 (2006)

- [14] Gohr, A., Hinneburg, A., Schult, R., Spiliopoulou, M.: Topic evolution in a stream of documents. In: SDM, pp. 859–872. SIAM (2009)
- [15] Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101(suppl. 1), 5228–5235 (2004)
- [16] Jo, Y., Hopcroft, J.E., Lagoze, C.: The web of topics: discovering the topology of topic evolution in a corpus. In: *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*, pp. 257–266. ACM, New York (2011)
- [17] Kennedy, A., LaVail, K., Nowak, G., Basket, M., Landry, S.: Confidence about vaccines in the united states: Understanding parents perceptions. *Health Affairs* 30(6), 1151–1159 (2011)
- [18] Liu, B.: Opinion observer: Analyzing and comparing opinions on the web. In: *WWW 2005: Proceedings of the 14th International Conference on World Wide Web*, pp. 342–351. ACM Press (2005)
- [19] McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), <http://mallet.cs.umass.edu>
- [20] Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD 2005*, pp. 198–207. ACM, New York (2005)
- [21] Wolfe, R.M., Sharp, L.K., Lipsky, M.S.: Content and design attributes of antivaccination web sites. *JAMA* 287(24), 3245–3248 (2002)
- [22] Zhou, D., Ji, X., Zha, H., Giles, C.L.: Topic evolution and social interactions: how authors effect research. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006*, pp. 248–257. ACM, New York (2006)
- [23] Zimmerman, R.K., Wolfe, R.M., Fox, D.E., Fox, J.R., Nowalk, M.P., Troy, J.A., Sharp, L.K.: Vaccine criticism on the World Wide Web. *J. Med. Internet Res.* 7(2), e17 (2005)

How Influential Are You: Detecting Influential Bloggers in a Blogging Community

Imrul Kayes¹, Xiaoning Qian¹, John Skvoretz², and Adriana Iamnitchi¹

¹ Computer Science & Engineering, University of South Florida, Tampa, FL, USA

² Department of Sociology, University of South Florida, Tampa, FL, USA
imrul@mail.usf.edu, {xqian, anda}@cse.usf.edu, jskvoretz@usf.edu

Abstract. Blogging is a popular activity with high impact on marketing, shaping public opinions, and informing the world about major events from a grassroots point of view. Influential bloggers are recognized by businesses as significant forces for product promotion or demotion, and by oppressive political regimes as serious threats to their power. This paper studies the problem of identifying influential bloggers in a blogging community, BlogCatalog, by using network centrality metrics. Our analysis shows that bloggers are connected in a core-periphery network structure, with the highly influential bloggers well connected with each others forming the core, and the non-influential bloggers at the periphery. The six node centrality metrics we analyzed are highly correlated, showing that an aggregate centrality score as a measure of influence will be stable to variations in centrality metrics.

Keywords: social networks, influence, network centrality, blogosphere.

1 Introduction

The new age of participatory web applications commonly known as Web 2.0 has enabled the transition of the traditional information consumers into information producers in a form of *grassroots journalism* [1]. This kind of web applications include blogs, wikis, social annotation and tagging, and media sharing.

Blogging, in particular, distinguishes itself through both popularity and impact. For example, WordPress alone, a free and open source blogging tool, is used by over 14.7% of Alexa Internet’s “top 1 million” websites and as of August 2011 manages 22% of all new websites [2]. Citizen journalism had high impact in major events such as South Asia tsunami, London terrorist bombings, and New Orleans Hurricane Katrina [1]. The *blogosphere*, the virtual universe of the blogs on the web, provides thus a conducive platform for different aspects of virtual and real life, such as viral marketing [3], sales prediction [4, 5], business models [6], and counter terrorism efforts [7].

A blog (also referred to as a “web log”) is a personal journal published on the World Wide Web consisting of discrete entries (“posts”) typically displayed in reverse chronological order. Blogs are usually the work of a single individual, occasionally of a small group, and are often themed on a focused topic. A conventional blog may combine text, images and links to other blogs and to web

pages. Blogging platforms allow the creation of online profiles in which links to other bloggers are specified. These blogger-to-blogger ties specify the blogger’s interest and endorsement of other bloggers, creating a social network through which blog updates are automatically disseminated.

The influence bloggers have on forming public opinions is significant. First, bloggers influence other bloggers’ opinions: in 2011, 68% of bloggers claimed to be influenced by the blogs they read [8]. Second, they can influence the opinions of the masses: 38% of bloggers talk about brands positively and negatively on their blogs. Studies [9] show that 83% of people prefer consulting family, friends or an expert over traditional advertising before trying a new restaurant, 71% of people do the same before buying a prescription drug or visiting a place.

The advantages of identifying influential bloggers are already evident: influential bloggers are often market-movers. Identifying these bloggers can help companies better understand key concerns, identify new trends, and smartly affect the market by targeting influential bloggers with additional information to turn them into unofficial spokespersons [10]. About 64% of the companies are shifting their focus to blogging [11].

This paper investigates the position of influential bloggers in the BlogCatalog blogging community. Based on previous research [12, 13] that correlated a node’s position in the network to its influence, we conjecture that the influence of a blogger is represented by its location in the blogging network.

The contributions of this work are the following. First, we propose a method that aggregates different network position measurements into an overall influence score and demonstrate quantitatively that variation in one metric is not likely to significantly affect the aggregate score. Second, we discover that the overall pattern of the BlogCatalog community is that of a core-periphery structure, in which the highly influential bloggers are tightly connected to each other and the non-influential bloggers form the periphery.

The remainder of this paper is organized as follows: Section 2 presents the methodology of our quantitative study. Section 3 presents empirical results and analysis. Section 4 describes related work. We conclude in Section 5 with a discussion of the consequences of our results.

2 Methodology

The influence a node has on other nodes in the network can be represented in social network analysis by different centrality metrics. For example, the larger the number of direct neighbors, the larger an audience the node has for direct communication. Alternatively, the larger the number of paths between other pairs of nodes a node is part of, the more it can control the communication between distant nodes. Based on this intuition, we conjecture that a blogger’s influence is determined by and manifests via its centrality in the blogging community.

We propose to aggregate different representative centrality metrics into a final influence score. We define the influence score of a node as the average of

the positions of that node in decreasing order of centrality scores over various centrality metrics. Specifically, each centrality metric assigns each node a score that can be used to order nodes in decreasing order of importance (according to that centrality). This allows each blogger to receive a rank according to each centrality metric: the first ranked blogger will be the most central one, the last ranked will be the one with the lowest centrality score. Bloggers having the same centrality score are given the same rank. A blogger's final rank is the average rank over all centrality measures. We selected six representative centrality metrics as the focus of our study: degree, betweenness, closeness, eigenvector, hub, and communicability centrality.

Degree centrality is defined as the number of links that a node has. Although simple, degree centrality intuitively captures an important aspect of blogger's potential influence: bloggers who have connections to many others are read by more people, have access to more information, and certainly have more prestige than those who have fewer connections. High degree centrality bloggers can reach many bloggers directly.

Betweenness centrality, which measures the extent to which a node lies on the shortest paths between other nodes, was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network [14]. Bloggers with high betweenness centrality may have considerable influence within a network by virtue of their control over information passing between others: they can comment, annotate, re-interpret the posts originating from a distant blogger and these altered views can be seen by other remote bloggers. The nodes with highest betweenness are also the ones whose removal from the network will most disrupt communications between other nodes because they lie on the largest number of paths taken by messages [15]. Formally, the betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through :

$$C(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (1)$$

where v is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t) paths, and $\sigma(s,t|v)$ is the number of those paths passing through some nodes v other than s,t . If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t|v) = 0$. Our implementation of betweenness for this research is based upon the Brandes algorithm [16].

Closeness centrality measures the mean distance from a node to other nodes, assuming that information travels along the shortest paths. Formally, the closeness centrality ($C(x)$) of a node x is defined as follows:

$$C(x) = \frac{n-1}{\sum_{y \in U, y \neq x} d(x,y)} \quad (2)$$

where $d(x,y)$ is the distance between node x and node y ; U is the set of all nodes; d is the average distance between x and the other nodes. In our blogging network this centrality measure estimates the amount of information a blogger

may have access to compared to other bloggers. Specifically, a blogger with lower mean distance to others can reach others faster.

To account for the fact that not all communications take place along the shortest path, we also consider communicability centrality. This centrality measure is defined as the sum of closed walks of all lengths starting and ending at the node [17].

The centrality of a node does not only depend on the number of its adjacent nodes, but also on their relative importance. Eigenvector centrality allocates relative scores to all nodes in the network such that high-score neighbors contribute more to the score of the node. Formally, Bonacich [18] defines the eigenvector centrality $C(v)$ of a node v as the function of the sum of the eigenvector centralities of the adjacent nodes, i.e.

$$C(v) = 1/\lambda \sum_{(v,t) \in E} c(t) \quad (3)$$

where λ is a constant. This can be rewritten in vector notation, resulting in an eigenvector equation with well-known solutions.

Hubs and authorities are other relevant centralities for the blogging community context. Authorities are nodes that contain useful information on a topic of interest; hubs are nodes that know where the best authorities are to be found [15]. A high authority centrality node is pointed to by many hubs, i.e., by many other nodes with high hub centrality. A high hub centrality node points to many nodes with high authority centrality. These two centralities can play a significant role also in our work of finding influential bloggers. They can infer that the bloggers that have high hub and authority centrality are not only influential but also they are connected with influential bloggers.

3 Quantitative Analysis

We computed the centrality metrics presented before on a real dataset from the BlogCatalog blogging community. We implemented the algorithms in Python 2.7 with the NetworkX library for graph processing, and used `awk` for result processing.

3.1 Dataset

For our experiments we used the declared social network of bloggers on BlogCatalog (www.blogcatalog.com) available at [19]. BlogCatalog is a blogging service that allows its members to create online profiles, post their blogs, and automatically receive blogging updates from the BlogCatalog users with whom they have declared “friend” relationships. At the time of data collection, BlogCatalog relationships were symmetrical; at the time of this writing, however, BlogCatalog maintains directed relationships, similar to follower–followed relationships in Twitter. The dataset thus represents an undirected social graph. Where needed

Table 1. Average path length, radius, diameter and clustering coefficient of the BlogCatalog network compared to other networks

| Network | Nodes | Avg. path len. | Radius | Diameter | Clustering coefficient |
|-------------|-----------|----------------|--------|----------|------------------------|
| BlogCatalog | 10,312 | 2.38 | 3 | 5 | 0.460 |
| Orkut | 3,072,441 | 4.25 | 6 | 9 | 0.171 |
| LiveJournal | 5,284,457 | 5.88 | 12 | 20 | 0.330 |
| Erdős-Renyi | 10,312 | 2.65 | 3 | 3 | 0.006 |
| Web | 200M | 16.12 | 475 | 905 | 0.081 |

for centrality metrics computations, we treated an undirected edge as two directed edges, as it is typically done and supported by the meaning of an edge in our dataset. The network size is 10,312 nodes and 333,983 edges (average degree 64.78 and density 0.00628). The structural properties of the network are presented in Table 1.

For a brief characterization, we compare the BlogCatalog network properties with other networks from diverse domains: the LiveJournal blogging network [20], the Orkut online social network [20], the Web graph [21], and the Erdős-Renyi random graph of the same size as the BlogCatalog dataset ($|V| = 10,312$, $p = 0.00628$). Table 1 shows the average path length, radius, diameter and clustering coefficient of all five networks. A notable characteristic of the blogging communities is the high clustering coefficient compared to other networks. Given a network $G = (V, E)$, the clustering coefficient C_i of a node $i \in V$ is the proportion of all the possible edges between neighbors of the node that actually exist in the network [15]. A high clustering coefficient in both blogging networks implies that a blogger’s connections are interconnected and have a greater effect on one another. The small average path length (2.38), comparable with that of the corresponding random graph (2.65), together with the high average clustering coefficient, places the BlogCatalog network in the category of small-world graphs [22]. As in many other real networks [23], BlogCatalog exhibits scale-free properties. Figure 1 shows the complementary cumulative degree distribution of bloggers. The distribution fits a power-law distribution with exponent $\alpha = 2.52$. Most real-world networks with power-law degree distributions have values of α in the range $2 \leq \alpha \leq 3$ [15]. The most notable characteristic of a scale-free network is the occurrence of hubs, which hold a much higher number of links than the average node. As hubs control the “connectedness” of the network, we expect that influential bloggers also will be hubs in BlogCatalog.

3.2 Centralities and Influential Bloggers

As described in Section 2, we use centrality metrics to rank blogger’s importance in the network. Figure 2 shows cumulative distributions of various ranks. One of the objectives of plotting these distributions is to show how granular the ranks are, more specifically, how successful these centrality metrics are in assigning distinct scores to different nodes in the network. To this end, analyzing

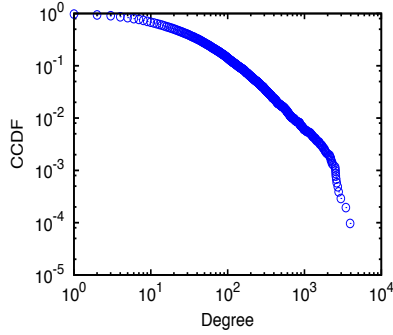


Fig. 1. Degree distribution in BlogCatalog

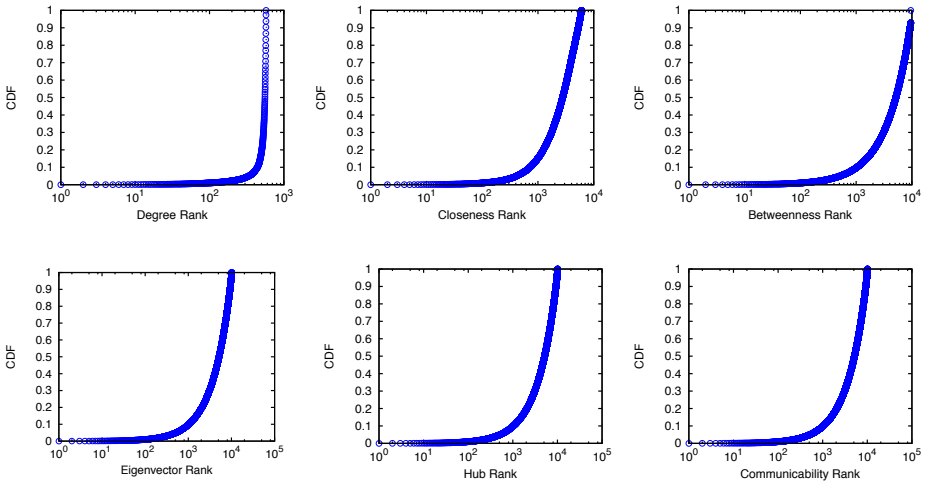


Fig. 2. Distribution of ranks of degree, closeness, betweenness, eigenvector, hub and communicability centralities

the distributions we get these facts: 5% of the bloggers cover the top 64% of the ranks in degree centrality scores, 12% of the bloggers correspond to top 12% ranks in closeness centrality, 10% of the bloggers correspond to top 10.80% ranks in betweenness centrality, 10% of the bloggers rank within 10.20% rank on eigenvector, hub rank distribution and 10% bloggers within top 10.19% rank on communicability rank distribution. So, we observe that all centrality measurements except degree centrality show granular scale of ranking, that is, they are typically capable of assigning a distinct score to each blogger (e.g., 10% bloggers within top 10.20% rank).

Bloggers that appear among the top 15 in multiple centrality metrics are represented in color in Table 2. The average rank of the top 10 most influential bloggers considering all centralities are shown in Table 3.

Table 2. The IDs of the top 15 bloggers according to each centrality measurement, sorted in increasing order by rank from left to right. DC: degree centrality, BC: betweenness centrality, CC: closeness centrality, EC: eigenvector centrality, HC: hub centrality, CoC: communicability centrality. Blogger IDs common to all centralities are colored the same. Black colored IDs represent bloggers who do not appear in the top 15 central bloggers from other centralities.

| | | | | | | | | | | | | | | | |
|------------|------|------|------|------|------|------|------|------|------|-----|------|------|------|------|------|
| DC | 4839 | 176 | 4374 | 8157 | 1226 | 4997 | 4984 | 8859 | 645 | 446 | 7098 | 7806 | 3198 | 2521 | 667 |
| BC | 176 | 4839 | 4374 | 8859 | 8157 | 645 | 1226 | 7806 | 233 | 446 | 3198 | 1932 | 4997 | 4984 | 7098 |
| CC | 4839 | 176 | 4374 | 8157 | 1226 | 4984 | 4997 | 8859 | 7098 | 645 | 7806 | 446 | 3198 | 2521 | 233 |
| EC | 4839 | 176 | 4374 | 1226 | 4984 | 8157 | 3198 | 4997 | 446 | 645 | 7098 | 2521 | 667 | 8859 | 4669 |
| HC | 4839 | 176 | 4374 | 1226 | 4984 | 8157 | 3198 | 4997 | 446 | 645 | 7098 | 2521 | 667 | 8859 | 4669 |
| CoC | 4839 | 176 | 4374 | 1226 | 4984 | 8157 | 3198 | 4997 | 446 | 645 | 7098 | 2521 | 667 | 8859 | 4669 |

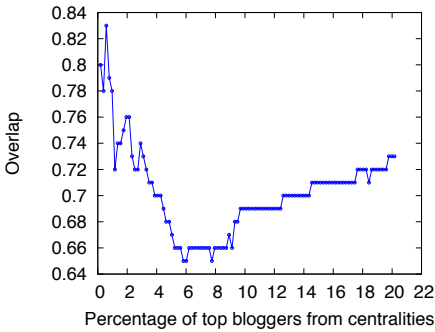


Fig. 3. Fraction of bloggers ranked in top x% by all centrality metrics

Table 3. Average rank of the top ten bloggers

| Bloggers' ID | Average Rank |
|---------------------|---------------------|
| 4839 | 1.17 |
| 176 | 1.83 |
| 4374 | 3.00 |
| 1226 | 4.83 |
| 8157 | 5.17 |
| 4984 | 7.00 |
| 4997 | 8.33 |
| 645 | 9.17 |
| 3198 | 9.17 |
| 446 | 9.83 |

3.3 Correlation of Centralities

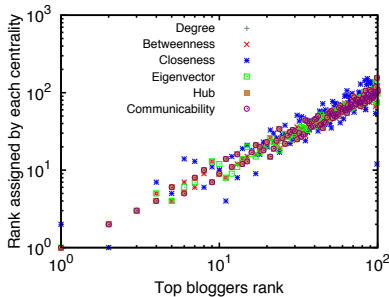
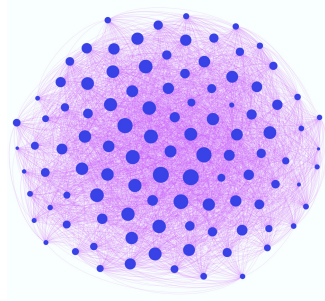
Out of the 15 most influential bloggers listed on each centrality, 12 bloggers (80%) are common in all the centralities. To better understand the correlation between these centrality measures, we run the following experiment.

We incrementally select the top bloggers according to each centrality metric (in increments of 0.2%, from 1 to 20%) and compute the fraction of bloggers who are common. The fraction of common top bloggers according to all centrality measures is shown in Figure 3. This fraction ranges from 0.65 to 0.83, showing that all centrality metrics considered tend to identify about the same individuals. More interestingly, the overlap is higher at the beginning, more specifically for the top 1% most central bloggers.

To observe more closely, we plot the ranks of top 1% of the bloggers assigned by all centralities, showed in Figure 4. As expected, a blogger’s assigned ranks from centralities form clusters and together with all the clusters we can visualize a straight line. This show that all the centralities tend to rank the same bloggers in the top.

Table 4. The correlation matrix of six centralities

| | Degree | Betweenness | Closeness | Eigenvector | Hub | Communicability |
|-----------------|--------|-------------|-----------|-------------|------|-----------------|
| Degree | 1.00 | 0.67 | 0.65 | 0.68 | 0.68 | 0.67 |
| Betweenness | 0.67 | 1.00 | 0.85 | 0.89 | 0.89 | 0.88 |
| Closeness | 0.65 | 0.85 | 1.00 | 0.98 | 0.98 | 0.97 |
| Eigenvector | 0.68 | 0.89 | 0.98 | 1.00 | 1.00 | 0.98 |
| Hub | 0.68 | 0.89 | 0.98 | 1.00 | 1.00 | 0.98 |
| Communicability | 0.67 | 0.88 | 0.97 | 0.98 | 0.98 | 1.00 |

**Fig. 4.** Assigned rank of top 1% most influential bloggers from all centralities**Fig. 5.** The subnetwork of the top 1% most influential bloggers, considering only ties among them. Node size is proportional to clustering coefficient.

We consider the blogger ranks as assigned by each centrality and calculate the Pearson correlation coefficient between each pair of centralities, as shown in Table 4. An entry (i, j) in the matrix denotes the correlation coefficient between $Centrality_i$ and $Centrality_j$. The values of the correlation coefficients are high, ranging from 0.65 to 1.00. The high values of correlation coefficients indicate a strong correlation among the centralities in terms of finding influential bloggers. This phenomenon has been observed by other studies: for example, Valente et al. [24] observed high correlation between four centralities: degree, betweenness, closeness, and eigenvector in a network of 58 users. Our study validates their findings using a significantly larger network and a larger set of centrality metrics.

3.4 Interrelations of Influential Bloggers

The average clustering coefficient of influential bloggers is low in BlogCatalog. For 1% of the influential bloggers the average clustering coefficient is 0.07, where the overall network average is 0.46. Figure 6 shows the clustering coefficients of the top 1% influential bloggers. The low clustering coefficients of the influential bloggers imply that they work as network ‘hubs’ in the BlogCatalog network.

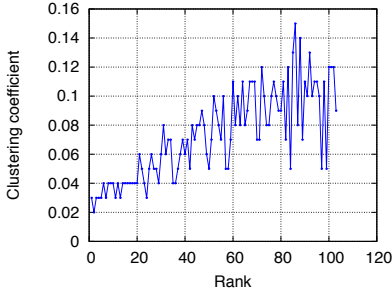


Fig. 6. The clustering coefficient of the top 1% most influential bloggers

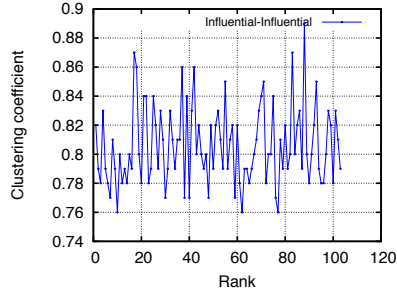


Fig. 7. The clustering coefficient of the top 1% most influential bloggers, considering only the ties among them

However, if we consider only the ties among influential bloggers, then the average clustering coefficient is very high, as shown in Figure 7: above 0.75, for an average of 0.81, thus significantly higher than the average of the entire network. Figure 5 depicts the sub-network of the most influential bloggers by representing the size of a node proportional to its clustering coefficient. Moreover, average path length of these bloggers is 1.22, where the network average is 2.38. This analysis shows that influential bloggers in BlogCatalog are highly connected, similar to the way influential users cluster in other communities (such as Facebook [13]). We define the subnetworks of the bloggers and compute assortativity coefficient as shown in Table 5. The assortativity coefficient is a measure of the likelihood for nodes with similar degree to connect to each other, and it ranges between -1 and 1 . A positive assortativity coefficient implies that nodes tend to connect to nodes of similar degree, while a negative coefficient implies the opposite. From the negative assortativity -0.25 of the entire network we can infer that nodes likely connect to nodes with very different degree from their own. Furthermore, exclusion of top 1% most influential bloggers from the network increases this trend of likelihood even more, which implicitly implies positive assortative mixing among influential bloggers. This implication is supported by the assortativity coefficient of $+0.07$ of the subnetwork of top 1% most influential bloggers, considering only ties among them. As such, the subnetwork of the top 1% most influential bloggers has negative assortativity (although less than the entire network) as they are connected with non-influential bloggers also. Along with the high clustering coefficient, we conclude that influential bloggers form a tightly-connected “core”, while the non-influential bloggers are located on the fringes of the network. A visualization of this phenomenon can be seen from Figure 8.

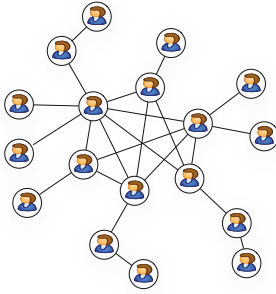


Fig. 8. Summary of interrelations of influential Bloggers

Table 5. Assortativity coefficient of different subnetworks

| Subnetwork Definition | Assortativity |
|--|---------------|
| Entire network | -0.25 |
| Subnetwork of bloggers excluding top 1% most influential bloggers | -0.67 |
| Subnetwork of top 1% most influential bloggers | -0.16 |
| Subnetwork of top 1% most influential bloggers, considering only ties among them | +0.07 |

4 Related Work

User influence is empirically elusive in social networks. Manski [25] states that user influence is difficult to identify in social observational data because influence is domain specific, thus domain-specific prior information is required. He argues that even if this information is available, the prospects for inference depend critically on the relationship between the variables defining the population. Inference is difficult to measure if these variables are statistically independent. In a similar vein, Aral et al. [26] observe diversity-bandwidth tradeoffs. The bandwidth of a tie is defined as the information transmission rate. Homophile nodes are connected by strong ties and interact more often, therefore have high bandwidth, but exchange little new information, whereas weak ties interact infrequently but are known to exchange new information. Both diversity of users and diversity of bandwidth are thus important for the diffusion of novel information. Since they are anti-correlated, there has to be a tradeoff to reach an optimal point in the propagation of new information.

Several approaches to identifying influential users have been proposed, including structural models [27], actor-oriented models [28], peer effects models [29], instrumental variable methods based on natural experiments [30], and ad hoc approaches based on specific data characteristics [31]. Our approach fits with structural models, as we used topological position as a measure of influence.

The problem of identifying influential bloggers in a blogging network has been studied empirically in BlogCatalog [32]. The authors compute a blogger’s influence score based on four measures: activity, recognition, novelty, and eloquence. The study finds that influential bloggers are not necessarily active bloggers, thus, only considering a blogger’s activity (e.g., number of posts or comments generated) may not reflect the blogger’s influence in the network. Our approach, instead, considers only the bloggers’ position in the network.

Domain-specific information has been used in other studies. Trusov et al. [33] identified influential users in online social networks based on longitudinal records of user log-in activity. They consider a user “influential” if her activity level, as captured by site log-ins over time, has a significant effect on the activity levels

of other users. They found that on average, approximately one-fifth of a user's friends actually influence the user's activity level on the site. By using users' real time activity correlated to that of their neighbors (thus, local network topology), this approach captures the local influence and disregards potential distant influences. Aral et al. [13] conducted an experiment to measure influence in the product adoption decisions of a representative sample of 1.3 million Facebook users. The experiment involved the random manipulation of influence-mediating messages sent from a commercial Facebook application. The application lets users share information and opinions about various social contexts. As users adopted and used the product, automated notifications of their activities were sent to randomly selected users of their social contacts. The study shows that influential individuals are less susceptible to influence than non-influential individuals and that they cluster in the network, while susceptible individuals do not. Influence in Twitter has been measured with TwitterRank [34], a variant of PageRank that also considers the topical similarity between users. Tang et al. [35] propose the UserRank algorithm which combines link analysis and content analysis techniques to identify influential users in an online healthcare forum. Han et al. [36] identify influential users in mobile social networks using fixed-length random walks.

New topology-aware centrality metrics have been proposed for measuring influence. Ilyas et al. [37] introduce the principal component centrality metric for identifying influential neighborhoods. The authors take eigenvector centrality as the *de facto* measure of node influence and identify influential nodes in Orkut that are not discovered by eigenvector centrality. This approach takes eigenvector centrality as the sole influence measure, while we consider multiple measures.

Customized ranks and topological similarities have been also studied in identifying influential users in different networks. Subbian et al. [38] propose the supervised Kemeny ranking aggregation method that combines different influence measures to produce a composite ranking mechanism. Ghosh and Lerman's [39] influence model use geodesic-path based distance measures and topological ranking measures. They introduce a normalized α -centrality algorithm that takes as input the score of a node (in this case, number of votes on Digg.com). This centrality measurement is domain dependent and can only be used in networks where voting feature is enabled.

Shetty et al. [40] proposed an entropy model for determining most influential nodes. Their social graph encodes nodes as persons or organizations and edges as the actions they are involved in. Influential nodes are those who affect the graph entropy most when they are removed from the graph (similar to hubs in our case). Zhang et al. [41] use PageRank or HITS link analysis algorithms for expert finding in a closed domain, assuming that the importance of a web page reflects the influence of its author in the social network. Our approach makes similar assumptions in that we also assume bloggers gain influence by virtue of staying structurally important in the network.

5 Summary

The blogging community has established itself as a fast growing and effective social media platform. Understanding influence within a blogging network is a problem with increasing relevance to marketing and information retrieval. We proposed a centrality aggregation method to measure relative influence scores of bloggers in the network. We apply our methodology to the BlogCatalog blogging community and learn the following: (1) some bloggers span significant influence on fellow bloggers due to their strategic location in the network; (2) the six network centrality metrics we studied (degree, betweenness, communicability, closeness, eigenvector and hub) are highly correlated in this community; and (3) influential bloggers form a densely connected core, while non-influential bloggers remain at the periphery of the network, less likely to connect to each other.

The core-periphery structure of the bloggers social network allows us to state the following hypothesis for future research: the structure of any discourse space will tend over time to a core-periphery pattern in which a small subset of contributors to the discourse will exercise hegemonic influence over the remaining vast majority of contributors. This hypothesis could apply, among others, to scientific disciplines viewed as discourse spaces.

Acknowledgements. We thank Nicolas Kourtellis and Xiang Zuo of University of South Florida for their feedback. The US National Science Foundation supported this research under grants CNS 0952420 and CNS 0831785.

References

- [1] Gillmor, D.: *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly (2006)
- [2] Rao, L.: Wordpress now powers 22 percent of new active websites in the u.s (2011), <http://techcrunch.com/2011/08/19/wordpress-now-powers-22-percent-of-new-active-websites-in-the-us/>
- [3] Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 61–70 (2002)
- [4] Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp. 78–87 (2005)
- [5] Mishne, G., de Rijke, M.: Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy. In: *Proceedings of the 15th International Conference on World Wide Web*, pp. 925–926 (2006)
- [6] Scoble, R., Israel, S.: *Naked conversations: how blogs are changing the way businesses talk with customers*. John Wiley (2006)
- [7] Coffman, T., Marcus, S.: Dynamic classification of groups through social network analysis and hmms. In: *Proceedings of IEEE Aerospace Conference*, pp. 3197–3205 (2004)

- [8] Technorati: State of the blogosphere 2011: Introduction and methodology (2011), <http://technorati.com/social-media/article/state-of-the-blogosphere-2011-introduction/>
- [9] Keller, E., Berry, J.: One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials. The Free Press (2003)
- [10] Aubrey, A.: Mcdonald's courts mom bloggers when changing the menu (2011), <http://www.npr.org/blogs/health/2011/07/27/138746335/mcdonalds-courts-mom-bloggers-when-changing-the-menu>
- [11] Elkin, T.: Just an online minute... online forecast (2005), <http://www.mediapost.com/publications/article/29803/just-an-online-minute-online-forecast.html/>
- [12] Aral, S., Muchnika, L., Sundararajana, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS* 106, 21544–21549 (2009)
- [13] Aral, S., Walker, D.: Identifying influential and susceptible members of social networks. *Science* 337, 337–341 (2012)
- [14] Freeman, L.: A set of measures of centrality based upon betweenness. *Sociometry* 40, 35–41 (1977)
- [15] Newman, M.E.J.: *Networks: An Introduction*. Oxford University Press (2010)
- [16] Brandes, U.: On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* 30, 136–145 (2008)
- [17] Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. *Physical Review E* 71, 056103 (2005)
- [18] Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* 2, 113–120 (1972)
- [19] Zafarani, R., Liu, H.: *Social computing data repository at ASU* (2009), <http://socialcomputing.asu.edu>
- [20] Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 29–42 (2007)
- [21] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web. In: *Proceedings of the 9th International World Wide Web Conference*, pp. 309–320 (2000)
- [22] Watts, D.J., Strogatz, S.: Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–442 (1998)
- [23] Barabasi, A.L.: *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume (2003)
- [24] Valente, T., Coronges, K., Lakon, C., Costenbader, E.: How correlated are network centrality measures? *Connections* 28, 16–26 (2008)
- [25] Manski, C.: Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60, 531–542 (1993)
- [26] Aral, S., Alstyn, M.V.: The diversity-bandwidth trade-off. *American Journal of Sociology* 117, 90–171 (2011)
- [27] Evans, D.: *Beyond Influencers: Social Network Properties and Viral Marketing*. Psychster Inc. (2009)
- [28] Snijders, T., van de Bunt, G., Steglich, C.: Introduction to actor-based models for network dynamics. *Social Networks* 32, 44–60 (2010)
- [29] Bramoull, Y., Djebbari, H., Fortin, B.: Identification of peer effects through social networks. *Journal of Econometrics* 150, 41–55 (2009)
- [30] Sacerdote, B.: Peer effects with random assignment: Results for dartmouth room-mates. *The Quarterly Journal of Economics* 116, 681–704 (2001)

- [31] Christakis, N., Fowler, J.: The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine* 357, 370–379 (2007)
- [32] Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: *Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 207–218 (2008)
- [33] Trusov, M., Bodapati, A., Bucklin, R.: Determining influential users in internet social networks. *Journal of Marketing Research* 47, 643–658 (2010)
- [34] Weng, J., Lim, E.P., Jiang, J., He, Q.: Twiterrank: finding topic-sensitive influential twitterers. In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 261–270 (2010)
- [35] Tang, X., Yang, C.: Identifying influential users in an online healthcare social network. In: *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, pp. 43–48 (2010)
- [36] Han, B., Srinivasan, A.: Your friends have more friends than you do: identifying influential mobile users through random walks. In: *Proceedings of the Thirteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 5–14 (2012)
- [37] Ilyas, M.U., Radha, H.: Identifying influential nodes in online social networks using principal component centrality. In: *Proceedings of IEEE International Conference on Communications*, pp. 1–5 (2011)
- [38] Subbian, K., Melville, P.: Supervised rank aggregation for predicting influencers in twitter. In: *SocialCom*, pp. 661–665 (2011)
- [39] Ghosh, R., Lerman, K.: Predicting influential users in online social networks. In: *Proceedings of KDD Workshop on Social Network Analysis* (2010)
- [40] Shetty, J., Adibi, J.: Discovering important nodes through graph entropy the case of enron email database. In: *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 74–81 (2005)
- [41] Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: *Proceedings of the 16th International Conference on World Wide Web*, pp. 221–230 (2007)

A Simulation Model Using Transaction Cost Economics to Analyze the Impact of Social Media on Online Shopping

Apratim Mukherjee¹, Shrabastee Banerjee², and Somprakash Bandyopadhyay^{3,*}

¹Dept. of Computer Science, BP Poddar Institute of Management & Technology,
Kolkata 700052, India
apratim2002@gmail.com

²Dept. of Economics, Lady Brabourne College, University of Calcutta,
Kolkata 700017, India
shrabasti.banerjee@gmail.com

³Management Information System Group, Indian Institute of Management Calcutta,
Kolkata 700104, India
somprakash@iimcal.ac.in

Abstract. In this paper, we have developed an agent-based simulation model to study the influence of social media on consumers' inclination towards on-line shopping. Social media includes web-based and mobile based technologies which are used to turn communication into interactive dialogue between organizations, communities, and individuals. Building upon the Transaction Cost Economics theory, the objective of our study is to examine the effect of social media on the "perceived transaction cost" of an individual, which determines his/her inclination to buy online. Transaction cost economics (TCE) theoretically explains why a transaction subject favors a particular form of transaction over others. Since purchasing from online stores can be considered a choice between the internet and traditional stores, it is reasonable to assume that consumers will go with the channel that has the lower transaction cost. Using agent-based models, we have studied the rate of adoption of on-line shopping by consumers and found it to be exponential, not linear.

Keywords: Transaction Cost, Social Network, Online shopping, Consumer Behavior.

1 Introduction

Since the advent of the internet, online shopping has progressively been gaining primacy throughout the world, drastically altering the established structure of markets [1,2,3]. This ascent of online stores is taking a toll on traditional markets. Amazon, one of the most successful online firms, is currently valued at over \$79 billion, which is 40 percent higher than the combined value of two large and successful offline

* Corresponding author.

retailers, Target and Kohl's, who have 2800 stores between them. Barnes & Noble, while still large, has also seen diminish of its market share [4].

Consumers interact, through processes such as imitation and conditioning, by means of individuals and groups of individuals (friends, family, etc.), which comprise the "social contacts" of the consumer [5, 6]. These contacts, according to their cohesion degree, influence more or less the consumer's purchase behavior. The online social network, which is a direct consequence of the same technological boom of the 90s that brought about the dominance of e-commerce, has revolutionized the way consumers interact with and influence each other. These interactions tend to change the "*transaction cost*" individuals associate with online shopping. A transaction cost is a cost incurred in making an economic exchange, i.e, it is the cost of participating in a market [7]. Transaction cost economics (TCE) is most commonly associated with the work of Oliver Williamson [8, 9, 10]. Using this transaction cost economics perspective, Teo, et al. [11] presents an empirical study for understanding consumers' on-line buying behavior. The results indicate that consumers' willingness to buy online is negatively associated with their perceived transaction cost.

In this paper, our objective is to demonstrate the adoption process of a consumer with regard to on-line shopping using an agent-based simulation model. We have studied the interaction among three entities to model this phenomenon: (i) Online Stores ($STORE_{online}$) (ii) online consumers ($CONSUMER_{online}$) and (iii) brick-and-mortar consumers who may be influenced into shopping online ($CONSUMER_{B\&M}$)

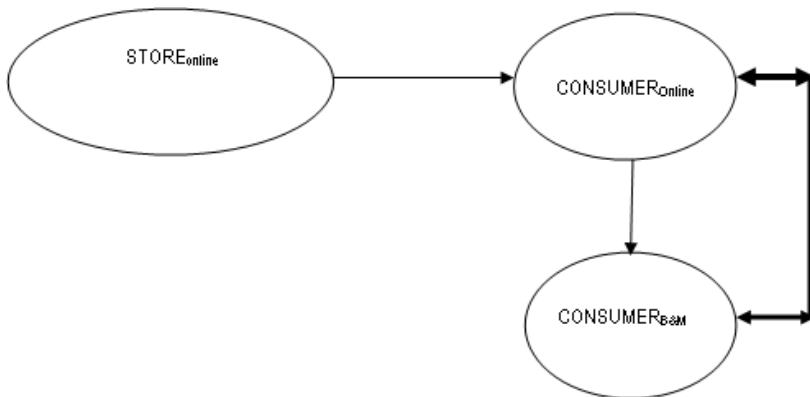


Fig. 1. An Interaction Model illustrating the adoption process

The influencing patterns between entities are illustrated below:

- ✓ $CONSUMER_{online}$ influences $CONSUMER_{B\&M}$ to move towards online shopping using online social networks. Thus, more consumers become online consumers (early majority); As this happens, these consumers in turn influence traditional consumers in their social circle, eventually turning them into online consumers (late majority).

- ✓ On some occasions online stores ($STORE_{online}$) may generate negative impacts through unacceptable service quality (delayed delivery, unacceptable product quality, etc) that may force some consumers in $CONSUMER_{Online}$ to go back to B&M shopping methods.

We will show how a consumer's perceived transaction cost is influenced through these interactions. Transaction cost economics (TCE) theoretically explains why a transaction subject favors a particular form of transaction over others. Since purchasing from online stores can be considered a choice between the internet and traditional stores, it is reasonable to assume that consumers will go with the channel that has the lower transaction cost. In this context, we study how online social networks accelerate this adoption process by constantly altering the transaction costs attached with online shopping, as perceived by individuals. The transition between $CONSUMER_{B\&M}$ to $CONSUMER_{Online}$ is covered in a considerably shorter period of time than it would have been in the absence of the social media.

To study these effects, agent-based simulation model is an invaluable tool [12]. Agent-based modeling is a bottom-up approach to understanding and analyzing complex, non-linear markets [13]. The method involves creating artificial agents designed to mimic the attributes and behaviors of their real-world counterparts. Using such a model, we may incorporate factors of social influence, heterogeneity, erratic rationality and in general present a fairly realistic picture of the consequences of inter-agent interaction. Agent-based simulations [14, 15] have offered during the last decade an interesting methodological issue and an innovative tool for specifying and validating behavioral individual models that are believed to be at the origin of emergent social and organizational phenomena. Using agent-based models, we create virtual populations including several hundreds of artificial consumers to study the rate of adoption of on-line shopping behavior by consumers.

2 A Transaction Cost Based Model

Transaction cost economics (TCE) theoretically explains why a transaction subject favors a particular form of transaction over others. The basic principle of TCE is that people like to conduct transactions in the most economic way. Since purchasing from online stores can be considered as a choice between the internet and traditional stores, it is reasonable to assume that consumers will go with the channel that has the lower transaction cost. Therefore, TCE becomes a viable theory for explaining the internet shopping decision of consumers. Specifically, whether a consumer would buy a product through the Internet is determined by the perceived transaction cost of the consumer [11]. Using this transaction cost economics perspective, Teo, et al. [11] shows that consumers' willingness to buy online is negatively associated with their perceived transaction cost.

Each consumer assigns a perceived transaction cost T to online shopping based on a set of factors like product uncertainty, convenience, economic utility, etc. The performance uncertainty of products bought online is one of the consumers' major concerns [11]. Thus, consumers' initial perception about high product uncertainty in

on-line shopping increases perceived transaction cost. Online buying, as an alternative to physical shopping, offers more convenience to consumers because they can save time and effort in searching for product information. Therefore, consumers perceive that convenience is high in on-line shopping and therefore perceived transaction cost is low in this respect. Also, the consumers perceive that economic utility is high in on-line shopping thus making perceived transaction cost low in this respect. Based on these factors, a consumer assigns an overall perceived transaction cost T to online shopping. The social interactions of a consumer will influence the values of T . A consumer becomes a $\text{CONSUMER}_{\text{Online}}$, once perceived transaction cost goes below a certain threshold value T^H .

Process of Influence and Adoption

Positive Influence. The adoption process is initiated when a consumer X , belonging to $\text{CONSUMER}_{\text{Online}}$, interacts with his/her friends within his/her social network framework. This process of interaction will influence another consumer Y , (who belongs to $\text{CONSUMER}_{\text{B\&M}}$ and is a friend of X ,) to alter his/her transaction cost. The nature of this interaction between the two is illustrated as follows:

Let $T(Y)$ is Y 's perception of Transaction costs associated with online shopping. Let us also assume that S is the *flexibility factor* of an individual ($0 < S < 1$) that determines how easily an individual can be influenced. S is closer to 0, when an individual is stubborn; on the other hand, S is closer to 1, when an individual is flexible and can easily be influenced.

After influence, each component of Y 's transaction cost will be reduced as follows:

$$T(Y_{\text{new}}) = T(Y) - S * T(Y)$$

So, when S is closer to 1, transaction cost of Y will reduce faster. Once $T(Y) < [\text{threshold } T^H]$, Y "crosses over" to become a member of the $\text{CONSUMER}_{\text{Online}}$ community. Gradually, the number of members in $\text{CONSUMER}_{\text{Online}}$ rises in the population.

Negative Influence. As indicated earlier, members belonging to the $\text{CONSUMER}_{\text{Online}}$ community have a perceived transaction cost below the threshold T^H that makes them inclined towards online shopping. However, this is not a static scenario. As indicated in figure 1, some on-line consumers may also experience negative influences from some on-line stores (delayed delivery, unacceptable product quality, etc) that may force them to go back to B&M shopping.

Let us assume that any consumer X belongs to $\text{CONSUMER}_{\text{Online}}$ community and let us assume that $T(X) = X$'s perception of Transaction cost associated with online shopping. Let us also assume that F is the negative Impact Factor s that depends on the nature of impact of the negative influence of $\text{STORE}_{\text{Online}}$ on X . After randomized influence (with a probability P) of $\text{STORE}_{\text{Online}}$ as stated above, X 's transaction cost will be increased as follows:

$$T(X_{\text{new}}) = T(X) + F * T(X)$$

For example, let us assume that X has received a product of unacceptable quality through on-line shopping and X associates an impact factor $F = 0.5$ for this event. So,

X's perceived transaction cost associated with online shopping ($T(X)$) will change as: $T(X_{\text{new}}) = T(X) + 0.5 * T(X)$. If this happens multiple times, $T(X)$ will eventually be greater than [threshold T^H] and X "crosses over" to become a member of $\text{CONSUMER}_{\text{B\&M}}$ community.

3 Agent Based Modeling and Simulation

Agent-based modeling and simulation (ABMS) is a new approach to modeling systems comprised of interacting autonomous agents [16]. Agents are autonomous decision-making entities or self-directed objects. Agent-based models are made up of agents and a framework for agent interactions. Agent-based modeling allows the behavior of system components (i.e., the agents) to be used to forecast the behavior of the overall system [17].

Economics is experiencing a paradigm shift in response to agent-based modeling. The field of Agent-based Computational Economics (ACE) has grown up around the application of ABMS to economic systems [12]. Some of the classical assumptions of micro-economic theory are: (1) Economic agents are rational; and, (2) Agents are homogeneous, having identical characteristics and rules of behavior. These assumptions are relaxed in ABMS applications to economic systems. Behavioral economics is a relatively new field that incorporates experimental findings on psychology and cognitive aspects of agent decision making to determine people's actual economic and decision making behavior. Thus, agent-modeling is a promising basis for modeling social life as interactions among adaptive agents who influence one another in response to the influences they receive [6].

The Simulation Framework

Using agent-based models, our goal is to create a virtual population of interacting Consumers and Stores to study the rate of adoption of on-line shopping by consumers. A variable transaction cost $T(n)$ is associated with each consumer n , which indicates his/her perception of the transaction cost associated with on-line shopping. When $T(n) < T^H$ (where T^H is the threshold transaction cost), the consumer will decide to purchase on-line.

Initially, there is a set of consumers $\in \text{CONSUMER}_{\text{Online}}$ with $T(n) < T^H$. They are the *early adoptors* of on-line shopping. During the course of interaction, they reduce the transaction cost of consumers $\in \text{CONSUMER}_{\text{B\&M}}$, as described in section 2. As a direct consequence of this, a consumer $\in \text{CONSUMER}_{\text{B\&M}}$ is now classified under $\text{CONSUMER}_{\text{Online}}$. Gradually, the number of $\text{CONSUMER}_{\text{Online}}$ rises in the population. However, some on-line consumers may also experience negative influences from some on-line stores that will increase their perceived transaction cost towards online shopping and may force them back to B&M shopping methods (depending on the value of the resultant transaction cost). The agent based simulation model that we develop will demonstrate the gradual changes in perceived transaction cost of an individual towards online shopping and will help us study this back and

forth movement between $\text{CONSUMERS}_{\text{online}}$ and $\text{CONSUMERS}_{\text{B\&M}}$ and its dependence on a set of specified parameters.

The proposed scheme is evaluated on a simulated social network environment under a variety of conditions to estimate the rate of adoption (of on-line shopping) against time. We present simulations for networks with 1000 and 10000 artificial consumers (the agents). Each consumer has an average number of friends N_n which is a variable simulation parameter. Also, we vary the initial number of online consumers (*Early Adopters*) who will act as *influencers* in the system. The adoption process is initiated when an agent $\in \text{CONSUMER}_{\text{Online}}$ interacts with its friends in the network. This process of interaction will influence an *agent* $\in \text{CONSUMER}_{\text{B\&M}}$ (provided they are friends) to reduce its transaction cost in favor of on-line shopping, as depicted in section 2.

These results are, of course, indicative and not validated by empirical studies. To demonstrate the usability of the model, we have generated T values randomly. For an initial set of agents $\in \text{CONSUMER}_{\text{Online}}$, the random values are chosen in such a fashion that T for each is less than the threshold $T^H (=10)$. For $\text{CONSUMER}_{\text{Online}}$ we choose random values between 1 to 10, while for $\text{CONSUMER}_{\text{B\&M}}$, random values between 11 to 50 (>10) are chosen.

We have studied the growth in number of online consumers against time as a function of following six parameters:

- Population Size participating in given social network environment (i.e total number of consumers under consideration, N)
- Average number of friends N_n per consumer
- Number of Early Adopters (i.e. initial number of online consumers), I , in given social network
- Flexibility factor, S ($0 < S < 1$) [section 2]
- Impact Factors F denoting magnitude of negative impact of $\text{STORE}_{\text{Online}}$ on $\text{CONSUMER}_{\text{Online}}$ (figure 1),
- Probability P of impact mentioned above. For example, $P=0.05$ means that for every 100 transactions, online stores create 5 negative impacts on online consumers with impact factor F .

4 Results and Discussions

4.1 Effect of N_n on Growth-Rate of Online Consumers

Figure 2 shows the growth-rate of online consumers against time with $I=10$ and average number of friends per consumer, i.e. $N_n = 10, 50$ and 100 respectively, where Impact Factors $F=1.0$ and Probability P of negative impact $=0.04$. The growth in number of online consumers depends on number of friends N_n per consumer. The growth saturates at 700 (70% of total population) at $N_n = 5$. This means that, under the given circumstance, the negative influence of $\text{STORE}_{\text{Online}}$ on consumers will hold back dynamically 30% consumers (300 out of 1000 on the average) towards B&M

shopping, when $N_n=5$. The fluctuating portion of the graph (between label B and C in figure 2) indicates dynamic transitions of consumers from $CONSUMER_{Online}$ to $CONSUMER_{B\&M}$, and vice versa, as shown and explained in figure 1.

But, as the number of friends per consumers increases to 10 and 15, the saturation point shifted up at 85 to 90%, (figure 2), indicating that stronger the influence of social network (more the number of friends), more will be the inclination towards online shopping.

Also, the growth-rate in figure 2 becomes sharper with increase in number of friends. Similarly, the starting point of the adoption process (point A in the graph) also decreases with increase in N_n , since a larger number of friends increases the probability of much faster adoption. For the same reason, positive influence of a large number of friends offsets the negative impact of online stores on consumers. As a result, the saturation point depends on the number of friends. However, the negative influence of $STORE_{Online}$ on consumers will always hold back dynamically a certain % of consumers towards B&M shopping.

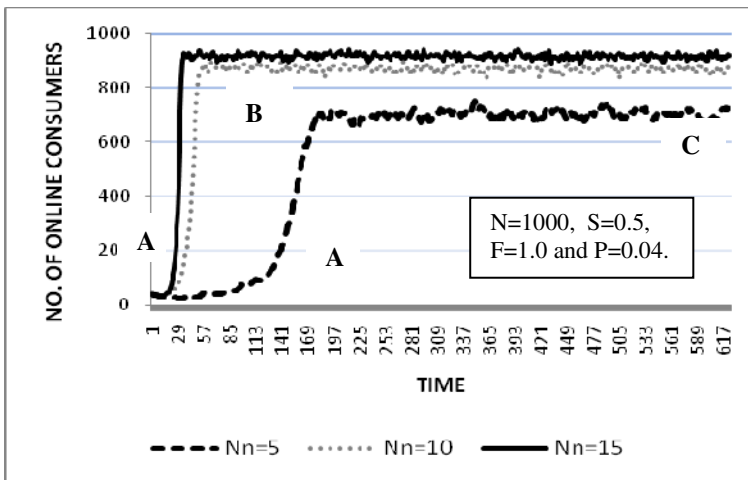


Fig. 2. Growth-rate of online consumers with $I=10$ at different N_n

4.2 Effect of F on Growth-Rate of Online Consumers

Negative Impact factor F and probability of negative impact P are responsible for increasing the transaction cost of an online consumer with respect to online shopping and consequently, pushing him/her back to B&M shopping method (when transaction cost $> T^H$). Fig 3 analyses the effect of F at $P=0.04$. As expected, when the negative impact factor of $STORE_{Online}$ is high (3.0), the number of consumers fluctuating between $CONSUMER_{Online}$ and $CONSUMER_{B\&M}$, and vice versa is quite high and both the growth-rate and saturation point are quite low (40%) compared to those at $F=1.0$ (where saturation point is at 70%).

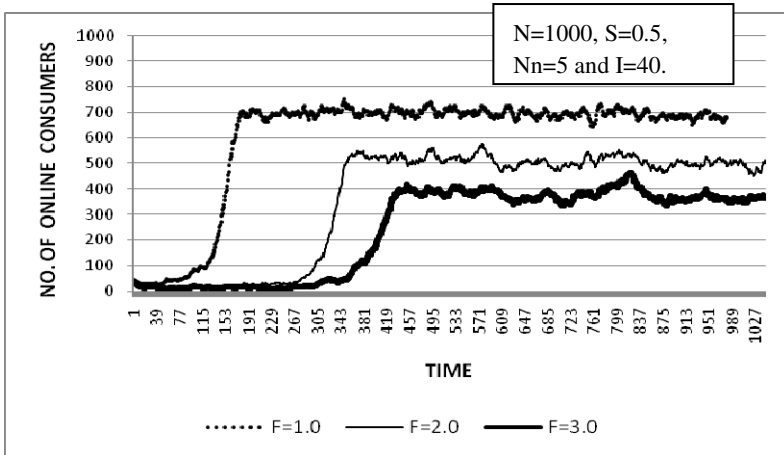


Fig. 3. Growth-rate of online consumers with $P=0.04$ at different F

4.3 Effect of P on Growth-Rate of Online Consumers

As indicated earlier, on-line consumers may experience negative impact from some on-line stores (delayed delivery, unacceptable product quality, etc) that may force them to go back to B&M shopping methods. P is the probability of such an impact. For example, $P=0.05$ signifies that online consumers suffer a negative impact once every 20 transactions. Fig. 4 depicts the effect of P at $F=1.0$. As expected, lower the probability of negative impact, higher will be the number of consumers inducted towards online shopping. So, at $P=0.01$, saturation level i.e. number of online consumers is almost 90% of the total population.

However, in the given situation, it is very interesting to note that if $P=0.05$ or higher, the entire population will move towards B&M shopping. Even the early adopters will not be able to tolerate the increased frequency of negative impacts (more than once in every 25 transactions) and they will become B&M shoppers.

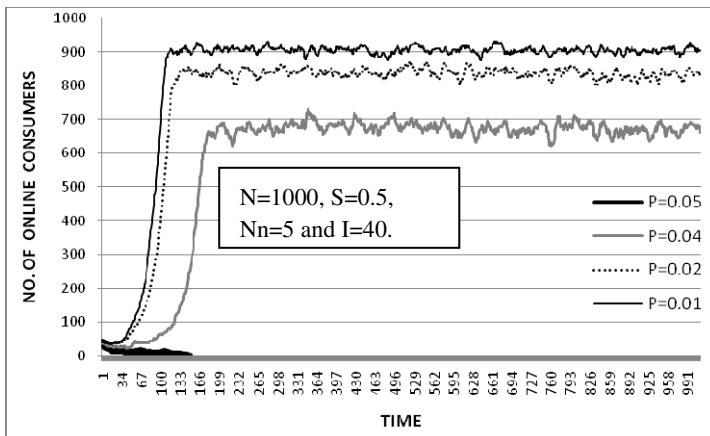


Fig. 4. Growth-rate of online consumers with $F=1.0$ at different P

4.4 Effect of N on the Growth Rate of Online Consumers

In order to test the scalability of our model, we have fixed all other parameters except N. Figure 5 shows the growth-rate of online consumers against time with $N=1000$, 3000, 5000 and 10000, where $N_n=5$, $I=0.4\%$, $S=0.5$, $P=0.04$ and $F=1.0$.

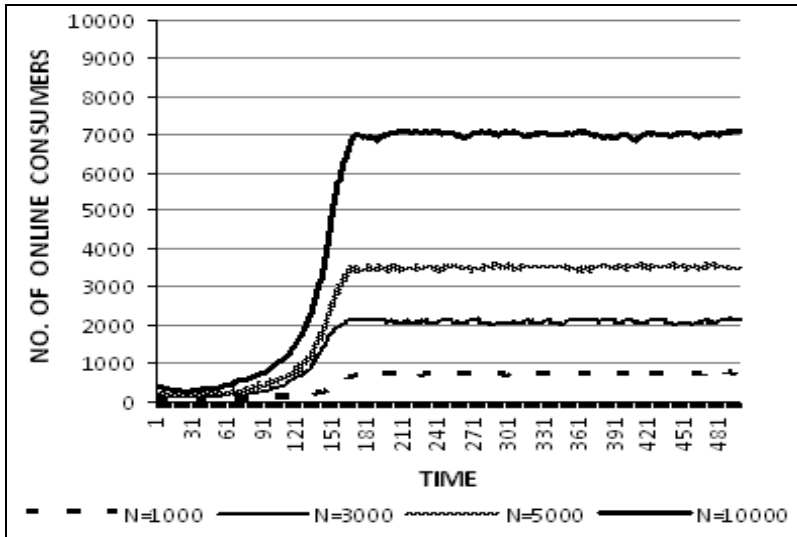


Fig. 5. Growth-rate of online consumers at different N

As shown, the *growth in number of online consumers does not depend on N*; it saturates at 70% of total population in all cases under the given circumstances. Hence, in the given scenario, the negative influence of $STORE_{Online}$ on consumers will always hold back dynamically 30% consumers on the average towards B&M shopping and it is independent of total number of consumers N.

4.5 Discussions

From the above observations, the following conclusions can be deduced:

- ✓ Social media accelerates online shopping adoption by constantly altering the transaction costs individuals associate with it. The transition between the “early adopters” (of online shopping) and the “early majority” is covered in a considerably shorter period of time than it would have been in the absence of the social networking effect. This is evident from fig 2 where the growth-rate becomes sharper with increase in number of friends and on-line social media is instrumental in increasing the average number of friends per consumer.
- ✓ However, the long-term success of online stores depends on their sustained performance. To keep up with the growing popularity, on-line stores will have to make constant upgradations to their services, and minimize errors as far as practicable. Otherwise, any negative impact they might generate will take a

heavy toll on their success, making it short lived. If they commit mistakes or fail to give sufficient attention to customers' needs, the negative Impact factor F and probability of negative impact P of online stores will increase the transaction cost of existing online consumers with respect to online shopping and consequently, will push them back to B&M shopping, as shown in fig.3 and fig. 4. Social media will propagate this negative impact much faster than it would have been without the presence of social media.

5 Conclusion

Several researchers have predicted the gradual growth in shares of online retailers to be linear [18]. However, those analyses have not considered the impact of on-line social networks, which is exerting a heavy influence on consumer purchase behavior and diffusion (from offline to online). Most on-line retailers are endeavoring to tie-up consumers' shopping activities with their presence in social networks (such as Facebook or Twitter). Thus, neutral consumers are quickly getting influenced in favor of shopping online. Hence, the predicted growth rate is exponential, and not linear. Brick-and-mortar firms, i.e conventional retailers must thus pursue a strategy of omnichannel retailing—an integrated approach that combines the advantages of physical stores with the experience of online shopping [18]. Only then can they survive the paradigm shift we have exemplified here.

It is to be noted that this is a model building exercise. However, once equipped with suitable empirical data, we can supply each of these values individually. By examining the parameters and altering them, we can obtain different trends to illustrate how fast influence takes over, how trends vary with country/ town/ population, etc.

References

1. Smith, M.D., Bailey, J., Brynjolfsson, E.: Understanding digital markets: review and assessment. In: Brynjolfsson, E., Kahin, B. (eds.) *Understanding the Digital Economy: Data, Tools and Research*, MIT Press, Cambridge (1999)
2. Wigand, R.T.: Electronic commerce: definition, theory, and context. *The Information Society* 13, 1–16 (1997)
3. Wang, Z.: Technological Innovation and Market Turbulence: The Dot-Com Experience. *Review of Economic Dynamics* 10(1) (2007)
4. Lieber, E., Syverson, C.: Online vs. Offline Competition. Prepared for the Oxford Handbook of the Digital Economy (January 2011)
5. Zielinski, J., Robertson, T.S., Ward, S.: *Consumer behavior*. Scott Foresman Series in Marketing (1984)
6. Antonides, G.: An attempt at integration of economic and psychological theories of consumption. *Journal of Economic Psychology* 10(1), 77–99 (1989)
7. Rindfleisch, A., Heide, J.B.: Transaction cost analysis: past, present and future applications. *Journal of Marketing* 61 (1997)

8. Williamson, O.E.: Transaction cost economics: the governance of contractual relations. *Journal of Law and Economics* 22 (1979)
9. Williamson, O.E.: The economics of organization: the transaction cost approach. *American Journal of Sociology* 87 (1981)
10. Williamson, O.E.: *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. The Free Press, New York (1985)
11. Teo, T.S.H., Wang, P., Leong, H.C.: Understanding online shopping behaviour using a transaction cost economics approach. *Int. J. Internet Marketing and Advertising* 1(1) (2004)
12. Tesfatsion, L.: *Agent-based Computational Economics (ACE)* (2005), <http://www.econ.iastate.edu/tesfatsi/ace.htm>
13. Paul, T., Cadman, R.: Agent-based modeling of customer behavior in the telecoms and media markets. *Info.* 4(1) (2002)
14. Ben Said, L., Drogoul, A., Bouron, T.: Multi-Agent Based Simulation of Consumer Behaviour: Towards a New Marketing Approach. In: Ghassemi, F., Post, D.A., Sivapalan, M., Vertessy, R. (eds.) *MODSIM 2001 International Congress on Modelling and Simulation*, vol. 1, pp. 71–76. Modelling and Simulation Society of Australia and New Zealand (2001)
15. Sallach, D., Macal, C.: The simulation of social agents: an introduction. *Social Science Computer Review* 19(3), 245–248 (2001)
16. Macal, C.M., North, M.J.: Tutorial on Agent-Based Modeling and Simulation Part 2: How to Model with Agents. In: *Proceedings of the IEEE 2006 Winter Simulation Conference*, Monterey, CA, December 3–6 (2006)
17. North, M.J., et al.: Multiscale Agent-Based Consumer Market Modeling. *Complexity* 15(5), 37–47 (2010)
18. Darrell, R. (2011). The future of shopping. *Harvard Business Review* (December 1, 2011)

Predicting Group Evolution in the Social Network

Piotr Bródka, Przemysław Kazienko, and Bartosz Kołoszczyk

Institute of Informatics, Wrocław University of Technology,
Wrocław, Poland

{piotr.brodka,kazienko}@pwr.wroc.pl,
bkoloszczyk@gmail.com

Abstract. Groups – social communities are important components of entire societies, analysed by means of the social network concept. Their immanent feature is continuous evolution over time. If we know how groups in the social network has evolved we can use this information and try to predict the next step in the given group evolution. In the paper, a new approach for group evolution prediction is presented and examined. Experimental studies on four evolving social networks revealed that (i) the prediction based on the simple input features may be very accurate, (ii) some classifiers are more precise than the others and (iii) parameters of the group evolution extracion method significantly influence the prediction quality.

Keywords: social network, group evolution, predicting group evolution, group dynamics, social network analysis, *GED*.

1 Introduction and Related Work

In most fields of science, researchers struggle to predict the future: the future consumption of power in electric network, future load of network grid, future consumption of goods etc. Social networks are no different. Recently, the main focus is on the link prediction [13], but there are also papers on (i) entire network structure modelling [18], (ii) modelling social network evolution [12], [15], or (iii) churn prediction and its influence on the network [10], [19]. However only few researchers have considered groups in the prediction process. Some of them like Zheleva et al. are using communities only for link prediction [20], the others like Kairam et al. tries to identify and understand the factors contributing in the growth and longevity of groups within social networks [9]. Unfortunately, there is no research directly regarding prediction of the entire group evolution. Probably, the main reason behind this is the fact that the methods for determining group history have not been good enough so far.

The approach presented in this paper, involves usage of the results produced by the *GED* method [3] to predict group evolution. The assumption is that using the information about preceding changes of a given group and its characteristic in the past, as the

input for the classifier, which was previously trained based on the historical changes of other groups in the social network, we can try to predict the next step in the given group evolution. Based on this assumption, a new approach for group evolution prediction was developed and it is presented and examined in this paper. The results of the first experiments on four evolving social networks revealed that (1) the prediction based on the proposed input features may be very accurate, (2) some classifiers like C4.5 decision trees or random forests are more precise than the others and (3) parameters of the group evolution identification method (GED) [3] significantly influence on the prediction quality.

2 GED Group Evolution Discovery

The concept of GED method and its full evaluation was presented in [3]. In this paper only the most important elements are presented, in order to help the reader understand the next chapters.

2.1 Temporal Social Network and Groups

Temporal social network TSN is a list of following timeframes (time windows) T . Each timeframe is in fact social network $SN(V, E)$ where: V – is a set of vertices and E is a set of directed edges $\langle x, y \rangle: x, y \in V$

$$\begin{aligned}
 TSN &= \langle T_1, T_2, \dots, T_m \rangle, \quad m \in N \\
 T_i &= SN_i(V_i, E_i), \quad i = 1, 2, \dots, m \quad , \\
 E_i &= \langle x, y \rangle: x, y \in V_i, \quad i = 1, 2, \dots, m
 \end{aligned}
 \tag{1}$$

2.2 Group Evolution

Group evolution is a sequence of events (changes) succeeding each other in the consecutive time windows (timeframes) within the social network. Possible events in social group evolution are:

1. *Continuing* (stagnation), when groups in the consecutive time windows are identical or when groups differ only by few nodes and their size remains the same.
2. *Shrinking*, when nodes has left the group, making its size smaller than in the previous time window. Like in case of growing, a group can shrink slightly as well as greatly.
3. *Growing* (opposite to shrinking), when new nodes has joined to the group, making its size bigger than in the previous time window. A group can grow slightly as well as significantly, doubling or even tripling its size.

4. *Splitting* occurs, when a group splits into two or more groups in the next time window. Like in merging, we can distinguish two types of splitting: equal and unequal, which might be similar to shrinking.
5. *Merging*, (reverse to splitting) when a group consist of two or more groups from the previous time window. Merge might be (1) *equal*, which means the contribution of the groups in merged group is almost the same, or (2) *unequal*, when one of the groups has much greater contribution into the merged group. In second case merging might be similar to growing.
6. *Dissolving*, when a group ends its life and does not occur in the next time window.
7. *Forming* of new group, which has not exist in the previous time window. In some cases, a group can be inactive over several timeframes, such case is treated as dissolving of the first group and forming again of the second one.

2.3 GED – A Method for Group Evolution Discovery in the Social Network

To discover group evolution in the social network a method called *GED* (Group Evolution Discovery) was used [3]. The most important component of this method is a measure called inclusion. This measure allows to evaluate the inclusion of one group in another. Therefore, inclusion $I(G_1, G_2)$ of group G_1 in group G_2 is calculated as follows:

$$I(G_1, G_2) = \frac{\overbrace{|\mathbf{G}_1 \cap \mathbf{G}_2|}^{\text{group quantity}}}{|\mathbf{G}_1|} \cdot \frac{\sum_{x \in (G_1 \cap G_2)} NI_{G_1}(x)}{\underbrace{\sum_{x \in (G_1)} NI_{G_1}(x)}_{\text{group quality}}} \quad (2)$$

where $NI_{G_1}(x)$ is the value reflecting importance of the node x in group G_1 .

As a node importance $NI_{G_1}(x)$ measure, any metric which indicate member position within the community can be used, e.g. centrality degree, betweenness degree, page rank, social position etc. The second factor in Equation 2 would have to be adapted accordingly to selected measure.

The *GED* method, used to discover group evolution, respects both the quantity and quality of the group members. The *quantity* is reflected by the first part of the *inclusion* measure, i.e. what portion of members from group G_1 is in group G_2 , whereas the *quality* is expressed by the second part of the *inclusion* measure, namely what contribution of important members from group G_1 is in G_2 . It provides a balance between the groups that contain many of the less important members and groups with only few but key members. The procedure for the *GED* is as follows:

Input: Temporal social network *TSN*, in which groups are extracted by any community detection algorithm separately for each timeframe T_i and any node importance measure is calculated for each group.

1. For each pair of groups $\langle G_1, G_2 \rangle$ in consecutive timeframes T_i and T_{i+1} inclusion $I(G_1, G_2)$ for G_1 in G_2 and $I(G_2, G_1)$ for G_2 in G_1 is computed
2. Based on both inclusions $I(G_1, G_2)$, $I(G_2, G_1)$ and sizes of both groups only one type of event may be identified:
 - a. *Continuing*: $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| = |G_2|$
 - b. *Shrinking*: $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| > |G_2|$ OR $I(G_1, G_2) < \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| \geq |G_2|$ OR $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) < \beta$ and $|G_1| \geq |G_2|$ and there is only one match between G_1 and groups in the next time window T_{i+1}
 - c. *Growing*: $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| < |G_2|$ OR $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) < \beta$ and $|G_1| \leq |G_2|$ OR $I(G_1, G_2) < \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| \leq |G_2|$ and there is only one match between G_2 and groups in the previous time window T_i
 - d. *Splitting*: $I(G_1, G_2) < \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| \geq |G_2|$ OR $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) < \beta$ and $|G_1| \geq |G_2|$ and there is more than one match between G_1 and groups in the next time window T_{i+1}
 - e. *Merging*: $I(G_1, G_2) \geq \alpha$ and $I(G_2, G_1) < \beta$ and $|G_1| \leq |G_2|$ OR $I(G_1, G_2) < \alpha$ and $I(G_2, G_1) \geq \beta$ and $|G_1| \leq |G_2|$ and there is more than one match between G_2 and groups in the previous time window T_i
 - f. *Dissolving*: for G_1 in T_i and each group G_2 in T_{i+1} $I(G_1, G_2) < 10\%$ and $I(G_2, G_1) < 10\%$
 - g. *Forming*: for G_2 in T_{i+1} and each group G_1 in T_i $I(G_1, G_2) < 10\%$ and $I(G_2, G_1) < 10\%$

For more detailed description of *GED* Method and its evaluation see [3].

3 The Concept of Using the *GED* Method for Prediction of Group Evolution

Presented approach, involves usage of the results of *GED* method. The assumption is that using a simple sequence, which consists only of several preceding groups' sizes and events, as an input for the classifier, the learnt model will be able to produce very good results even for simple classifiers.

The sequences of groups sizes and events between timeframes can be extracted from the *GED* results. In this paper 4-step sequences were used (Figure 1). Obviously, the event types varied depending on the individual groups, but the time frame numbers were fixed. It means that for each event four group profiles in four previous time

frames together with three associated events were identified as the input for the classification model, separately for each group. A single group in a given time frame (T_n) was a case (instance) for classification, for which its event $T_n T_{n+1}$ was predicted.

The sequence presented in Figure 1 was used as an input for classification. The first part of the sequence was used as 7 input features (variables), i.e. (1) **Group size in T_{n-3}** , (2) **Event type $T_{n-3}T_{n-2}$** , (3) **Group size in T_{n-2}** , (4) **Event type $T_{n-2}T_{n-1}$** , (5) **Group size in T_{n-1}** , (6) **Event type $T_{n-1}T_n$** , (7) **Group size in T_n** . A predictive variable was the next event for a given group. Thus, the goal of classification was to predict (classify) **Event $T_n T_{n+1}$ type** – out of the six possible classes: i.e. (1) growing, (2) continuing, (3) shrinking, (4) dissolving, (5) merging and (6) splitting. Forming was excluded since it can only start the sequence.

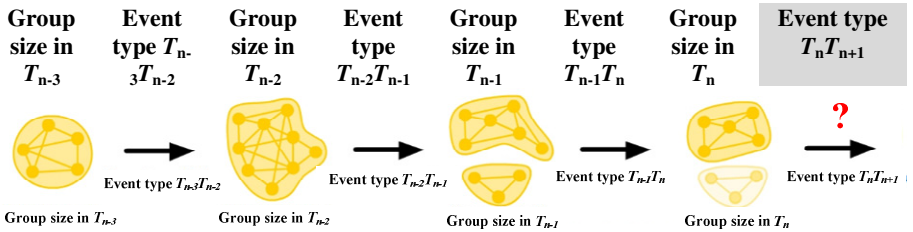


Fig. 1. The sequence of events for a single group together with its intermediate sizes (descriptive input variables) as well as its target class - event type in $T_n T_{n+1}$. It corresponds to one case in classification

4 Experiment Setup

As mentioned before, the notion, which was checked during the experiments, was that using the information about preceding changes of a given group as well as its description in the past as the input for the classifier, trained based on the historical transitions of other groups, we can try to predict the next step in the given group evolution.

To check this four temporal social networks TSN have been extracted from four different datasets to perform and evaluate prediction of group evolution.

1. The first network was extracted from Wrocław University of Technology email communication. The whole data set was collected within the period from February 2006 to October 2007 and consists of 5,845 nodes (distinct university employees' email addresses) and 149,344 edges (emails sent from one address to another). The temporal social network consisted of fourteen 90-days timeframes extracted from this source data. Timeframes have the 45-days overlap, i.e., the first timeframe begins on the 1st day and ends on the 90th day, the second begins on the 46th day and ends on the 135th day and so on.
2. The third social network was extracted from the portal www.salon24.pl, which is dedicated especially to political discussions, but also some other subjects from different domains may be brought up there. The network consists of 3,775 nodes and 77,932 edges. There are 12 non-overlapping timeframes representing 12 months of the 2009 year.

3. The fourth one is the well-known Enron e-mail network with 150 nodes and 2,144 edges. The network was split into twelve, 90-days timeframes without overlap.
4. The fifth network was extracted from the portal extradom.pl. It gathers people, who are engaged in building their own houses in Poland. It helps them to exchange best practices, experiences, evaluate various constructing projects and technologies or simply to find the answers to their questions provided by others. The data covers a period of 17 months and contains 3,690 users and 34,082 relations. 33 timeframes were extracted, each of them 30 days long with 15 days overlap, similarly to the first data set.

For each timeframe social communities were extracted using CFinder [16] and for each *TSN* the *GED* method [3] was utilized to extract groups evolution. The *GED* method was run 36 times for each *TSN* with all combination of α and β parameters from the set {50%, 60%, 70%, 80%, 90%, 100%}. As a node importance measure the social position measure [21] (measure similar to page rank) was utilized.

Next, the 4-step sequences were separately extracted from the *GED* results for all networks and every combination of α and β parameters, see an example sequence in Figure 1.

Experiment was performed in WEKA Data Mining Software [7]. Ten different classifiers were utilized with default settings: (see Table 1). For the method of validation 10-fold cross-validation was utilized as the most commonly used [14]. In WEKA, this means 100 calls of one classifier with training data and tested against the test data in order to get statistically meaningful results.

Table 1. WEKA classifiers used

| WEKA name | Name |
|---------------|-----------------------------------|
| BayesNet | Bayes Network classifier [7] |
| NaiveBayes | Naive Bayesian classifier [8] |
| IBk | k-nearest neighbor classifier [1] |
| KStar | Instance-Based classifier [4] |
| AdaBoost | Adaboost M1 method [6] |
| DecisionTable | Decision table [11] |
| JRip | RIPPER rule classifier [5] |
| ZeroR | 0-R classifier |
| J48 | C4.5 decision tree [17] |
| RandomForest | Random forest [2] |

5 Results

All classifiers were utilized for each of 4 networks and each combination of α and β parameters. The measure selected for presentation and analysis of the results is F measure which is the harmonic mean of precision and recall.

At the beginning, the classifiers were compared for each dataset separately in order to indicate which one is the best. The results are presented in Table 2 and Figures 2-5.

Table 2. The classifiers comparison for each dataset

| Data set | Classifier | Max F measure | Min F measure | Diff. |
|-------------|---------------|---------------|---------------|-------|
| salon24.pl | BayesNet | 1.00 | 1.00 | 0.00 |
| | NaiveBayes | 1.00 | 1.00 | 0.00 |
| | IBk | 1.00 | 1.00 | 0.00 |
| | KStar | 1.00 | 1.00 | 0.00 |
| | AdaBoostM1 | 1.00 | 0.70 | 0.30 |
| | DecisionTable | 1.00 | 0.90 | 0.11 |
| | JRip | 1.00 | 0.97 | 0.03 |
| | ZeroR | 0.82 | 0.60 | 0.23 |
| | J48 | 1.00 | 0.99 | 0.01 |
| | RandomForest | 1.00 | 1.00 | 0.00 |
| Enron | BayesNet | 0.83 | 0.69 | 0.15 |
| | NaiveBayes | 0.81 | 0.72 | 0.08 |
| | IBk | 0.79 | 0.71 | 0.08 |
| | KStar | 0.79 | 0.72 | 0.07 |
| | AdaBoostM1 | 0.51 | 0.32 | 0.20 |
| | DecisionTable | 0.78 | 0.64 | 0.14 |
| | JRip | 0.80 | 0.73 | 0.07 |
| | ZeroR | 0.27 | 0.15 | 0.11 |
| | J48 | 0.92 | 0.80 | 0.13 |
| | RandomForest | 0.89 | 0.76 | 0.13 |
| extradom.pl | BayesNet | 0.87 | 0.54 | 0.32 |
| | NaiveBayes | 0.87 | 0.50 | 0.37 |
| | IBk | 0.88 | 0.55 | 0.33 |
| | KStar | 0.88 | 0.52 | 0.36 |
| | AdaBoostM1 | 0.83 | 0.50 | 0.33 |
| | DecisionTable | 0.88 | 0.48 | 0.39 |
| | JRip | 0.88 | 0.35 | 0.53 |
| | ZeroR | 0.88 | 0.33 | 0.54 |
| | J48 | 0.88 | 0.33 | 0.55 |
| | RandomForest | 0.88 | 0.40 | 0.48 |
| WrUT emails | BayesNet | 0.86 | 0.76 | 0.10 |
| | NaiveBayes | 0.86 | 0.73 | 0.13 |
| | IBk | 0.88 | 0.79 | 0.09 |
| | KStar | 0.88 | 0.81 | 0.08 |
| | AdaBoostM1 | 0.68 | 0.54 | 0.14 |
| | DecisionTable | 0.88 | 0.74 | 0.14 |
| | JRip | 0.83 | 0.78 | 0.05 |
| | ZeroR | 0.53 | 0.21 | 0.32 |
| | J48 | 0.91 | 0.84 | 0.07 |
| | RandomForest | 0.90 | 0.82 | 0.08 |

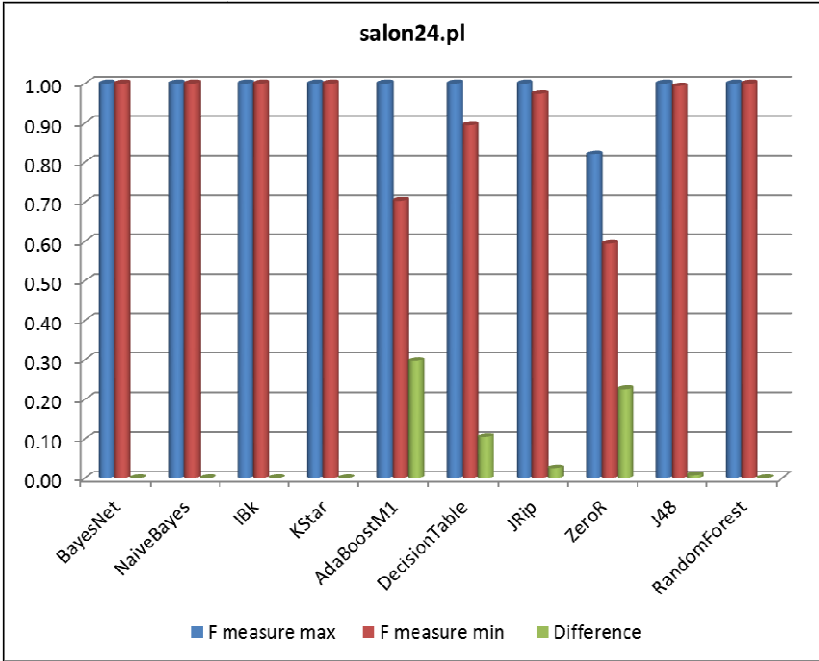


Fig. 2. The classifiers comparison for salon24.pl

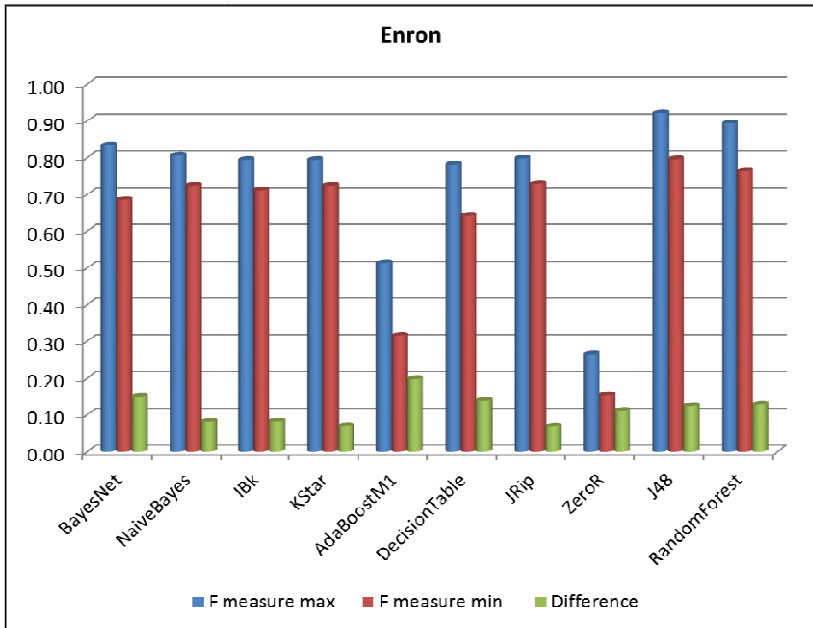


Fig. 3. The classifiers comparison for Enron

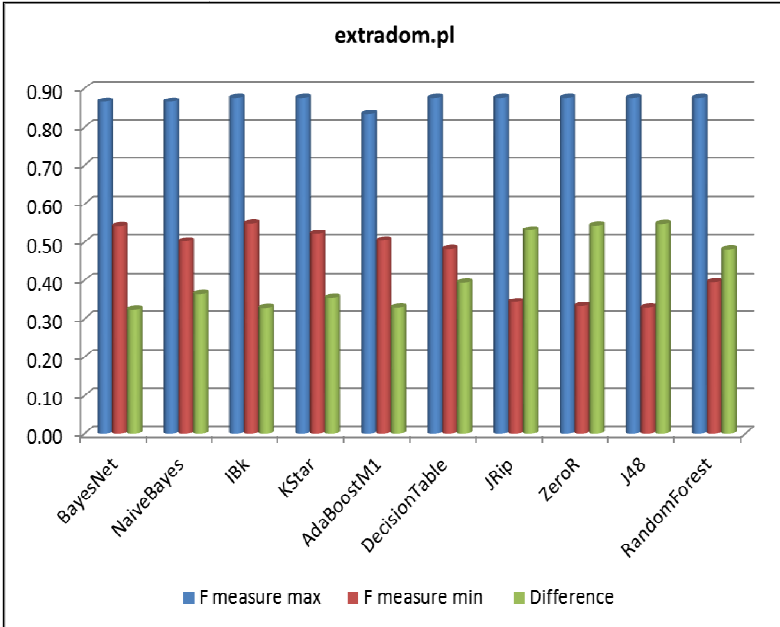


Fig. 4. The classifiers comparison for extradom.pl

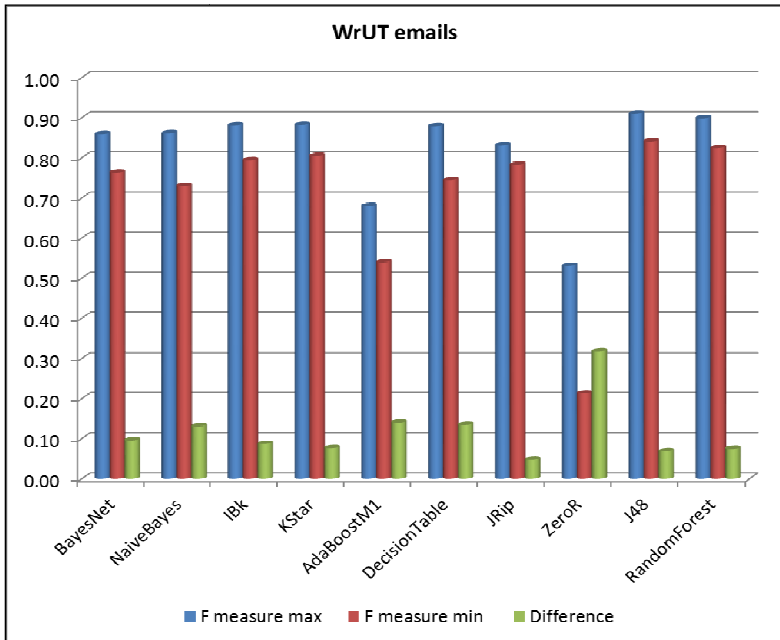


Fig. 5. The classifiers comparison for WrUT emails

Table 2 clearly indicates that for each dataset the best two classifiers are J48 (C4,5) decision trees and Random Forest ensemble of decision trees, thus, both classifiers were used for further analyses. Additionally, the results for these two classifiers are quite impressive since F measure for both of them is always around 0.8-0.9.

Now, it is necessary to analyse how the α and β parameters affect the classification. This was done for the WrUT dataset. The first analysis was for J48 and is presented in Table 3 and Figures 6, 7.

Table 3. The weighted average of F-measure measure (weighted by the contribution of the class–event in the dataset) for F48 decision tree for all six possible classes

| $\beta \backslash \alpha$ [%] | 50 | 60 | 70 | 80 | 90 | 100 |
|-------------------------------|-------|-------|-------|-------|-------|-------|
| 50 | 0.881 | 0.85 | 0.887 | 0.889 | 0.884 | 0.888 |
| 60 | 0.884 | 0.879 | 0.898 | 0.885 | 0.883 | 0.91 |
| 70 | 0.886 | 0.89 | 0.897 | 0.902 | 0.897 | 0.884 |
| 80 | 0.879 | 0.885 | 0.889 | 0.91 | 0.886 | 0.882 |
| 90 | 0.87 | 0.882 | 0.871 | 0.913 | 0.892 | 0.887 |
| 100 | 0.852 | 0.869 | 0.848 | 0.907 | 0.869 | 0.841 |

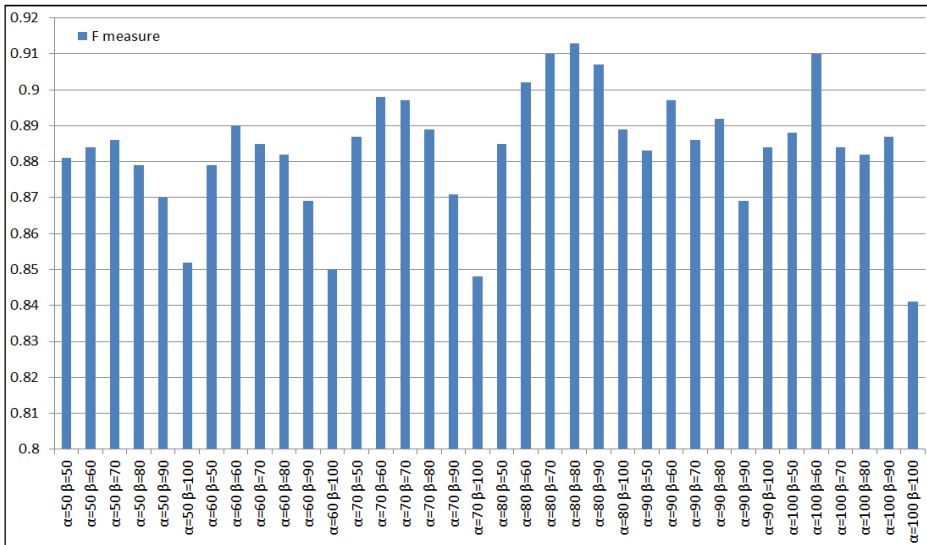


Fig. 6. F-measure values in relation to β and α

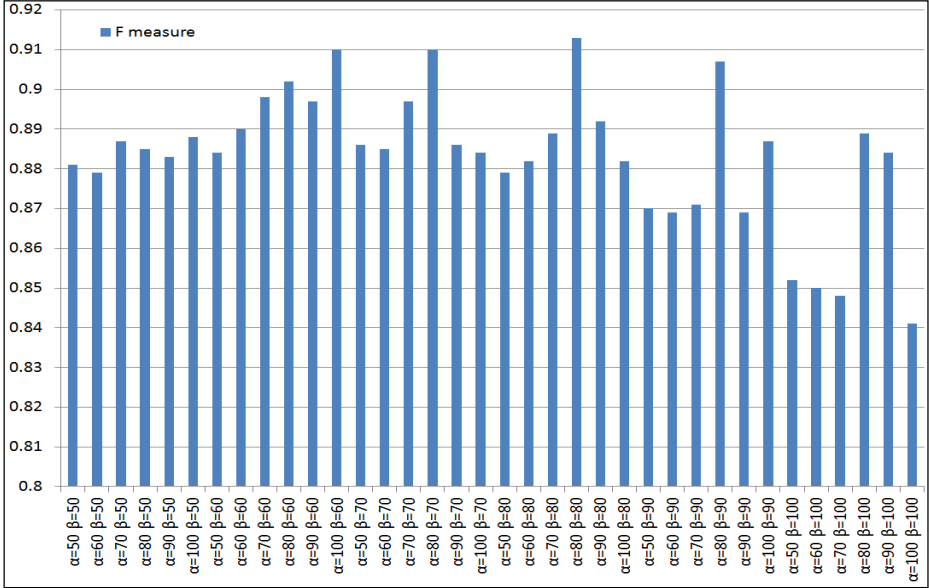


Fig. 7. F-measure values in relation to α and β

While analysing Figures 6 and 7 for the constant α , we can observe the best results are when β is around 80%. However, for the constant β , it is hard to see any regular pattern. In general, the highest F-measure is for $\alpha = 80\%$. So, if the J48 decision tree is used as a classifier, it is recommended to use $\alpha = 80\%$ and β from the set $\{70\%, 80\%, 90\%\}$ for the *GED* method parameters. The reason behind such a result can be quite simple. If we look at results presented in [3] we can see that the high α and β reduce the number of split and merge events. Thus, the number of those events is similar to the number of other events. On the other hand, for the low α and β the number of splits and merges overshadow the number of the other events. It means that value of about 80% appears to be the best with respect to classification quality evaluated by the F-measure.

Quite similar results were achieved by the Random Forest classifier. The parameter α can be from the set $\{80\%, 90\%, 100\%\}$ and β from $\{60\%, 70\%, 80\%, 90\%\}$. Hence, the conclusion is: the *GED* method with the high α and β produces better input features for classification, also if applied to the Random Forest classifier. The evaluation of α and β influence with the Random Forest classifier was presented in Table 4, Figure 8 and 9. Not like for J48 tree, for Random Forest tree a specific pattern can be found for both α and β . For the constant α the best results are if β is equal to 60%, 70% or 80, see Figure 5.8, and for the constant β the best results are when α is equal to 80%, 90% or 100%, see Figure 9.

Table 4. The weighted average of F measure for Random Forest tree for all six classes

| $\beta \backslash \alpha$ [%] | 50 | 60 | 70 | 80 | 90 | 100 |
|-------------------------------|-------|-------|-------|-------|-------|-------|
| 50 | 0.846 | 0.848 | 0.857 | 0.874 | 0.868 | 0.87 |
| 60 | 0.848 | 0.852 | 0.865 | 0.881 | 0.875 | 0.899 |
| 70 | 0.846 | 0.853 | 0.872 | 0.891 | 0.879 | 0.897 |
| 80 | 0.849 | 0.854 | 0.862 | 0.893 | 0.882 | 0.867 |
| 90 | 0.843 | 0.848 | 0.849 | 0.896 | 0.872 | 0.887 |
| 100 | 0.828 | 0.824 | 0.828 | 0.869 | 0.869 | 0.849 |

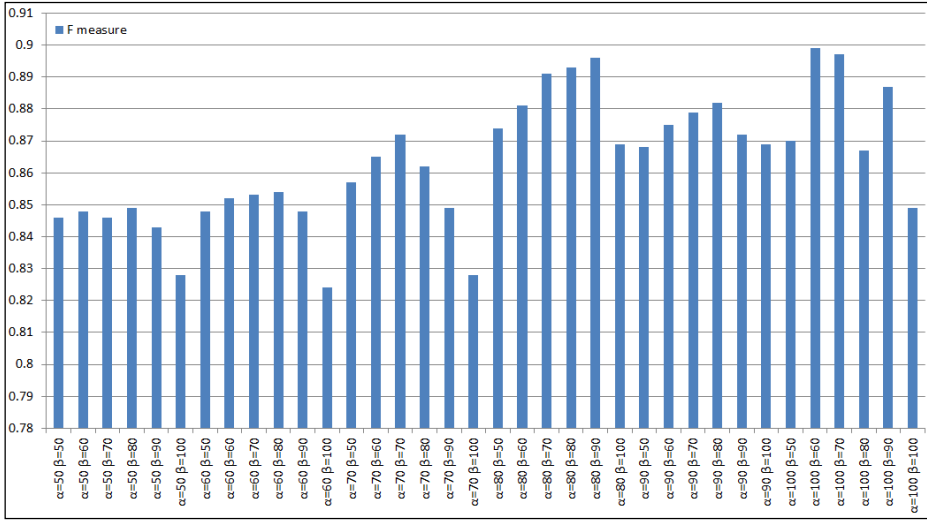


Fig. 8. F-measure values in relation to β and α

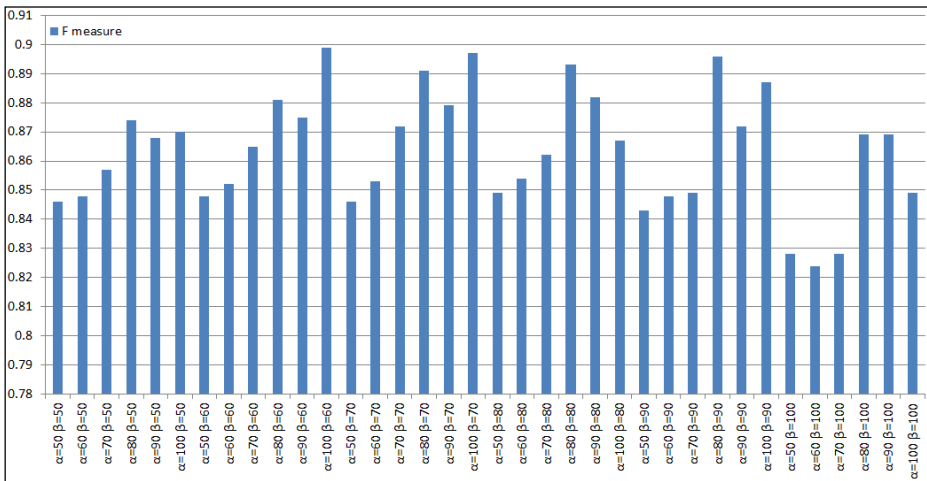


Fig. 9. F-measure values in relation to α and β

6 Conclusions and Future Work

It was shown that using a simple sequence which consists only of several preceding groups' sizes and events as an input for the classifier, the learnt model is able to produce very good results even for simple classifiers. It means that such prediction of group evolution can be very efficient in terms of prediction quality. The experimental analyses on six evolving social networks have revealed that decision trees and random forest as classifiers usually provide the most accurate results. Additionally, we can observe that the GED method used for change identification can be successfully used as a right indicator. However, its two parameters α and β significantly influence on the classification quality and the best results can be achieved for their values at the level of about 80%.

Of course, many questions remain unsolved, in particular:

- Are similar prediction results achievable for every kind of network?
- What would happen, if we use different classifiers or more advanced classification concepts like competence areas (clustering of groups and application of separate classifiers to each cluster)?
- What would be the influence of adding more input features (measures) describing the group like its diameter, average degree, percentage of network members which are in this group, the number of core members etc. as well as their various aggregations, e.g. average size for last 6 time frames?
- What would be the results, if we use shorter/longer sequences (more preceding events and group measures)?
- What would happen, if we use different node importance measure used in *GED*?

All above questions will be addressed in future research. This paper however, aimed only to present that predicting group evolution using the *GED* method with some common classifiers is both possible and effective.

References

1. Aha, D., Kibler, D.: Instance-based learning algorithms. *Machine Learning* 6, 37–66 (1991)
2. Breiman, L.: Random Forests. *Machine Learning* 45(1), 5–32 (2001)
3. Bródka, P., Saganowski, S., Kazienko, P.: GED: The Method for Group Evolution Discovery in Social Networks. *Social Network Analysis and Mining* (2012), Open Access, <http://www.springerlink.com/content/d6771886878t8p10>, doi:10.1007/s13278-012-0058-8
4. Cleary, J.G., Trigg, L.E.: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, pp. 108–114 (1995)
5. Cohen, W.W.: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, pp. 115–123 (1995)
6. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, San Francisco, pp. 148–156 (1996)

7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1) (2009)
8. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338–345 (1995)
9. Kairam, S., Wang, D., Leskovec, J.: The life and death of online groups: predicting group growth and longevity. In: *WSDM 2012*, pp. 673–682. ACM (2012), doi:10.1145/2124295.2124374
10. Kazienko, P., Ruta, D., Bródka, P.: The Impact of Customer Churn on Social Value Dynamics. *International Journal of Virtual Communities and Social Networking* 1(3), 60–72 (2009)
11. Kohavi, R.: The Power of Decision Tables. In: Lavrač, N., Wrobel, S. (eds.) *ECML 1995*. LNCS, vol. 912, pp. 174–189. Springer, Heidelberg (1995)
12. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: *KDD 2008*, pp. 462–470. ACM (2008)
13. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci.* 58, 1019–1031 (2007), doi:10.1002/asi.20591
14. McLachlan, G.J., Do, K.A., Ambrose, C.: *Analyzing Microarray Gene Expression Data*. Wiley Series in Probability and Statistics (2004) ISBN-10: 0471226165
15. Michalski, R., Palus, S., Bródka, P., Kazienko, P., Juszczyszyn, K.: Modelling Social Network Evolution. In: Datta, A., Shulman, S., Zheng, B., Lin, S.-D., Sun, A., Lim, E.-P. (eds.) *SocInfo 2011*. LNCS, vol. 6984, pp. 283–286. Springer, Heidelberg (2011)
16. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818 (2005)
17. Quinlan, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo (1993)
18. Singh, L., Getoor, L.: Increasing the Predictive Power of Affiliation Networks. *IEEE Data Eng. Bull (DEBU)* 30(2), 41–50 (2007)
19. Wai-Ho, A., Chan, K.C.C., Xin, Y.: A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation* 7(6), 532–545 (2003)
20. Zheleva, E., Getoor, L., Golbeck, J., Kuter, U.: Using Friendship Ties and Family Circles for Link Prediction. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) *SNAKDD 2008*. LNCS, vol. 5498, pp. 97–113. Springer, Heidelberg (2010)
21. Bródka, P.: *Key Users in Social Network. How to find them?* LAP Lambert Academic Publishing (2012) ISBN-13: 978-3-659-19597-6, ISBN-10: 3659195979

Interpolating between Random Walks and Shortest Paths: A Path Functional Approach

François Bavaud and Guillaume Guex*

Department of Computer Science and Mathematical Methods
Department of Geography
University of Lausanne, Switzerland

Abstract. General models of network navigation must contain a deterministic or drift component, encouraging the agent to follow routes of least cost, as well as a random or diffusive component, enabling free wandering. This paper proposes a thermodynamic formalism involving two path functionals, namely an energy functional governing the drift and an entropy functional governing the diffusion. A freely adjustable parameter, the temperature, arbitrates between the conflicting objectives of minimising travel costs and maximising spatial exploration. The theory is illustrated on various graphs and various temperatures. The resulting optimal paths, together with presumably new associated edges and nodes centrality indices, are analytically and numerically investigated.

1 Introduction

Consider a network together with an agent wishing to move (or wishing to move goods, money, information, etc.) from source node s to target node t . The agent seeks to minimise the total cost or duration of the move, but the ideal path may be difficult to realise exactly, in absence of perfect information about the network.

The above context is common to many behavioral and decision contexts, among which “small-world” social communications (Travers and Milgram 1969), spatial navigation (e.g. Farnsworth and Beecham 1999), routing strategy on internet networks (e.g. Zhou 2008, Dubois-Ferrière et al. 2011), and several others (e.g. Borgatti 2005; Newman 2005).

Trajectories can be coded, generally non-univocally, by $X = (x_{ij})$ where $x_{ij} =$ “number of direct transitions from node i to node j ”. The use of the flow matrix X is central in Operational Research (e.g. Ahuja et al. 1993) and Markov Chains theory (e.g. Kemeny and Snell 1976); four optimal st -paths have in particular been extensively analysed *separately* in the litterature, namely the shortest-path, the random walk, the maximum flow (Freeman et al. 1991) and the electrical current (Kirchhoff 1850; Newman 2005; Brandes and Fleischer 2005).

This paper investigates the properties of st -paths resulting from the minimisation of a *free energy functional* $F(X)$, over the set $X \in \mathcal{X}$ of admissible

* The specific remarks of two anonymous reviewers are gratefully acknowledged.

solutions. $F(X)$ contains a resistance component privileging shortest paths, and an entropy component favouring random walks. The conflict is arbitrated by a continuous parameter $T \geq 0$, the *temperature* (or its *inverse* $\beta := 1/T$), and results in an analytically solvable unique optimum *continuously interpolating* between shortest-paths and random walks. See Yen et al. (2008) and Saerens and al. (2009) for a close proposal, yet distinct in its implementation.

Section 2 introduces the formalism, in particular the *energy functional* (based upon an edge resistance matrix R , symmetrical or not) and the *entropy functional* (based upon a Markov transition matrix W , reversible or not, related to R or not). Section 2.5 provides the analytic form of the unique solution minimising the free energy. Section 4 proposes the definition of edge and vertex betweenness centrality indices directly based upon the flow X . They are illustrated in sections 3 and 5 for various network geometries at various temperatures.

2 Definitions and Solutions

2.1 Admissible Paths

Consider a connected graph $G = (V, E)$ involving $n = |V|$ nodes together with two distinguished and distinct nodes, the source s and target t . The st -path or flow matrix, noted $X^{st} = (x_{ij}^{st})$ or simply $X = (x_{ij})$, counts the number of transitions from i to j along conserved unit paths starting at s , possibly visiting s again, and absorbed at t . Hence

$$x_{ij} \geq 0 \quad \text{positivity} \quad (1)$$

$$x_{i\bullet} - x_{\bullet i} = \delta_{is} - \delta_{it} \quad \text{unit flow conservation} \quad (2)$$

where δ_{ij} is the Kronecker delta, the components of the identity matrix. Here and in the sequel, \bullet denotes the summation over the values of the replaced index, as in $x_{i\bullet} = \sum_{j=1}^n x_{ij}$. In particular, $x_{s\bullet} = x_{\bullet s} + 1$. Also,

$$x_{t\bullet} = 0 \quad \text{absorbtion at } t \quad (3)$$

entailing $x_{tj} = 0$ for all j , and $x_{\bullet t} = 1$. Normalisation (2) can be extended to *valued flows*

$$x_{i\bullet} - x_{\bullet i} = v(\delta_{is} - \delta_{it}) \quad \text{conservation for valued flow} \quad (4)$$

where $v \geq 0$, the amount sent through the network, is the *value* of the flow. Further familiar constraints consist of

$$x_{ij} \leq c_{ij} \quad \text{capacity, where } c_{ij} \geq 0 \quad (5)$$

$$x_{ij} \geq b_{ij} \quad \text{minimum flow requirement, } b_{ij} \geq 0 \quad (6)$$

$$x_{\bullet j_0} = 0 \quad \text{forbidden node } j_0 \quad (7)$$

$$x_{i_0 j_0} = 0 \quad \text{forbidden arc } (i_0 j_0) \quad (8)$$

2.2 Mixtures and Convexity

Any of the above constraints (1) to (8) or combinations thereof defines a *convex* set \mathcal{X} of admissible *st*-paths: if X and Y are admissible, so is their *mixture* $\alpha X + (1 - \alpha)Y$ for $\alpha \in [0, 1]$. Mixture of paths are generally non-integer, and can be given a probabilistic interpretation, as in

- $x_{\bullet\bullet}$ = “average time (number of transitions) for transportation from s to t ”
- $x_{ij}/x_{i\bullet}$ = “conditional probability to jump to j from i ”.

From now on, one considers by default unit flows X , generally non-integer, obeying (1), (2) and (3).

2.3 Path Entropy and Energy

Let $W = (w_{ij})$ denote the $(n \times n)$ transition matrix of some irreducible Markov chain. A *st*-path constitutes a random walk (as defined by W) iff $x_{ij}/x_{i\bullet} = w_{ij}$ for all visited node i , i.e. such that $x_{i\bullet} > 0$. Random walk *st*-paths X minimise the *entropy* functional

$$G(X) := \sum_{ij} x_{ij} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} = \sum_i x_{i\bullet} K_i(X||W) = x_{\bullet\bullet} \sum_i \frac{x_{i\bullet}}{x_{\bullet\bullet}} K_i(X||W)$$

where $K_i(X||W) := \sum_j \frac{x_{ij}}{x_{i\bullet}} \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} \geq 0$ is the Kullback-Leibler divergence between the transition distributions X and W from i , taking on its minimum value zero iff $\frac{x_{ij}}{x_{i\bullet}} = w_{ij}$. Note $G(X)$ to be *homogeneous*, that is $G(vX) = vG(X)$ for $v > 0$, reflecting the *extensivity* of $G(X)$ in the thermodynamic sense.

By contrast, shortest-paths and other alternative optimal paths minimize *resistance* or *energy* functionals of the general form

$$U(X) := \sum_{ij} r_{ij} \varphi(x_{ij})$$

where $r_{ij} > 0$ represent a cost or resistance associated to the directed arc ij , and $\varphi(x)$ is a smooth non-decreasing function with $\varphi(0) = 0$. In particular, minimizing $U(X)$ yields

- *st*-shortest paths for the choice $\varphi(x) = x$, where r_{ij} is the length of the arc ij
- *st*-electric currents from s to t for the choice $\varphi(x) = x^2/2$, where r_{ij} is the resistance of the conductor ij (see section 2.8).

As in Statistical Mechanics, we consider in this paper the class of admissible paths minimizing the *free energy*

$$F(X) := U(X) + T G(X) . \tag{9}$$

Here $T > 0$ is a free parameter, the *temperature*, controlling for the importance of the fluctuation around the trajectory of least resistance or energy (ground

sate), realised in the low temperature limit $T \rightarrow 0$. In the high temperature limit $T \rightarrow \infty$ (or $\beta \rightarrow 0$, where $\beta := 1/T$ is the *inverse temperature*), the path consists of a random walk from s to t governed by W . Hence, minimising the free energy (9) generates for $T > 0$ “heated extensions” of classical minimum-cost problems $\min_X U(X)$, with the production of random fluctuations around the classical, “ground state” solution.

Derivating the free energy with respect to x_{ij} , and expressing the conservation constraints (2) through Lagrange multipliers $\{\lambda_i\}$ yields the optimality condition

$$T \ln \frac{x_{ij}}{x_{i\bullet} w_{ij}} + r_{ij} \varphi'(x_{ij}) = \lambda_j - \lambda_i \quad (10)$$

that is

$$x_{ij} = x_{i\bullet} w_{ij} \exp(-\beta[r_{ij} \varphi'(x_{ij}) + \lambda_i - \lambda_j]) . \quad (11)$$

The multipliers are defined up to an additive constant (see 15). In any case, $x_{ij} = 0$ when $w_{ij} = 0$ or $i = t$.

2.4 Minimum Free Energy and Uniqueness

Multiplying (10) by x_{ij} and summing over all arcs yields an identity involving the entropy $G(X)$ of the optimal path X . Substitution in the free energy together with (2) demonstrates in turn the identity

$$\min_X F(X) = \sum_{ij} r_{ij} [\varphi(x_{ij}) - \varphi'(x_{ij}) x_{ij}] + \lambda_t - \lambda_s . \quad (12)$$

The first term is negative for $\varphi(x)$ convex, positive for $\varphi(x)$ concave, and zero for the heated shortest-path problem $\varphi(x) = x$, for which $\min_X F(X) = \lambda_t - \lambda_s$.

Also, the entropy functional is convex, that is $G(\alpha X + (1 - \alpha)Y) \leq \alpha G(X) + (1 - \alpha)G(Y)$ for two admissible paths X and Y and $0 \leq \alpha \leq 1$. The energy $U(X)$ is convex (resp. concave) iff $\varphi(x)$ is convex (resp. concave).

When a strictly convex functional $F(X)$ possesses a local minimum on a convex domain \mathcal{X} , the minimum is unique. In particular, we expect the optimal flows for $\varphi(x) = x^p$ to be unique for $p > 1$, but not anymore for $0 < p < 1$, where local minima may exist; see Alamgir and von Luxburg (2011) on “ p -resistances”.

In the shortest-path problem $p = 1$, the solution is unique if $T > 0$ (Section 2.5); when $T = 0$, local minima of $U(X)$ may coexist, yet all yielding the same value of $U(X)$.

2.5 Algebraic Solution

Solving (11) is best done by considering separately the target node t . Define $v_{ij} := w_{ij} \exp(-\beta r_{ij} \varphi'(x_{ij}))$ as well as the $(n - 1) \times (n - 1)$ matrix $V = (v_{ij})_{i,j \neq t}$. Also, define the $(n - 1)$ dimensional vectors

$$\begin{aligned} a_i &:= x_{i\bullet} \exp(-\beta \lambda_i)|_{i \neq t} & b_j &:= \exp(\beta \lambda_j)|_{j \neq t} \\ q_i &:= v_{it}|_{i \neq t} & e_j &:= \delta_{js}|_{j \neq t} \end{aligned} \quad (13)$$

Summing (11) over all i (for $j \neq t$, resp. $j = t$), then over all j for $i \neq t$ yields, using (2) and (3)

$$V'a = a - \exp(-\beta\lambda_s) e \quad a'q = \exp(-\beta\lambda_t) \quad Vb + \exp(\beta\lambda_t) q = b$$

Define the $(n-1) \times (n-1)$ matrix $M = (m_{ij})$ and the $(n-1)$ vector z as

$$M := (I - V)^{-1} = I + V + V^2 \dots \quad z := Mq \quad (14)$$

Then a and b express as

$$a_i = \exp(-\beta\lambda_s) m_{si} \quad b_j = \exp(\beta\lambda_s) \frac{z_j}{z_s} = \exp(\beta\lambda_j)$$

implying incidentally

$$\lambda_j = T \ln z_j + C \stackrel{\text{(Section 2.6)}}{=} T \ln z_j + \lambda_t . \quad (15)$$

Finally

$$x_{i\bullet} = m_{si} \frac{z_i}{z_s} \quad x_{ij} = m_{si} v_{ij} \frac{z_j}{z_s} \quad (i \neq t) \quad (16)$$

$$x_{it} = m_{si} \frac{q_i}{z_s} \quad x_{\bullet\bullet} = \frac{(Mz)_s}{z_s} = \frac{(M^2q)_s}{(Mq)_s} . \quad (17)$$

In general, V , M , q and d depend upon X . Hence (16) and (17) define a recursive system, whose fixed points may be multiple if $U(X)$ is not convex (Section 2.4), but converging to a unique solution for $p > 1$.

In the heated shortest-path case $p = 1$, the above quantities are independent of X . Hence the solution is unique, and particularly easy to compute in one single $O(n^3)$ step, involving matrix inversion, as illustrated in Sections 3 and 5.

2.6 Probabilistic Interpretation

In addition to the absorbing target node t , let us introduce another ‘‘cemetery’’ or absorbing state 0, and define an extended Markov chain P on $n+1$ states with transition matrix

$$P = \left(\begin{array}{c|cc|c} & \mathbf{i \neq t, 0} & \mathbf{t} & \mathbf{0} \\ \hline \mathbf{i \neq t, 0} & V & q & \rho \\ \hline \mathbf{t} & 0 & 1 & 0 \\ \hline \mathbf{0} & 0 & 0 & 1 \end{array} \right)$$

where $\rho_i = 1 - \sum_{k=1}^n v_{ik}$ is the probability of being absorbed at 0 from i in one step.

$M = (m_{ij})$ is the so-called *fundamental matrix* (see (14) and Kemeny and Snell 1976 p.46), whose components m_{ij} give the *expected number of visits from i to j* , before being eventually absorbed at 0 or t . Also, z_i (with $i \neq t, 0$) is the *survival probability*, that is to be, directly or indirectly, eventually absorbed at

t rather than killed at 0, when starting from i . The higher the node survival probability, the higher the value of its Lagrange multiplier in view of (15).

Extending the latter to $j = t$ entails the consistency condition $z_t = 1$, making $\lambda_t \geq \lambda_i$ for all i . In particular, the free energy of the heated shortest-path case is, in view of (12),

$$F(X^{st}) = -T \ln z_s(T) ,$$

increasing (super-linearly in T) with the risk of being absorbed at 0 from s .

2.7 High-Temperature Limit

The energy term in (9) plays no role anymore in the limit $T \rightarrow \infty$ (that is $\beta \rightarrow 0$), and so does the absorbing state 0 above in view of $\rho_i = 0$. In particular, $z_i \equiv 1$ and $x_{ij}^{st} = m_{si} w_{ij}$ for $i \neq t$.

Also, $x_{\bullet\bullet}^{st}$ is the expected number of transitions needed to reach t from s . The *commute time distance* or *resistance distance* $x_{\bullet\bullet}^{st} + x_{\bullet\bullet}^{ts}$ is known to represent a *squared Euclidean distance* between states s and t : see e.g. Fouss et al. 2007, and references therein; see also Yen et al. (2008) and Chebotarev (2010) for further studies on resistance and shortest-path *distances*.

2.8 Low-Temperature Limit

Equations (11), (16) and (17) show the positivity condition $x_{ij} \geq 0$ to be automatically satisfied, thanks to the entropy term $G(X)$. However, the latter disappears in the limit $T \rightarrow 0$, where one faces the difficulty that the optimality condition (10) $r_{ij} \varphi'(x_{ij}) = \lambda_j - \lambda_i$ is still justified only if x_{ij} is freely adjustable, that is if $x_{ij} > 0$.

For the st -shortest path problem $\varphi(x) = x$, one gets, assuming the solution to be unique, the well-known characterisation (see e.g. Ahuja et al. (1993) p.107):

$$\begin{cases} r_{ij} = \lambda_j - \lambda_i & \text{if } x_{ij} > 0 \\ r_{ij} > \lambda_j - \lambda_i & \text{if } x_{ij} = 0 \end{cases}$$

occurring in the dual formulation of the st -shortest path problem, namely “*maximize* $\lambda_t - \lambda_s$ *subject to* $\lambda_j - \lambda_i \leq r_{ij}$ *for all* i, j ”. Here λ_i is the shortest-path distance from s to i .

For the st -electrical circuit problem $\varphi(x) = x^2/2$, one gets $r_{ij} x_{ij} = \lambda_j - \lambda_i$ if $x_{ij} > 0$, in which case $x_{ji} > 0$ cannot hold in view of the positivity of the resistances, thus forcing $x_{ji} = 0$. Hence

$$\begin{cases} x_{ij} = \frac{\lambda_j - \lambda_i}{r_{ij}} > 0 & \text{if } \lambda_j > \lambda_i \\ x_{ij} = 0 & \text{otherwise} \end{cases}$$

expressing *Ohm's law* for the current intensity x_{ij} (Kirchhoff 1850), where λ_i is the electric potential at node i .

3 Illustrations and Case Studies: Simple Flow and Net Flow

Let us restrict on st -shortest path problems, i.e. $\varphi(x) = x$, whose free energy is homogeneous in the sense $F(vX) = vF(X)$ where $v > 0$ is the value of the flow in (4).

Graphs are defined by a $n \times n$ Markov transition matrix W together with a $n \times n$ positive resistance matrix R . Fixing in addition s, t and β , yields an unique *simple flow* x_{ij}^{st} , computable for any W (reversible or not) and any R (symmetric or not) - a fairly large set of tractable weighted networks.

An obvious class of networks consists of binary graphs, defined by a symmetric, off-diagonal adjacency matrix, with unit resistances and uniform transitions on existing edges (i.e. a simple random walk in the sense of Bollobás 1998).

Such are the graphs A (Figure 1) and B (Figure 2) below. Graph C (Figure 3) penalises in addition two edges forming short-cut from the point of view of W , but with increased values of their resistance.

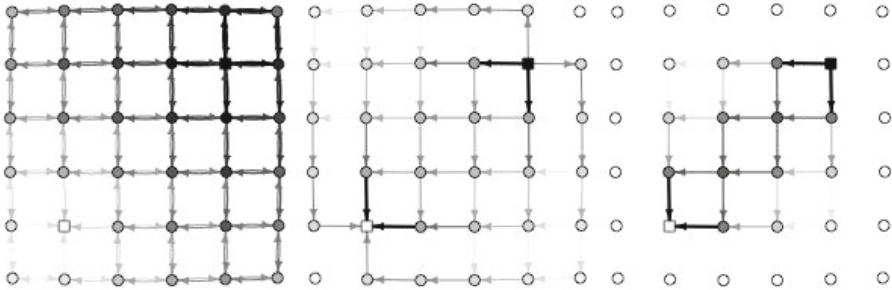


Fig. 1. Graph A is a square grid with uniform transitions and resistances. The resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$ (shortest-path dominance). Note the simple flow and net flow to be identical at low temperatures.

Among the wide variety of graphs defined by a (W, R) pair, the plain graphs A, B and C primarily aim at illustrating the basic fact that, at high temperature, reverberation among neighbours of the source may dramatically lengthen the shortest path - an expected phenomenon (Figure 4).

Another quantity of interest is the *net flow*

$$\nu_{ij}^{st} := |x_{ij}^{st} - x_{ji}^{st}| \quad (18)$$

discounting “back and forth walks” inside the same edge, as discussed by Newman (2005): as a matter of fact, the presence of such alternate moves mechanically increases the simple flow inside an edge or node, especially near the source

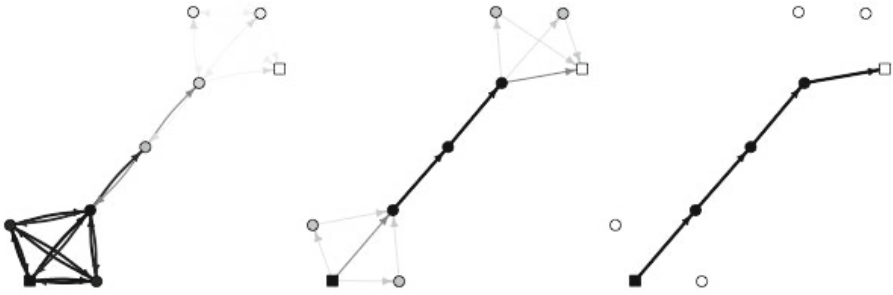


Fig. 2. Graph B consists of two cliques K_4 joined by two edges, with uniform transitions and resistances. Again, the resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$.

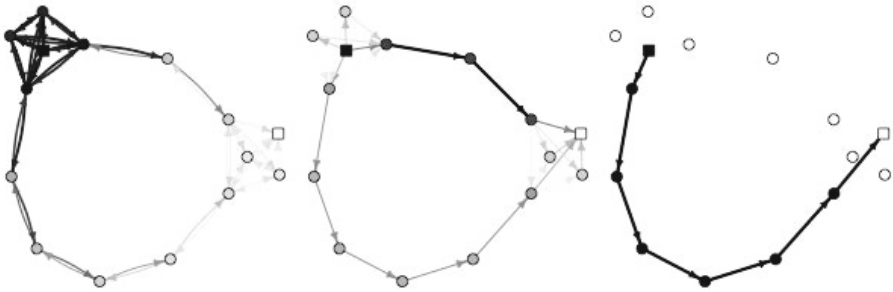


Fig. 3. Graph C consists of two cliques K_5 joined by two paths: the upper one consists of five edges, each with unit resistance, while the upper one contains two edges, each with resistance tenfold larger. The resulting (high values in black, low values in light grey) simple flow x_{ij}^{st} and net flow ν_{ij}^{st} from s (black square) to t (white square) are depicted respectively on the left and middle picture with $\beta = 0$ (random walk) and on the right with $\beta = 50$.

at high temperature (Figures 1, 2 and 3, left), giving the false impression the behaviour is more entropic (that is, random-walk dominated) around the source, which is erroneous.

The net flow “filters out” reverberations and hence captures the resulting “trend” of the agents within their random movements, who rarely go back along the edge from where they came if there is another way; cf. the circulation of “used goods” as defined in Borgatti (2005) along trails exempt of edges repetition. At low temperatures, the simple flow is directed in one way and hence converges to the simple flow (Figures 1, 2 and 3, right).

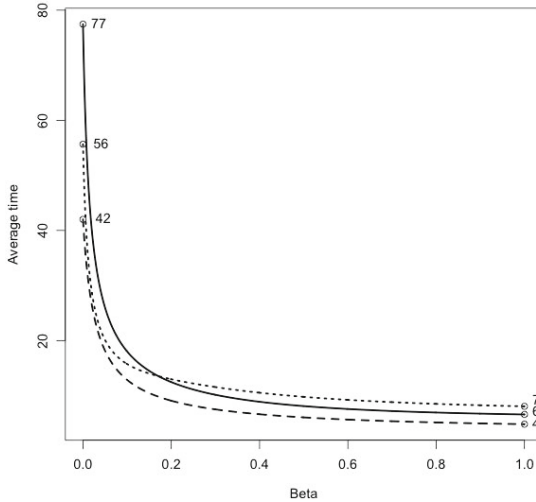


Fig. 4. The average time $x_{\bullet\bullet}^{st}$ to reach t from s is minimum for $T = 0$, and decreases with the inverse temperature β . Solid line: graph A ; Dashed line: graph B ; Dotted line: graph C .

4 Edge and Vertex Centrality Betweenness

Several flow-based indices of betweenness centrality have been proposed ever since the shortest-path centrality pioneering proposal of Freeman (1977). In particular, random-walk centrality indices have been discussed by Noh and Rieger (2004) and Newman (2005). In this paper, we study the (unweighted) *mean flow betweenness*, defined for edges and vertices respectively (with complexity $O(n^5)$) as

$$\langle x_{ij} \rangle := \frac{1}{n(n-1)} \sum_{s,t|s \neq t} x_{ij}^{st} \quad \langle x_{i\bullet} \rangle := \sum_j \langle x_{ij} \rangle = \langle x_{\bullet i} \rangle \quad (19)$$

where the latter identity results from the conservation condition (2). Definition (19) is intuitive enough: an edge is central if it carries a large amount of flow *on average*, that is by considering *all pairs of distinct source-targets couples*, thus extending the formalism to flows without specific source or target, such as monetary flows.

A more formal motivation arises from sensitivity analysis, with the result

$$\frac{\partial F(X(R))}{\partial r_{ij}} = \sum_{kl} \frac{\partial F(X(R))}{\partial x_{kl}(R)} \frac{\partial x_{kl}(R)}{\partial r_{ij}} + x_{ij}(R) = x_{ij}$$

where $F(X(R)) = \sum_{ij} r_{ij} x_{ij}(R) + TG(X(R))$ is the minimum free energy (9) under the constraints of Section 2.1 and r_{ij} the resistance of the edge ij .

Note that $\langle x_{\bullet\bullet} \rangle := \sum_j \langle x_{\bullet j} \rangle$ represents the average time to go from a vertex s to another vertex t and to return to s , averaged over all distinct pairs st . One can also define the *relative mean flow betweenness* as

$$c_{ij} := \frac{\langle x_{ij} \rangle}{\langle x_{\bullet\bullet} \rangle} \qquad c_i := \frac{\langle x_{i\bullet} \rangle}{\langle x_{\bullet\bullet} \rangle}$$

with the property $c_{ij} \geq 0$, $\sum_{ij} c_{ij} = 1$ and $c_i = c_{i\bullet} = c_{\bullet i}$.

Another candidate for a flow-based betweenness index is the *mean net flow*, again defined for edges and vertices as

$$\langle \nu_{ij} \rangle := \frac{1}{n(n-1)} \sum_{s,t|s \neq t} \nu_{ij}^{st} \qquad \langle \nu_{i\bullet} \rangle := \sum_j \langle \nu_{ij} \rangle = \langle \nu_{\bullet i} \rangle \quad (20)$$

Middle pictures in Figures 5, 6 and 7 below demonstrate how the mean net flow “subtracts” the mechanical contribution arising from back and forth walks inside the same edge, in better accordance to a common sense notion of centrality.

Also, the sensitivity of the trip duration with respect to the edge resistance

$$\sigma_{ij} := \frac{\partial \langle x_{\bullet\bullet}(R) \rangle}{\partial r_{ij}}$$

constitutes yet another candidate, amenable to analytic treatment, whose study is beyond the size of the paper.

5 Case Studies (Continued): Mean Flow and Mean Net Flow

Figures 5, 6 and 7 depict the mean flow betweenness and the mean net flow betweenness (19) for the three graphs of Section 3, at high temperatures (left and middle) and low temperatures (right). Here $\langle x_{ij} \rangle = \langle x_{ji} \rangle$ due to the symmetry of R and the reversibility of W . Visual inspection confirms the role of the mean flow as a betweenness index, approaching the shortest-path betweenness at low temperatures.

At high temperatures, the mean flow $\langle x_{ij} \rangle$ turns out to be *constant* for all edges ij , a consistent observation for all “random-walk type” networks we have examined so far. As a consequence, the mean flow centrality of a node $\langle x_{i\bullet} \rangle$ is *proportional to its degree* for $\beta \rightarrow 0$, and identical to the shortest-path betweenness for $\beta \rightarrow \infty$. The former simply measures the local connectivity of the node, while the latter also takes into account the contributions of the remote parts of the network, in particular penalising high-resistance edges in comparison to low-resistance ones (Figure 7).

At low temperature, the net mean flow converges (together with the simple flow) to the shortest-path betweenness (Figures 5, 6 and 7, right). At high temperatures, the net mean flow betweenness is large for edges connecting clusters,

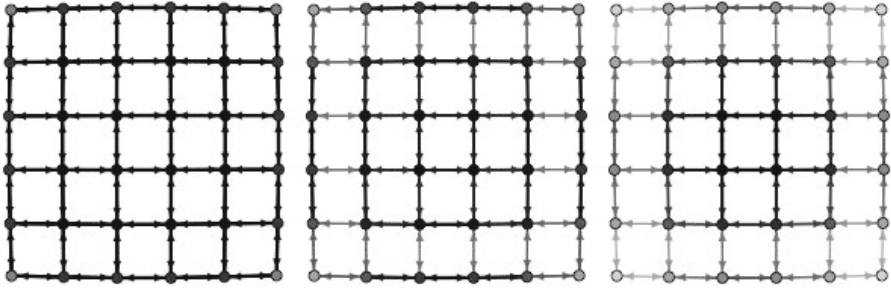


Fig. 5. Graph A: mean flow $\langle x_{ij} \rangle$ and mean net flow $\langle \nu_{ij} \rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey

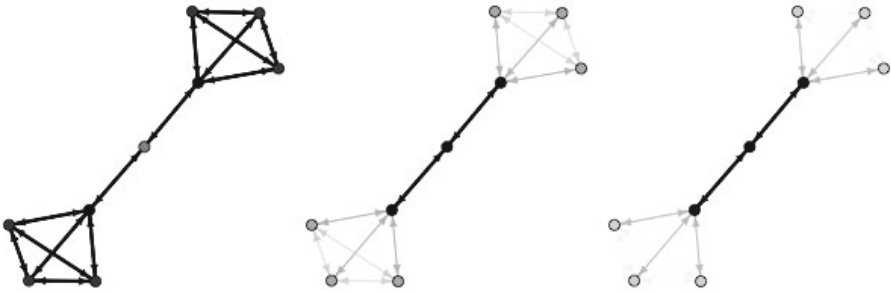


Fig. 6. Graph B: mean flow $\langle x_{ij} \rangle$ and mean net flow $\langle \nu_{ij} \rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey



Fig. 7. Graph C: mean flow $\langle x_{ij} \rangle$ and mean net flow $\langle \nu_{ij} \rangle$, with $\beta = 0$ (left and middle) and $\beta = 50$ (right); high values in black, low values in light grey

but, as expected, small for edges inside clusters. Hence an original kind of centrality, the “net random walk betweenness”, differing from shortest-path and degree betweenness, can be identified (Figures 5, 6 and 7, middle). As suggested in Figure 8 (right), contributions of both origins manifest themselves in the mean flow node centrality, for intermediate values of the temperature.

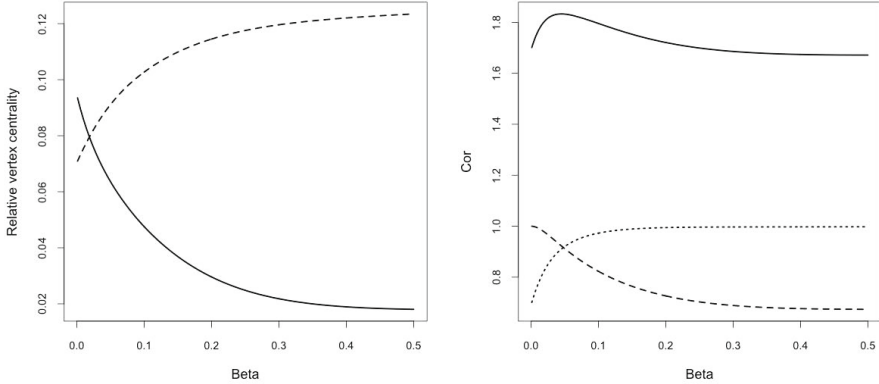


Fig. 8. Left: mean net flow centrality for the vertex in the “high-resistance path” (solid line) of network C, and for one of the nodes in the “low-resistance path” (dashed line) of network C. Right: inter-nodes correlation between the mean net flow centrality with itself at $\beta = 0$ (net random walk centrality; dashed line) and at $\beta = \infty$ (shortest-path node centrality; dotted line), in function of the inverse temperature β , for graph C. The sum of the two lines (solid line) is maximum for $\beta = 0.04$, arguably indicating a transition between an high- and a low-temperature regime.

6 Conclusion

The paper proposes a coherent mechanism, easy to implement, interpolating between shortest paths and random walks. The construction is controlled by a temperature T and applies to any network endowed with a Markov transition matrix W and a resistance matrix R . The two matrices can be related, typically as (componentwise) inverses of each other (e.g. Yen et al. 2008) *or not*, in which case continuity at $T = 0$ and $T = \infty$ however requires $w_{ij} > 0$ whenever $r_{ij} < \infty$.

Modelling empirical *st*-paths necessitates to define W and R . The “simple symmetric model”, namely unit resistances and uniform transitions on existing edges (Section 3) is, arguably, already meaningful in social phenomena and otherwise. For more elaborated applications, one can consider a possible model of tourist paths exploring Kobe (Iryio et al. 2012), consisting in choosing street directions as W with a bias towards “pleasant” street segments identified by low entries in R . Or the situation where a person at s wishes to be introduced to another person at t , by moving over an existing social network (defined by W) of friends, friends of friends, etc., where the resistance r_{ij} can express the difficulty that actor i introduces the person to actor j . One can also consider general situations where W expresses an average motion, mass circulation, and R captures an individual specific shift, biased towards preferentially reaching a peculiar outcome t , such as a specific location, or an a-spatial goal such as fortune, power, marriage, safety, etc.

By contrast, the construction seems little adapted to the simulation of replicant agents (such as viruses, gossip or e-mails) violating in general the flow conservation condition (2).

The paper has defined and investigated a variety of centrality indices for edges and nodes. In particular, the mean flow betweenness interpolates between degree centrality and shortest-path centrality for nodes. Regarding edges, the mean net flow embodies various measures ranging from simple random-walk betweenness (as defined in Newman 2005) to shortest-path betweenness, again. The average time needed to attain another node, respectively being attained from another node

$$T_s^{\text{out}} := \frac{1}{n-1} \sum_{t \mid t \neq s} x_{\bullet\bullet}^{st} \qquad T_t^{\text{in}} := \frac{1}{n-1} \sum_{s \mid s \neq t} x_{\bullet\bullet}^{st}$$

constitute alternative centrality indices, generalising Freeman's *closeness centrality* (Freeman 1979), incorporating a drift component when $T > 0$.

Maximum-likelihood type arguments, necessitating a probabilistic framework not exposed here, suggest for W and R fixed the estimation rule for T

$$U(X^{st}) = U(X^{st}(T))$$

where U is the energy functional in Section 2.3. Here X^{st} is the observed, empirical path, and $X^{st}(T)$ is the optimal path (16, 17) at temperature T . Alternatively, T could be calibrated from the observed total time, using Figure 4 as an abacus.

References

- Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network Flows. Theory, algorithms and applications. Prentice Hall (1993)
- Alamgir, M., von Luxburg, U.: Phase transition in the family of p-resistances. In: Neural Information Processing Systems (NIPS 2011), pp. 379–387 (2011)
- Bollobás, B.: Modern Graph Theory. Springer (1998)
- Borgatti, S.P.: Centrality and network flow. Social Networks 27, 55–71 (2005)
- Brandes, U., Fleischer, D.: Centrality Measures Based on Current Flow. In: Diekert, V., Durand, B. (eds.) STACS 2005. LNCS, vol. 3404, pp. 533–544. Springer, Heidelberg (2005)
- Chebotarev, P.: A class of graph-geodetic distances generalizing the shortest-path and the resistance distances. Discrete Applied Mathematics 159, 295–302 (2010)
- Dubois-Ferrière, H., Grossglauser, M., Vetterli, M.: Valuable Detours: Least-Cost Any-path Routing. IEEE/ACM Transactions on Networking 19, 333–346 (2011)
- Iryo, T., Shintaku, H., Senoo, S.: Experimental Study of Spatial Searching Behaviour of Travellers in Pedestrian Networks. In: 1st European Symposium on Quantitative Methods in Transportation Systems, EPFL Lausanne (2012) (contributed talk)
- Kemeny, J.G., Snell, J.L.: Finite Markov Chains. Springer (1976)
- Farnsworth, K.D., Beecham, J.A.: How Do Grazers Achieve Their Distribution? A Continuum of Models from Random Diffusion to the Ideal Free Distribution Using Biased Random Walks. The American Naturalist 153, 509–526 (1999)
- Fouss, F., Pirotte, A., Renders, J.-M., Saerens, M.: Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation. IEEE Transactions on Knowledge and Data Engineering 19, 355–369 (2007)

- Freeman, L.C.: Centrality in networks: I. Conceptual clarification. *Social Networks* 1, 215–239 (1979)
- Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks* 13, 141–154 (1991)
- Kirchhoff, G.: On a deduction of Ohm's laws, in connexion with the theory of electrostatics. *Philosophical Magazine* 37, 463 (1850)
- Newman, M.E.J.: A measure of betweenness centrality based on random walks. *Social Networks* 27, 39–54 (2005)
- Noh, J.-D., Rieger, H.: Random walks on complex networks. *Phys. Rev. Lett.* 92, 118701 (2004)
- Saerens, M., Achbany, Y., Fouss, F., Yen, L.: Randomized Shortest-Path Problems: Two Related Models. *Neural Computation* 21, 2363–2404 (2009)
- Travers, J., Milgram, S.: An experimental study of the small world problem. *Sociometry* 32, 425–443 (1969)
- Yen, L., Saerens, M., Mantrach, A., Shimbo, M.: A Family of Dissimilarity Measures between Nodes Generalizing both the Shortest-Path and the Commute-time Distances. In: *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–793 (2008)
- Zhou, T.: Mixing navigation on networks. *Physica A* 387, 3025–3032 (2008)

Dynamic Targeting in an Online Social Medium

Peter Laflin¹, Alexander V. Mantzaris², Fiona Ainley¹, Amanda Otley¹,
Peter Grindrod³, and Desmond J. Higham²

¹ Bloom Agency: Green Sand Foundry, 99 Water Lane, Leeds,
LS11 5QN, United Kingdom
plaflin@bloomagency.co.uk

² Department of Mathematics and Statistics: University of Strathclyde,
26 Richmond Street Glasgow, G1 1XH, United Kingdom
alexander.mantzaris@strath.ac.uk

³ Department of Mathematics: University of Reading, Whiteknights,
P.O. Box 220, Reading RG6 6AX, United Kingdom
p.grindrod@reading.ac.uk

Abstract. Online human interactions take place within a dynamic hierarchy, where social influence is determined by qualities such as status, eloquence, trustworthiness, authority and persuasiveness. In this work, we consider topic-based Twitter interaction networks, and address the task of identifying influential players. Our motivation is the strong desire of many commercial entities to increase their social media presence by engaging positively with pivotal bloggers and Tweeters. After discussing some of the issues involved in extracting useful interaction data from a Twitter feed, we define the concept of an *active node subnetwork sequence*. This provides a time-dependent, topic-based, summary of relevant Twitter activity. For these types of transient interactions, it has been argued that the flow of information, and hence the influence of a node, is highly dependent on the timing of the links. Some nodes with relatively small bandwidth may turn out to be key players because of their prescience and their ability to instigate follow-on network activity. To simulate a commercial application, we build an active node subnetwork sequence based on key words in the area of travel and holidays. We then compare a range of network centrality measures, including a recently proposed version that accounts for the arrow of time, with respect to their ability to rank important nodes in this dynamic setting. The centrality rankings use only connectivity information (who Tweeted whom, when), but if we post-process the results by examining account details, we find that the time-respecting, dynamic, approach, which looks at the follow-on flow of information, is less likely to be ‘mised’ by accounts that appear to generate large numbers of automatic Tweets with the aim of pushing out web links. We then benchmark these algorithmically derived rankings against independent feedback from five social media experts who judge Twitter accounts as part of their professional duties. We find that the dynamic centrality measures add value to the expert view, and indeed can be hard to distinguish from an expert in terms of who they place in the top ten. We also highlight areas where the algorithmic approach can be refined and improved.

1 Motivation

Centrality measures have proved to be extremely useful for identifying important players in an interaction network [1]. Although the fundamental ideas in this area were developed to analyse a single, static network, there is a growing need to develop tools for the *dynamic* case, where links appear and disappear in a time-dependent manner. Key application areas include voice calls [2, 3], email activity [4, 3], online social interaction [5], geographical proximity of mobile device users [6], voting and trading patterns [7, 8] and neural activity [9, 10].

This work focuses on the use of centrality measures to discover influential players in a dynamic Twitter interaction network, with respect to a given topic, with the aim of finding suitable targets from a marketing perspective. In this social interaction setting, the idea of key players, who influence the actions of others, is intuitively reasonable. Empirical evidence is given in [11] for *discussion catalysts* in an on-line community who are “responsible for the majority of messages that initiate long threads.” Further, Huffaker [12] identifies *on-line leaders* who “trigger feedback, spark conversations within the community, or even shape the way that other members of a group ‘talk’ about a topic.” Experiments in [13] on email and voice mail data found evidence of individuals “punching above their weight” in terms of having an ability to disseminate or collect information that cannot be predicted from static or aggregate summaries of their activity. These people were termed *dynamic communicators*, and an explanatory model, based on an inherent hierarchy among the nodes, was suggested. Such concepts make it clear that the dynamic nature of the links plays a key role—the *timing* and *follow on effect* of an interaction must be quantified if key players are to be identified. A recent business-oriented survey [14, Section 4] lists network dynamics as a key technical challenge, and the authors in [15] argue that “the temporal aspects of centrality are underrepresented.”

Several recent articles have addressed the issue of discovering important or influential players in networks derived from Twitter data. The work in [16] focused on how a shortened URL is passed through the network. Using the premise that a person who passes on such a URL has been influenced by the sender, it studies the structure of cascades. Related work in [17] looked at large scale information spread on the Twitter follower graph in order to measure global activity. The authors in [18] studied a large scale Twitter follower graph and compared three measures that quantify types of influence: number of followers (out degree), number of retweets and number of mentions, finding little overlap between the top Tweepers in each category. Similarly, [19] also ranked users by the number of followers and compared with ranking by PageRank, finding the two measures to be similar. By contrast, they found that the retweet measure produces a very different ranking. We note that none of the influence measures considered in [18, 19] fully respect the time-ordering of Twitter interactions. For example, reversing the arrow of time does not change the count of followers, retweets or mentions. In this sense, they overlook a crucial aspect of the interaction data. Our work differs from that described above by (a) focussing on subject-specific Tweets of interest in a typical business application, (b) building

the interactions between Tweeters on this topic and recording them in a form that we call the active node subnetwork sequence, and (c) comparing a range of centrality measures in this dynamic setting, including one that respects the arrow of time, against independent hand curated rankings from social media experts exposed to the same data.

2 Building the Active Node Subnetwork Sequence

The Twitter business home page at <https://business.twitter.com/basics/what-is-twitter/> explains that

“Anyone can read, write and share messages of up to 140 characters on Twitter. These messages, or Tweets, are available to anyone interested in reading them, whether logged in or not. Your followers receive every one of your messages in their timeline—a feed of all the accounts they have subscribed to or followed on Twitter. This unique combination of open, public, and unfiltered Tweets delivered in a simple, standardized 140-character unit, allows Twitter users to share and discover what’s happening on any device in real time. ”

The number of active Twitter users currently exceeds 140 Million, with over 340 Million Tweets generated per day. Of direct relevance to our work, the business home page adds that

“Businesses can also use Twitter to listen and gather market intelligence and insights. It is likely that people are already having conversations about your business, your competitors or your industry on Twitter. ”

Twitter is a means to send out information over a well-defined network. This brings to life a scenario that social scientists have for many years been using as a theoretical tool to develop concepts and measures. In particular, given only a network interaction structure, perhaps describing social acquaintanceship, it has proved extremely useful to imagine that information flows along the links and thereby to identify important actors [20, 1]. In this setting, most centrality measures are defined through, or can be motivated from, the idea of studying random walks along the edges [21], or deterministically counting geodesics, paths, trails or walks [22]. These ideas have been extremely well accepted and widely used, despite the obvious simplifications that the methodology involves. For example, even if we accept that social acquaintanceship is a reasonable proxy for the links along which information flows, there are issues concerning

Link Types: if A and B are acquainted professionally and A passes on some work-related news to B, then it is reasonable to expect that B is more likely to pass this news on to professional colleagues than other friends. So we could argue that some $A \rightarrow B \rightarrow C$ paths have a greater chance of being traversed than others.

Link Dynamics: if A and B meet only on a Sunday evening, and B and C meet only on a Monday morning, then we could argue that even though the undirected path $A \leftrightarrow B \leftrightarrow C$ exists in the network, the route $A \rightarrow B \rightarrow C$ is a more likely conduit for news than $C \rightarrow B \rightarrow A$. This is because B meets C soon after an $A \rightarrow B$ exchange, and hence is more likely to (a) remember and (b) regard as topical, any information received from A. This gives another sense in which paths are not created equal.

By exploiting features of the Twitter data, we can, to some extent, sidestep the shortcomings above while retaining the elegance and simplicity of the network-based view:

Link Types: each link represents a physical exchange of information that is known to have taken place (rather than a proxy such as social acquaintanceship), and moreover, by filtering based on Tweet content, we can, in principle, record only links that are relevant to a specific topic of interest,

Link Dynamics: the Twitter data gives us access to the time at which each piece of information was disseminated.

Twitter’s follower graph, where nodes represent users and a directed link connects a user to a follower, has been studied, for example, in [18, 19, 17]. In our work, we wish to focus on users who are engaging with a particular topic, so a natural first step is to look at those who send Tweets containing a predefined set of phrases. In principle, the followers of all such users are exposed to the information in those Tweets. However, in practice we do not know if or when a follower reads a Tweet or acts upon it outside the Twitter platform. In this work, we focus on clearly *active* nodes, that is, users who send out at least one Tweet on the required topic. We then focus on directed user-to-follower connections that involve these active nodes. As well as ruling out those Tweets that land on ‘stony ground’ this pruning exercise generally has the effect of reducing the size of the network considerably; an issue that is of importance if we wish to consider global Tweets about popular topics over long time scales.

To be precise, we use the Twitter feed to construct an *active node subnetwork sequence* as follows.

Definition 1. *The active node subnetwork sequence:*

- Start the clock at time t_{start}
- Listen to all Tweets that contain the required phrase(s)
- Each time a new Tweet is recorded, make sure the sender and all the sender’s followers are nodes in the network (i.e. add them if necessary), and add a time-stamped directed link from the sender node to all follower nodes.
- Stop the clock at time t_{end}
- Post-process the network by removing all nodes that have zero aggregate out degree, i.e., remove those people who did not send out any relevant Tweets.
- Slice the data into M windows of size $\Delta t = (t_{\text{end}} - t_{\text{start}})/M$. We will let $t_k = t_{\text{start}} + (k - 1)\Delta t$. Then, for $k = 1, 2, \dots, M$, the k th window covers the time period $[t_k, t_{k+1}]$ and is represented by an integer-valued matrix $A^{[k]}$.

Here $(A^{[k]})_{ij}$ records the number of links from node i to node j that appeared in this time period.

- Binarize each $(A^{[k]})_{ij}$, that is, set all positive integers to the value 1. (See the remark below for a discussion of this step.)

Implicit in this definition is the simplifying assumption that a Tweet has an influence over a fixed period of time, Δt . It may be argued that a Tweet, once sent, exists for ever and should create a permanent link that perpetuates across all subsequent time windows. However, we believe that a more compelling argument is that Tweets are time-sensitive and fairly rapidly disappear down a typical follower’s timeline. The choice of Δt then quantifies the typical “read and respond” time.

We emphasize in particular that reducing Δt does not necessarily give a more accurate representation of reality—although we know the precise time that the Tweet was sent, we do not know if or when each follower digests the content. On the other hand taking Δt too large (e.g. one giant window) causes us to lose information about the time-ordering of the Tweets.

We constructed an active node subnetwork sequence by listening to Tweets containing the phrases **city break**, **cheap holiday**, **travel insurance**, **cheap flight** and two phrases relating to specific travel brands. This simulates a typical client-driven investigation on behalf of a travel company wishing to improve its social media presence. The collection took place from 17 June 2012 at 14:41 to 18 June 2012 at 12:41. We took Δt equal to 66 minutes, producing 20 time windows. The total number of Tweeters and followers associated with this data set is 442,948. Restricting attention to active nodes, with nonzero out degree, reduced the network size to $N = 590$.

We observed that some accounts can Tweet a lot in a short space of time. One account Tweeted 104 times in timeframe 10 and a further 23 times in timeframe 11. This account released a total of 127 Tweets in 68 minutes. This motivates our decision to binarize the data within each window—in this way we have not taken account of how many times an account Tweets, but rather we represent the fact they did Tweet in that timeframe. This is done to try to stop the overall result being influenced by accounts using a high volume of automated Tweets. This choice is a balance between allowing a “noisy” account broadcasting automated Tweets to score higher than we would like in our calculations against our ability to pick out influential people by observing a natural increase in the rate of conversation because something interesting or relevant is happening.

To give a feel for the data, Figure 1 visualizes two portions of the the network at the end of the first time window. We will return to this data set in section 4 when we compare centrality measures.

3 Centrality Measures

In the case of a single time point, with binary adjacency matrix $A \in \mathbb{R}^{N \times N}$, the resolvent matrix $(I - \alpha A)^{-1}$ was proposed by Katz [23] as a means to summarize pairwise “influence” under “attenuation through intermediaries.” Here the fixed

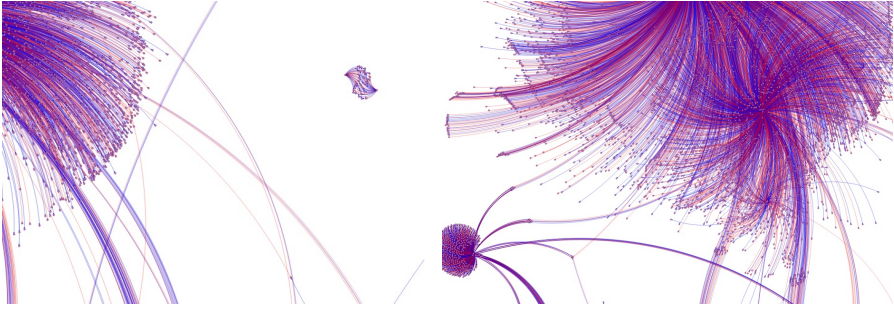


Fig. 1. Two details from the active subnode network sequence at the end of the first time window; in particular, showing the existence of an isolated community

parameter α governs the strength of the attenuation, and for $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ denotes the spectral radius of A , we have

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \alpha^3 A^3 + \dots$$

Using the fact that $(A^p)_{ij}$ records the number of distinct walks¹ of length p from node i to node j [20], we see that the (i, j) element of $(I - \alpha A)^{-1}$ counts the total number of walks of all possible length, with walks of length p downweighted by α^p . The idea of attaching less importance to longer walks is intuitively reasonable, and Katz [23] also points out that α may be interpreted probabilistically, as the chance that a message successfully traverses an edge. It follows that the row sums and column sums of the resolvent quantify the ability of nodes to broadcast and receive information, respectively. Rather than inverting $I - \alpha A$, it is more efficient and numerically accurate to solve a linear system. Hence in our tests we will compute vectors \mathbf{Kb} and \mathbf{Kr} in \mathbb{R}^N satisfying

$$(I - \alpha A)\mathbf{Kb} = \mathbf{1}, \quad (I - \alpha A^T)\mathbf{Kr} = \mathbf{1}, \tag{1}$$

where $\mathbf{1} \in \mathbb{R}^N$ is the vector with all entries equal to one. In this case the i th components of \mathbf{Kb} and \mathbf{Kr} measure the ability of node i to broadcast and receive messages, respectively, across the static network represented by the binary adjacency matrix A , in the sense of Katz. The nodes may then be ranked according to these scores.

In the limit $\alpha \rightarrow 0$, longer walks make a negligible contribution in (1), and, ignoring uniform shifts and scalings that do not alter the rankings, the measures collapse to out degree and in degree, respectively, that is,

$$(\text{deg}_{\text{out}})_i = \sum_{j=1}^N a_{ij}, \quad (\text{deg}_{\text{out}})_j = \sum_{i=1}^N a_{ij}. \tag{2}$$

¹ A walk of length w from node i to node j is characterized by a sequence of w edges $i \rightarrow i_1, i_1 \rightarrow i_2, \dots, i_{w-1} \rightarrow j$. There is no requirement for the edges, or the nodes that they connect, to be distinct.

We note that these two quantities are also widely used as centrality measures in their own right [20, 1].

In recent years, several authors have pointed out that concepts such as geodesics, paths and walks can be extended to the case of a time-ordered sequence of networks [24–26, 8]. We focus here on the dynamic walk notion from [3] which produces generalizations of the Katz centrality measures (1) that are feasible for large-scale network computations. Using the notation introduced in section 2, the following definition was made in [3].

Definition 2. A dynamic walk of length w from node i_1 to node i_{w+1} consists of a sequence of edges $i_1 \rightarrow i_2, i_2 \rightarrow i_3, \dots, i_w \rightarrow i_{w+1}$ and a non-decreasing sequence of times $t_{r_1} \leq t_{r_2} \leq \dots \leq t_{r_w}$ such that $A_{i_m, i_{m+1}}^{[r_m]} \neq 0$.

Dynamic walks are easily counted by forming appropriate matrix powers. For example, with the (i, j) component relating to walks from node i to node j ,

- $A^{[1]}A^{[2]}$ counts all dynamic walks of length two that use one edge at time t_1 followed by one edge at time t_2 ,
- $A^{[5]}A^{[5]}A^{[9]}A^{[10]}$ counts all dynamic walks of length four that use two edges at time t_5 , and then an edge at time t_9 and finally an edge at time t_{10} .

Following the Katz idea of downweighting walks of length w by α^w , this leads to the expression

$$\left(I - \alpha A^{[1]}\right)^{-1} \left(I - \alpha A^{[2]}\right)^{-1} \dots \left(I - \alpha A^{[M]}\right)^{-1}$$

as a summary of the number of dynamic walks that exist between each pair of nodes. In this case, α should be chosen below the reciprocal of $\max_{1 \leq k \leq M} \rho(A^{[k]})$.

Expressing these computations in terms of sparse linear systems, rather than matrix inversions, and normalizing to prevent underflow and overflow, we arrive at the dynamic broadcast and receive centralities from [3] given by

$$\text{Db} := \text{Db}^{[1]}, \quad \text{Dr} := \text{Dr}^{[M]}, \quad (3)$$

where the vector sequence $\{\text{Db}^{[r]}\}_{r=1}^{M+1}$ is computed iteratively by setting $\text{Db}^{[M+1]} = \mathbf{1}$ and then solving

$$\left(I - \alpha A^{[r]}\right) \text{Db}^{[r]} = \text{Db}^{[r+1]}$$

and normalizing

$$\text{Db}^{[r]} \mapsto \frac{\text{Db}^{[r]}}{\|\text{Db}^{[r]}\|_2},$$

for $r = M, M - 1, \dots, 1$.

Similarly, receive centralities may be computed by transposing the adjacency matrices.

4 Experimental Results

4.1 Comparison of Network Centrality Measures

Using the holiday travel based active node network sequence described in section 2, we now compare the six centrality measures outlined in section 3. In order to apply the measures designed for static networks, we formed a single thresholded binarized network, B . To do this, we first formed the time-aggregate matrix $A_{\text{sum}} := \sum_{k=1}^M A^{[k]}$. Then we thresholded based on a value θ , so that $(B)_{ij} = \begin{cases} 1 & \text{if } (A_{\text{sum}})_{ij} \geq \theta, \\ 0 & \text{otherwise.} \end{cases}$ Here θ is chosen so that the number of edges in B matches, as closely as possible, the average number of edges in $\{A^{[k]}\}_{k=1}^M$. For convenience, we use the following descriptors:

- **Katz broadcast** and **Katz receive** denote the centrality measures in (II) applied to the thresholded binarized network. We used $\alpha = 0.9/\rho(B)$.
- **Dynamic broadcast** and **dynamic receive** denote the centrality measures (3) on the active node subnetwork sequence. We used $\alpha = 0.9/\max_k \rho(A^{[k]})$.
- **Out degree** and **in degree** denote the row sums and column sums of A_{sum} respectively; the rankings based on these measures are equivalent to the $\alpha \rightarrow 0$ rankings from dynamic broadcast and receive.

Because our aim is to identify influential Tweeters, we intuitively expect the three broadcast-based measures (out degree, Katz broadcast and dynamic broadcast) to be more useful than the three receive-based measures (in degree, Katz receive and dynamic receive) in this context.

Each of these six centrality measures produces a vector in \mathbb{R}^{590} , which can be used to determine (up to ties) a ranking, that is, a permutation of the integers 1 to 590. There are, of course, many ways to compare these different measures. The upper panel in Table I shows the Kendall tau and Spearman rho correlation coefficients for each pairwise combination of measures. In the context of using the measures to identify important nodes, rather than looking at correlation across the entire set of centralities it is perhaps more meaningful to focus on those nodes that are identified as important. The lower panel in Table II therefore shows the overlap, that is, the number of common nodes, among the top ten and top twenty lists in a pairwise manner. The tables indicate a slightly higher match within, rather than across, the broadcast-based measures and the receive-based measures, although this is not completely consistent; for example Katz broadcast and Katz receive have the highest pairwise correlations.

For a visual overview, Figures 2 and 3 scatter plot the dynamic broadcast centrality against each other measure. In Figure 2 we see that dynamic broadcasting and dynamic receiving are quite different achievements. One node comes top in both measures, and from Table I we see that 16 nodes appear in both top 20 lists. However, the orderings within the top twenty are clearly different. Perhaps most noticeably, the fourth highest dynamic broadcaster ranks relatively poorly according to dynamic receive. Further investigation revealed that this account

Table 1. Upper panel shows Kendall tau correlation across pairs of node rankings in upper triangle and Spearman rho correlation across pairs of node rankings in lower triangle. Lower panel shows overlap between top 10 across pairs of node rankings in upper triangle and overlap between top 20 across pairs of node rankings in lower triangle.

| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
|-------------------|------------|-----------|----------------|--------------|-------------------|-----------------|
| out degree | | 0.48 | 0.34 | 0.35 | 0.60 | 0.46 |
| in degree | 0.48 | | 0.43 | 0.46 | 0.47 | 0.64 |
| Katz broadcast | 0.31 | 0.42 | | 0.87 | 0.34 | 0.42 |
| Katz receive | 0.33 | 0.47 | 0.88 | | 0.36 | 0.45 |
| dynamic broadcast | 0.69 | 0.52 | 0.32 | 0.35 | | 0.49 |
| dynamic receive | 0.47 | 0.73 | 0.41 | 0.45 | 0.54 | |

| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
|-------------------|------------|-----------|----------------|--------------|-------------------|-----------------|
| out degree | | 2 | 5 | 2 | 6 | 3 |
| in degree | 6 | | 1 | 1 | 2 | 2 |
| Katz-broadcast | 11 | 3 | | 3 | 6 | 3 |
| Katz-receive | 4 | 7 | 4 | | 3 | 9 |
| dynamic broadcast | 6 | 4 | 7 | 15 | | 4 |
| dynamic receive | 4 | 5 | 5 | 18 | 16 | |

belongs to a travel insurance brand. The account (id = 34²) appears to supply automated Tweets on the subject of insurance. (In the exercise reported in subsection 4.2, the social media experts ranked this account as mid-range because the Tweets generated were not personalised according to best practice.)

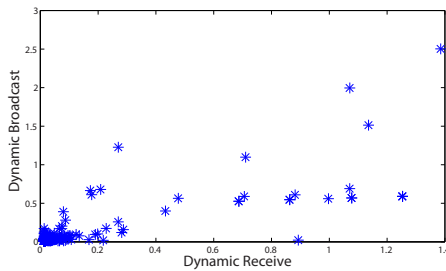


Fig. 2. Dynamic broadcast against dynamic receive for the active nodes

In the upper left picture of Figure 3 the second highest dynamic broadcaster stands out as having a relatively low Katz broadcast measure. This account (id = 398) Tweets stories about travel. As with account 34 discussed above, there were a lot of automated Tweets. This appears to be an account that is looking to send out, rather than receive, links, and most Tweets contain links to websites—however the content of the Tweets was felt to be relevant to the topic, which is why the account appears in third place in the overall expert summary of subsection 4.2 (Table 4).

² The id numbers are local to this experiment and have no further significance.

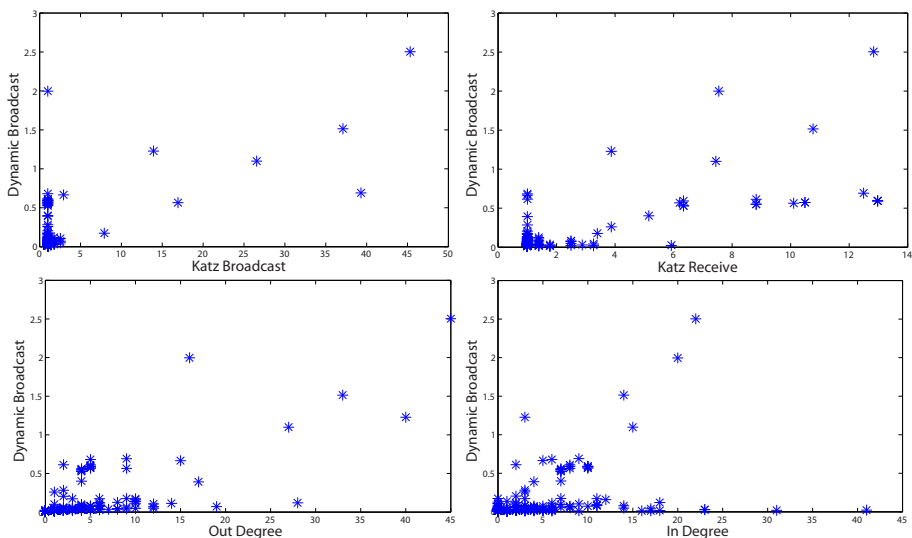


Fig. 3. Dynamic broadcast against: upper left: Katz broadcast, upper right: Katz receive, lower left: out degree, lower right: in degree, for the active nodes

In the upper right picture of Figure 3 the first and third best Katz receivers ($id = 388$ and 394 , respectively) are seen to be poor dynamic broadcasters. These accounts belong to news aggregators Tweeting about travel and other news. They passed on similar information and have a similar follower profile.

The fourth highest out degree node is seen in the lower left picture of Figure 3 to be a very poor dynamic broadcaster. This unusual account ($id = 341370$) Tweeted about lots of different topics but has only 35 followers. This case caused an interesting split between the social media experts during the exercise discussed in subsection 4.2. Two experts rated the account as mid range and three rated it lowest of those considered. On closer inspection, we found that the accounts which were subsequently retweeting exhibited some strange behaviour that was not obvious at first glance. Figure 4 illustrates one set of retweets, suggesting that an automated process is at work in the retweeting operation, in an effort leverage influence.

More generally, it is clear from Figure 3 and Table 1 that high out degree nodes can have very poor dynamic broadcast centrality—generating a high bandwidth does not directly translate into effective communication in this sense.

In the lower right picture of Figure 3 there are three accounts with very high in degree that are not good dynamic broadcasters. The highest in degree account ($id = 172$) belongs to a holiday company based in Kauai, Hawaii, Tweeting about holidays there. The account produces some automated Tweets but they do not appear to be designed simply to publicize links. The next ($id = 158$) was regarded by the experts as the most heavily automated of those considered, generating Tweets on a wide range of subjects, not focused in any area, with the apparent aim of link distribution. The third ($id = 31$) was a news aggregator in the manner of accounts 388 and 394 discussed above.

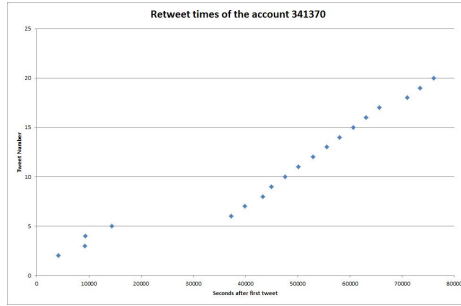


Fig. 4. Retweet times for a Tweet emerging from account id 341370

4.2 Results from Social Media Experts

In order to benchmark the centrality results, we enlisted the help of five professionals working in social media who have day-to-day experience of ranking and targeting accounts based on Twitter data. It is not feasible to study by eye the full set of dynamic interaction data across the 590 active nodes—indeed, this is a key motivation for the use of automated tools. Hence, in collaboration with social media professionals, and with the aid of the six centrality measures, we focused attention on a list of 41 accounts that were felt to be highly relevant. The five experts were then given access to the full details of the Tweets from this list, including the content of their messages, and asked to rank them in order of importance. They had no knowledge of the six centrality rankings.

Table 2 records the level of consistency between the five experts, in terms of Kendall tau correlations across the 41 accounts and overlap between the top 10 in each list. We see that although the correlation is generally positive, there is some considerable variation between the views. Hence, although we regard this information as providing a very useful guide, we do not see it as a “gold standard” with which to judge centrality measures in this context.

Table 2. Upper: Kendall tau correlation between rankings of the 41 Tweets from pairs of experts. Lower: overlap amongst top ten in rankings of the 41 Tweets from pairs of experts.

| | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 |
|----------|----------|----------|----------|----------|----------|
| Expert 1 | | -0.10 | 0.93 | 0.19 | 0.33 |
| Expert 2 | 5 | | -0.10 | 0.31 | 0.14 |
| Expert 3 | 10 | 3 | | 0.20 | 0.37 |
| Expert 4 | 6 | 5 | 6 | | 0.55 |
| Expert 5 | 6 | 5 | 6 | 5 | |

For Table 3 we merged the five different expert rankings of the 41 nodes, giving equal weight to each, into a single list. We then compared this ‘average

expert’ with the rankings of these 41 nodes produced by each of the six centrality measures. We show the top ten overlap. Comparing with the results in Table 2, it may be argued that at least three of the centrality measures are almost indistinguishable from experts in this sense. To give more insight, Table 4 shows the top 10 list for the averaged expert and the three broadcast-based centralities. We see that dynamic broadcast has a top three that includes two of the experts’ top three. Out degree and Katz broadcast have one such ‘correct’ answer in their top three. We also note that the centrality rankings are closer to each other than to the average expert, in terms of overlap.

Table 3. Overlap amongst top ten for each of the six centrality measures against the average over five experts

| | | | | | | |
|---------|------------|-----------|----------------|--------------|-------------------|-----------------|
| | out degree | in degree | Katz broadcast | Katz receive | dynamic broadcast | dynamic receive |
| Overlap | 4 | 3 | 2 | 1 | 3 | 2 |

Table 4. Account ids in rank order from 1 to 10. Column 1: average over five experts. Column 2: out degree. Column 3: Katz broadcast. Column 4: dynamic broadcast.

| average expert | out degree | Katz broadcast | dynamic broadcast |
|----------------|------------|----------------|-------------------|
| 397 | 74 | 74 | 74 |
| 362 | 34 | 302 | 398 |
| 398 | 362 | 362 | 362 |
| 341 | 341370 | 358 | 34 |
| 289 | 358 | 375 | 358 |
| 345 | 71 | 34 | 302 |
| 462 | 345 | 341 | 397 |
| 212 | 398 | 352 | 352 |
| 71 | 352 | 200 | 373 |
| 18 | 484 | 409 | 380 |

5 Summary and Future Work

Our aim in this work was to investigate the use of network centrality measures on appropriately processed Twitter data as a means to target influential nodes. We found that these measures can extract value, both in isolation and when combined, especially when the time-dependent nature of the interactions is incorporated. In particular, benchmarking against the views of five experts in social media showed that the dynamic broadcast centrality results are, in the sense of overlap at the important upper end, hard to distinguish from hand curated expert rankings.

There are many open questions and remaining challenges in this area. Obvious issues include the best way to choose algorithmic parameters, such as the time

window size, Δt , and Katz downweighting parameter, α . For long time periods, or real-time monitoring, it would also be of interest to consider downweighting information over time, as described in [27]. A bigger challenge is detecting, categorizing and dealing with accounts that generate automated Tweets. Here, it may be preferable to leave the elegant but simplified network viewpoint and dig down into the precise correlations over time of account activity.

Acknowledgments. Alexander V. Mantzaris, Desmond J. Higham and Peter Grindrod thank the EPSRC and RCUK Digital Economy programme for support through the project Mathematics of Large Technological Evolving Networks (MOLTEN). Peter Laffin, Fiona Ainley and Amanda Otley thank the Technology Strategy Board of the UK for funding the SMART project entitled Digital Business Analytics for Decision Makers. Work done on that project has contributed to the knowledge shared in this paper, especially with regard to building networks from the data. They also thank colleagues at Bloom Agency for allowing them time to work on this project, outside of their usual client workload. We thank Alex Craven, Phil Jefferies and Claire Hunter-Smith for liaising with social media experts and coordinating their feedback and comments.

References

1. Newman, M.E.J.: Networks an Introduction. Oxford University Press, Oxford (2010)
2. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. *Proc. Nat. Acad. Sci.* 106(36), 15274–15278 (2009)
3. Grindrod, P., Higham, D.J., Parsons, M.C., Estrada, E.: Communicability across evolving networks. *Physical Review E* 83, 046120 (2011)
4. Barabási, A.L.: The origin of bursts and heavy tails in human dynamics. *Nature* 435, 207–211 (2005)
5. Tang, J., Scellato, S., Musolesi, M., Mascolo, C., Latora, V.: Small-world behavior in time-varying graphs. *Phys. Rev. E* 81, 05510 (2010)
6. Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., Gesualdo, F., Pandolfi, E., Rav, L., Rizzo, C., Tozzi, A.E.: Close encounters in a pediatric ward: Measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS ONE* 6(2), e17144 (2011)
7. Bajardi, P., Barrat, A., Natale, F., Savini, L., Colizza, V.: Dynamical patterns of cattle trade movements. *PLoS ONE* 6(5), e19869 (2011)
8. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. *Science* 328, 876–878 (2010)
9. Bassett, D.S., Wymbs, N.F., Porter, M.A., Mucha, P.J., Carlson, J.M., Grafton, S.T.: Dynamic reconfiguration of human brain networks during learning. *Proc. Nat. Acad. Sci.* 108 (2011), doi: 10.1073/pnas.1018985108
10. Grindrod, P., Higham, D.J.: Evolving graphs: Dynamical models, inverse problems and propagation. *Proc. Roy. Soc. A* 466, 753–770 (2010)

11. Gleave, E., Welser, H.T., Lento, T.M., Smith, M.A.: A conceptual and operational definition of ‘social role’ in online community. In: Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1–11. IEEE Computer Society, Los Alamitos (2009)
12. Huffaker, D.: Dimensions of leadership and social influence in online communities. *Human Communication Research* 36, 593–617 (2010)
13. Mantzaris, A.V., Higham, D.J.: A model for dynamic communicators. *European Journal of Applied Mathematics* (to appear, 2012)
14. Bonchi, F., Castillo, C., Gionis, A., Jaimes, A.: Social network analysis and mining for business applications. *ACM Trans. Intell. Syst. Technol.* 2(3), 22:1–22:37 (2011)
15. Shamma, D.A., Kennedy, L., Churchill, E.F.: In the limelight over time: Temporalities of network centrality. In: Proceedings of the 29th International Conference on Human Factors in Computing Systems, CSCW 2011, ACM (2011)
16. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone’s an influencer: quantifying influence on Twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM 2011, pp. 65–74. ACM, New York (2011)
17. Lerman, K., Ghosh, R., Surachawala, T.: Social contagion: An empirical study of information spread on digg and twitter follower graphs. CoRR abs/1202.3162 (2012)
18. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in Twitter: The million follower fallacy. In: ICWSM 2010: Proceedings of International AAAI Conference on Weblogs and Social (2010)
19. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 591–600. ACM, New York (2010)
20. Estrada, E.: *The Structure of Complex Networks*. Oxford University Press, Oxford (2011)
21. Newman, M.: A measure of betweenness centrality based on random walks. *Social Networks* 27(1), 39–54 (2005)
22. Borgatti, S.P.: Centrality and network flow. *Social Networks* 27, 55–71 (2005)
23. Katz, L.: A new index derived from sociometric data analysis. *Psychometrika* 18, 39–43 (1953)
24. Holme, P.: Network reachability of real-world contact sequences. *Physical Review E* 71(4), 046119 (2005)
25. Kim, H., Tang, J., Anderson, R., Mascolo, C.: Centrality prediction in dynamic human contact networks. *Comput. Netw.* 56(3), 983–996 (2012)
26. Kossinets, G., Kleinberg, J., Watts, D.: The structure of information pathways in a social communication network. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Datamining, KDD 2008, pp. 435–443. ACM, New York (2008)
27. Grindrod, P., Higham, D.J.: A matrix iteration for dynamic network summaries. *SIAM Review* (to appear, 2012)

Connecting with Active People Matters: The Influence of an Online Community on Physical Activity Behavior

Maartje Groenewegen¹, Dimo Stoyanov¹, Dirk Deichmann¹,
and Aart van Halteren^{1,2}

¹ VU University Amsterdam, The Netherlands
maartjegroenewegen@live.nl, d.stoyanov@student.vu.nl,
{d.deichmann,a.t.van.halteren}@vu.nl

² Philips Research, Eindhoven, The Netherlands

Abstract. This paper discusses the impact of online social networks as means to motivate people to become more physically active. Based on a data set from 4333 participants we show that the activity level of people that participated in the online community (for 14 weeks) is significantly higher compared to people that choose not to become a member of that community. Detailed analyses show that the number of contacts in the online community does not have a significant effect on the physical activity level while network density even has a significant, negative effect. On the other hand, the activity level of a participant is higher when his or her friends also have a high average activity level. This effect is even higher when a participant's amount of friends increases. Theoretical and managerial implications concerning the impact of online social networks on offline behavior are discussed.

Keywords: Physical activity promotion, social network analyses, motivation.

1 Introduction

Physical activity is regarded as an important element of a healthy lifestyle. Abundant scientific research connects physical activity to many psychological and physical health benefits [26]. Physical inactivity increases the risk of developing several diseases such as diabetes, obesity, cardiovascular diseases, high blood pressure and some cancers [25]. Adults are supposed to be physically active for at least 30 minutes on five days a week, according to public health guidelines [38]. Unfortunately, only 48.1 percent of the U.S. population achieves these physical activity recommendations [11]. In other industrialized countries, the percentage is about the same [38]. Promoting physical activity is a cost-effective way to reduce avoidable healthcare costs.

Over the last couple of years, there has been research on different intervention strategies. Especially the role of the social networks - through mechanisms of social support, social engagement, social influence, social pressure and access to

resources - is emphasized in the literature [39,19,4,23]. Moreover, the internet is seen as an opportunity to promote physical activity to a wide range of people [38]. Social networks play an important role in this. A fabric of social relations is generally seen as a resource available to an individual. This resource can be mobilized to steer or facilitate action and behavior [11,10]. Among others, this is why McNeill, Kreuter and Subramanian [23] advocate a physical promotion program which includes social environmental factors that are able to support change in physical activity behavior.

Earlier studies did research on the influence of social networks on the behavior of actors in the network. A recent study by Christakis and Fowler [14] noticed an increase in obesity over the past 30 years in the United States. They investigated whether we can speak of an obesity epidemic, i.e. obesity spreading throughout the social networks of people. They found considerable evidence of the social network influencing the weight of actors up to three degrees of separation. We build on the research of Christakis and Fowler and advance it further by grounding our work on objective measures of the physical activity level of people. The data is collected through an activity monitor. Additionally, we rely on observed, objective network data from which network characteristics can be calculated.

Vandelanotte et al. [38] reviewed fifteen studies focusing on online interventions to improve physical activity. They conclude that 8 out of 15 studies found an increase in the activity level of people. However, this effect was short lived and when focusing on specific intervention mechanisms it turned out that the number of contacts of the participants with others (in the form of emails, telephone contact, discussion boards, chat sessions or contact with online coaches) was the only factor that could be associated with an increase in the physical activity level of people. Their most important recommendation is therefore in line with earlier literature: Increased interactivity in physical activity promotion programs or websites will enhance engagement and retention of the participants in the physical activity program. However, Vandelanotte et al. focused mainly on website based interventions, which is different from our study. Website based interventions focus on interactions between the service and the user instead of a social network in which people can become friends and observe the behavior of other users. Additionally, website based intervention studies have a broad definition of interactions since email contact as well as online coaches and discussion boards are included. Our research focuses specifically on the online community attached to a promotion program. We examine the influence of peer's activity level and network structural characteristics on a person's activity level.

Taken together, our study provides data-driving insights in the ways through which online social networks can influence people's offline physical activity behavior. Our research question is:

How does the introduction of an online community into an online physical activity promotion program influence the physical activity levels of its participants?

2 Social Connectedness as a Motivator

Over the past decades many theories and models of health behavior change have been introduced. These theories show that for an individual to engage in new behavior, that person first needs to build the motivation and then needs to translate that motivation into behavior [32]. Motivation has been reported as an important element in sports and physical activity [36]. According to the Social Determination Theory (SDT), motives associated with more self-determination are more powerful in changing and maintaining behavior [30]. SDT claims that people have a basic need for relatedness which motivates people to behave in a certain way. People have the basic psychological need to 'feel connected to others within a social milieu' [16] and have a sense of belonging with other people [22]. The higher the extent to which the need for relatedness is fulfilled when people engage in certain behavior, the stronger their motivation will be to continue this behavior.

How does the need for relatedness motivate people? In order to understand the social network structures through which the behavior of people is influenced, we draw on social network theory. According to social network theory actors are part of a structure of interconnected relationships which offer possibilities and constraints to the actor's behavior [8]. The social network in which actors are embedded has certain characteristics that flow from the interpersonal relationships between actors. Degree centrality and density are two important structural characteristics [29] and are the focus of this research.

Degree

Degree centrality concerns 'the extent to which a given individual is connected to others in a network' [33]. When talking about the degree of friends we mean the amount of direct ties going to and from a person [29]. Prior research shows that the higher the number of friends in a network, the larger the increase in the physical activity level of people [23]. There are three ways through which the amount of friends influences the physical activity behavior of people.

First, the amount of friends has an influence on the level of engagement of participants. Burke, Marlow and Lento [9] state that Facebook users with many friends are more engaged with the community. Additionally, Ellison, Steinfield and Lampe [17] use the amount of Facebook friends as a measure for engagement. Higher engagement in a community is connected to a higher sense of belongingness in a community [37]. People with lots of friends are more engaged, feel more related to, and identify more with the community. Social identity theory states that if people identify themselves within a certain group, they will behave according to the norms of this group [35].

Second, social pressure increases when people have more friends [17] and this causes people to behave in a certain way [15]. The actor experiences a higher pressure to conform to the group norms [35] and will be motivated to adapt his/her behavior.

Third, more friends increases the amount of social support. Friends encourage and motivate people to reach their goals and to change their behavior. 'Social

support specific to physical activity may provide the initial motivation to increase physical activity levels' [18, p. 786].

Hence, having more friends increases engagement with the community, causes higher social pressure which motivates people to conform to group norms and lead to higher levels of social support which motivates participants to be more physically active. We hypothesize:

H1: There is a positive association between the number of friends a person has in the online community of a health promotion program and his/her physical activity level

Density

Another structural characteristic of the network is density. Density 'describes the general level of linkage among [people in a network]' [31, p. 69]. There might be two reasons for why density should influence people's physical activity level.

First, a dense network might increase the feeling of attachment and belongingness. This is because a dense network facilitates a common identity [21]. In such networks, the participants' friends are connected to each other and therefore form a (sub) community which provides the feeling of attachment. A dense network will influence the behavior of its members stronger than a less dense network [21]. Haynie [21] argues that this is the case because the opportunities to communicate and interact are higher and therefore it is more likely that views on physical activity behavior are expressed more frequently. Moreover, Friedkin [20] argues that if the network is dense, the likelihood that actors have influence on each other increases, because people are more aware of each others' opinions in dense groups [20]. Additionally, a social identity is easier created in a dense, small network rather than in a big network full of structural holes [28]. As said earlier, social identity and feeling of belongingness causes people to conform to group norms.

Besides the feeling of belongingness and the social identity that are created through a dense network, the social pressure is also higher in a dense network. When a participant's friends are connected to each other they can exert more pressure on the participant which causes participants to conform to group norms. We expect a direct influence of density of the ego-network on the physical activity level of the participant.

H2: There is a positive association between the density of a person's network in the online community of a health promotion program and his/her physical activity level.

Friends Physical Activity Levels

Instead of looking at network characteristics such as the amount of friends and density, the behavior of a participant's friend is also of great importance. Berkman et al. [4] mention social influence as one of the potential mechanisms of social networks effecting physical activity behavior. They argue that people 'obtain normative guidance' when they compare their attitudes or behavior to that of the

group [4, p. 849]. When the behavior or attitude is confirmed, the behavior will not be changed, while it will be changed when it is not congruent with the group's behavior or attitude. Again, the norms and values of the group are thus important. Previous research in other health contexts has found that the behavior of peers is the most important predictor for the behavior of people themselves [4]. Moreover, experimental research by Centola [12] confirms that the healthy behavior of friends increases the level of a person's engagement towards healthy behavior. Through social interaction, attitudes and behavior are formed [20]. Hence, people compare their own behavior with that of others and adapt to the behavior of others. Drawing on these earlier insights, we therefore argue that the average activity level of a person's friends influences the person's own activity level.

H3: There is a positive association between the average physical activity level of a person's friends in the online community of a health promotion program and his/her physical activity level.

Interactions

Earlier we discussed the ways through which social networks influence people's behavior. We argued that the amount of friends, the density of a network and the behavior of friends all influence a person's behavior. However, the behavior of friends is considered as the most powerful influence [4]. It is important to have a lot of friends, but it is even more important to have a lot of friends with a high average physical activity level. Additionally the density of a network is important. However, when the average physical activity of a participant's friends is high, the impact of the density of a network on a participant's physical behavior will be strengthened. Participants conform to the physical activity behavior of their friends and if this is high, they will thus be motivated to increase their activity level. This leads to the following two hypotheses:

H4: The average physical activity level of a person's friends in the online community of a health promotion program enhances the positive relationship between the number of friends a person has and his/her physical activity level.

H5: The average physical activity level of a person's friends in the online community of a health promotion program enhances the positive relationship between the density of a person's network and his/her physical activity level.

3 Methods

Sample, Setting and Procedure

A multinational company offering a physical activity promotion program was approached and they provided de-identified data from a subset of participants who are between 18 and 65. Due to confidentiality reasons, we do not have access to personal information except for gender and country of residence. The program uses an activity monitor to measure activity energy expenditure [6]. The monitor

has to be connected regularly to a computer, so that the data is uploaded to the online system. The promotion program has three different phases. The first phase is a one week assessment period. Its purpose is to evaluate the user's activity level during his/her daily routine. The assessment is followed by a 12 week plan which aims to gradually increase the participants' activity level towards an end-goal. The goal is determined from the activity reported during the assessment week. After the plan, the members of the program can opt to start a new 12 week plan to further increase their activity level or simply continue with the activity goal set during the last week of their program. The activity promotion program provides an online community and joining this social network is optional for the users. Each member of the community can connect and become friends online with others, exchange messages and see the relative achievements of themselves and of others (which are only visible after participants become friends).

We extracted all the information of 6291 participants, of which 2418 opted into the community. Due to missing values we are left with a total of 4423 participants of which 1674 were community members. Including the cases with missing values does not show significant differences in our key variables. The participants are based in 9 different countries. All the data has been recorded from 01-05-2010 to 01-08-2010. The starting day of our data collection is also the day when the online community feature became available. We aggregated the data on a per week basis.

33.9% of all the women in the sample opted into the community versus 39.7% of the men. Members of the community login significantly more often and have a significantly lower BMI than participants who did not opt into the community. The average physical activity level (PAL) of all the participants (including community and non-community members) is 1.60. 57.9% of the participants are female, 42.1% are male. 78.5% are based in the US, 20.1% are in The Netherlands, 1.4% are in other countries.

Dependent Variable

Physical Activity Level (PAL) is measured on a per day basis and averaged for a given week. PAL [27] is a standard unit for measuring the intensity of one's physical activity. Usually it is computed as $PAL = TEE \times BMR^{-1}$, where TEE represents the total energy expenditure and BMR is the basal metabolic rate, which is age dependent. A physical activity monitor was used for precise measurement of PAL.

Independent Variables

Degree is operationalized as the number of connections which a member of the online community has during a particular week.

Density measures the number of actual ties of a participant divided by the number of possible ties [31]. in the ego-network. This means that only the ties between ego's friends are taken into account. Ego-network density is thus the amount of existing ties between ego's friends divided by the amount of total ties possible between ego's friends (UCINET [7]).

Peer PAL is the average of the PAL of one's peers (i.e., friends) in the online community during a week.

Control Variables

Prior PAL is the lagged PAL value for a person. We control for prior PAL because focal activity values might significantly be driven by peoples' past behavior. Ajzen and Madden [3] use past behavior as a control variable when testing the theory of planned behavior and state that past behavior has an influence on the contemporary behavior.

Furthermore, we control for Body Mass Index (BMI) as it might influence how seriously people take the activity promotion program and therefore their physical activity level. People with a high BMI can be more motivated to lose weight than people with a lower BMI [5] and therefore display higher levels of activity. We include the average BMI value of a person during the week since people are able to adjust this value constantly. BMI is based on two self-reported measures; weight and height and is calculated by dividing weight by the squared height of a person [5].

As a further indication of involvement of people in the activity promotion program we controlled for the number of times users logged into the online system during a corresponding week. Logging in entails connecting the activity monitor to a computer but not necessarily mean that the community was used.

Another variable that needs to be controlled for is the status of the participant. Status refers to the phase of a participant in the program. Participants with status 'plan' might be more active since they want to achieve a certain goal, different from after the plan when sustaining an activity level is more important. Goal setting theory predicts that when a specific, quantified goal is set, this will lead to a higher performance [34].

Additionally, we controlled for gender of the participant (self-reported), their country and the number of weeks that have passed since the user joined the online community (community week).

Analysis

We use a panel approach to model that each participant has multiple, non-independent observations. To better account for unobserved heterogeneity, we estimate random effects for panel models [2]. In doing so, we insert additional participant related error terms that allow observations of the same participant to be correlated. Before including the interactions we mean centered the main effects.

4 Results

Tables 1 and 2 present summary statistics and the correlation matrix for the combined data set and community data set, respectively.

Table 1. Correlations Complete Data Set

| | Mean | Std. Dev. | Min. | Max | 1 | 2 | 3 | 4 |
|--------------|-------|--------------|------|-------|-------|-------|------|------|
| 1. PAL | 1.60 | 0.24 | 1.11 | 3.83 | | | | |
| 2. Prior PAL | 1.60 | 0.24 | 1.11 | 3.83 | 0.75 | | | |
| 3. Gender | 1.58 | 0.49 | 1.00 | 2.00 | -0.13 | -0.13 | | |
| 4. Logins | 2.38 | 3.78 | 0.00 | 95.00 | 0.32 | 0.25 | 0.00 | |
| 5. BMI | 28.67 | 6.47 | 6.80 | 86.63 | -0.14 | -0.13 | 0.01 | 0.02 |

Everything above $r=0.00$ is significant at $P<0.05$

Checking for variance inflation factors we observe that all stayed well below the critical value of $VIF = 10$, so there is no concern of multicollinearity.

From Table 4 we can conclude that the community influences the physical activity level of program participants. For this regression the data set containing both community and non-community users was used. The regression contains the control variables and the dummy variable for community. As we can observe from Table 4, the dummy variable for community is significant. This indicates that there is a difference between non-community and community users and their physical activity level: Participants who are members of the community have a significantly higher increase in PAL than non-community members.

From Table 3 we can observe the coefficients and significance of every variable. The first model contains the direct effect of the control variables, the second model adds the main effects of degree, density and peer PAL, the third model contains the interaction effect of degree and peer PAL, the fourth model adds the interaction effect of density and peer PAL and the fifth model contains all variables and interaction effects. As can be observed from Model 2, degree has a positive but non-significant effect on PAL. Hypothesis 1 is therefore rejected.

Density, however, has a negative, significant effect. This contradicts our Hypothesis 2. We expected a positive and significant effect of density on PAL. The findings illustrate that when the ego network of a participant becomes denser, his/her physical activity level decreases.

Peer PAL has a positive and significant effect. This shows that if the average PAL of a participant's friends increases, the PAL of the participant himself will also increase. Therefore Hypothesis 3 is confirmed.

Model 3 shows that the interaction of degree with peer PAL is significant and positive. This confirms Hypothesis 4. The significant interaction effect is also shown in Figure 1. The amount of friends together with a high average level of physical activity of one's friends is associated with an increase in a person's PAL.

The interaction of density and peer PAL is negative but non-significant. We therefore reject Hypothesis 5.

Table 2. Correlations

| | Mean | Std. Dev. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-------------------|-------|-----------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1. PAL | 1.65 | 0.25 | 1.11 | 3.83 | | | | | | |
| 2. Prior PAL | 1.65 | 0.25 | 1.11 | 3.83 | 0.78 | | | | | |
| 3. Gender | 1.54 | 0.50 | 1.00 | 2.00 | -0.17 | -0.18 | | | | |
| 4. Community Week | 6.01 | 3.39 | 1.00 | 14.00 | -0.02 | 0.00 | -0.07 | | | |
| 5. Logins | 2.97 | 4.38 | 0.00 | 95.00 | 0.33 | 0.25 | 0.02 | -0.26 | | |
| 6. BMI | 28.49 | 6.55 | 11.67 | 70.51 | -0.16 | -0.16 | 0.06 | -0.05 | 0.00 | |
| 7. Degree | 7.88 | 7.67 | 2.00 | 56.00 | 0.01 | 0.02 | 0.16 | 0.15 | 0.04 | 0.05 |
| 8. Density | 0.52 | 0.34 | 0.00 | 1.00 | -0.05 | -0.02 | -0.07 | -0.04 | -0.06 | 0.00 |
| 9. Peer PAL | 1.65 | 0.16 | 1.11 | 2.90 | 0.18 | 0.17 | -0.09 | -0.05 | 0.08 | -0.10 |
| | | | | | | | | | | -0.07 |
| | | | | | | | | | | 0.01 |

Everything above $r=0.00$ is significant at $P<0.05$

Table 3. Panel Regression Model of Physical Activity Level (PAL)

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|----------------------------|---------------------|-----------------------|---------------------|---------------------|---------------------|
| Constant | 0.47 *** (-0.02) | 0.37 *** (-0.02) | 0.48 *** (0.03) | 0.48 *** (0.03) | 0.38 *** (0.04) |
| <i>Control Variables</i> | | | | | |
| Prior PAL | 0.74 *** (-0.01) | 0.74 *** (-0.01) | 0.74 *** (-0.01) | 0.74 *** (-0.01) | 0.73 *** (0.01) |
| Gender | -0.02 *** (0.00) | -0.02 *** (0.00) | -0.02 *** (0.00) | -0.02 *** (0.00) | -0.02 *** (0.00) |
| Community Week | 0.00 ** (0.00) | 0.00 * (0.00) | 0.00 *** (0.00) | 0.00 * (0.00) | 0.00 *** (0.00) |
| Logins | 0.01 *** (0.00) | 0.01 *** (0.00) | 0.01 *** (0.00) | 0.01 *** (0.00) | 0.01 *** (0.00) |
| BMI | 0.00 *** (0.00) | 0.00 *** (0.00) | 0.00 *** (0.00) | 0.00 *** (0.00) | 0.00 *** (0.00) |
| <i>Community Variables</i> | | | | | |
| Degree | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) |
| Density | -0.01 (0.00) | ** (0.00) (0.00) | -0.01 ** (0.00) | 0.01 ** (0.04) | 0.01 ** (0.04) |
| Peer PAL | 0.07 (0.00) | *** (-0.01) (0.00) | 0.07 *** (-0.01) | 0.07 *** (0.02) | 0.08 ** (0.02) |
| <i>Interactions</i> | | | | | |
| Degree x Peer PAL | | 0.00 * (0.00) | 0.00 * (0.00) | 0.00 * (0.03) | 0.00 * (0.03) |
| Density x Peer PAL | | | | -0.00 (0.03) | -0.03 (0.03) |
| Wald $\chi^2 =$ | 20327.56 | 20479.78 | 20486.93 | 20478.06 | 20485.20 |

Note. N = 11893 observations of a total of 1674 adults. Standard Errors in parentheses. Two tailed tests (one tailed for directional hypotheses)

* $P < .05$

** $P < .01$

*** $P < .001$

Table 4. Panel Regression Model of Physical Activity Level(PAL) - Complete Dataset

| | Model 1 |
|-----------------|------------------|
| Constant | 0.64 *** (0.05) |
| Prior PAL | 0.63 *** (0.00) |
| Gender | -0.02 *** (0.00) |
| Logins | 0.01 *** (0.00) |
| BMI | 0.00 *** (0.00) |
| Community | 0.01 *** (0.00) |
| Wald $\chi^2 =$ | 29000.05 |

N = 29516 observations of total 4423 adults.

Standard Errors in parantheses.

*P<0.05

**P<0.01

***P<0.001

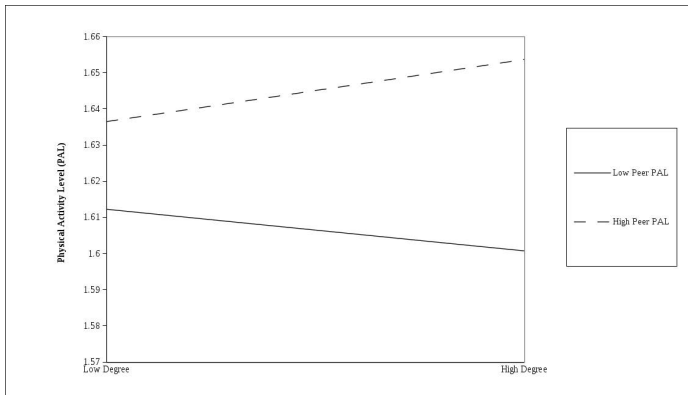


Fig. 1. Interaction Degree and Peer PAL

5 Discussion

The purpose of this study was to examine the relationship between an online social network and a person’s behavior. We can conclude that there is a positive relation between the online community and the physical activity level that participants display offline. The online community therefore matters. Offline behavior is mainly driven by the behavior of a person’s friends in the online network. People want to conform to the norms of the group and therefore they might be motivated to change their behavior [35]. We found a positive effect of the behavior of people’s friends and people’s own behavior. Additionally, we found a significant interaction effect of degree and the behavior of friends being: Participants benefit more from the program when they build relationships with people that are physically active. However, the direct effect of degree was not significant. This indicates that having a lot of friends does not matter per se, but that the amount of friends only matters in combination with the behavior of the surrounding friends.

We also formulated a hypothesis on the influence of a dense network on the behavior of people. However, we needed to reject Hypotheses 2 and 5. Density turned out to have a significant, but negative effect on the behavior of people. This contradicts our expectations which were that density would have a positive influence on physical activity because of social control within the network and because of the stronger feeling of belongingness to the network. Although we expected this social control to be positive for the behavior of the person, it could also be that the person experienced too much pressure from the network leading to a reduction in motivation. However, further research is needed to examine the mechanisms through which density might negatively influence the behavior of people. The interaction effect of density with friends' behavior is also not significant. We expected that a dense network together with a positive behavior of friends would influence a person's own behavior. This, however, was not the case in this study.

Practical Implications

According to our analyses, it is clear that when participants take the opportunity to build online relations with other participants, positive effects for physical activity levels can be expected. The behavioral outcomes of people who participated in the community online were significantly more positive than the results of people who did not participate in the community. It is thus recommended to have online community features added to activity programs. Moreover, the average physical activity of a person's friends seems to be an important network predictor for a person's own physical activity increase. In fact, being connected with a large group of active people can really make a difference for the physical activity level of a participant. Therefore, we also suggest that participants should form or be supported to establish online relations with successful, active, participants from the community. Our findings might also be relevant and can be applied at other online health promotion programs, such as those focusing on weight or stress management.

Strengths, Limitations and Further Research

Our study suggests that participants in the program benefit from building online relationships with people that are successful in the program. However, some may argue that there is a concern for causality: Successful people might also like to connect to others that are successful. We actually can make a difference between selection and influence in this research since we have longitudinal data with the increase in PAL as a dependent variable. Hence, we observe how the difference in PAL between two weeks is caused by active friends. Moreover, the selection argument stating that people chose their friends based on their behavior instead of adapting their behavior to that of their friends [24] is not applicable in this research since participants do not have insight into the activity level of their peers until they decide to connect to them. This allows us to conclude that an increase in the average physical activity level of a person's friends has a positive influence on this person's increase in physical activity from one week to the next. However, it could be that certain personality traits - for which we do not control - cause people to opt-in for the community. Future research should take this into account.

In additional analyses (available upon request) we also tested the possibility that homophily might play a role. Specifically, we tested whether active people that have active friends become even more physically active over time. However, we could not detect such an effect meaning that in our context, actors are not more likely to be influenced by alters who are similar to them [13]. From this we can conclude that being connected to active alters is important, also when participants are not active themselves.

Although we describe the ways in which degree, density and behavior of friends in the network can influence the behavior of a person, further research should focus on mediating mechanisms in these relations. The degree of social identity with the group could be of influence on someone's behavior, since this could mediate the relation between the community structure and physical activity. While the regression coefficients show small effect sizes, some of them are, however, significant. Independent of effect size, our results thus provide evidence of the influence of social networks on people's physical activity behavior. Moreover, it can be expected that the effects are even stronger when investigating the effect of offline social networks on physical activity levels. An interesting direction for future research is therefore also to take the influence of people's offline networks on their online networks and physical activity level into account.

Finally, in the program that we investigated, participants have the choice to opt-in to the community. A follow-up study could also focus on what happens when everybody is by default part of the community compared to a program that does not have a community feature. These limitations notwithstanding, our research illustrates how relationships in online social communities are an important facilitator of offline physical activity.

References

1. Adler, P.S., Kwon, S.: Social capital: Prospects for a new concept. *Academy of Management Review* 27, 17–40 (2002)
2. Ahuja, G.: Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative Science Quarterly* 45, 425–455 (2000)
3. Ajzen, I., Madden, T.J.: Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *Journal of Experimental Social Psychology* 22(5), 453–474 (1986)
4. Berkman, L.F., Glass, T., Brissette, I., Seeman, T.E.: From social integration to health: Durkheim in the new millennium. *Social Science & Medicine* 51, 843–857 (2000)
5. Bish, C.L., Blanck, H.M., Serdula, M.K., Marcus, M., Kohl, H.W., Kahn, L.K.: Diet and physical activity behaviors among americans trying to lose weight: 2000 behavioral risk factor surveillance system. *Obesity Research* 13, 596–607 (2005)
6. Bonomi, A.G., Plasqui, G., Goris, A.H., Westerterp, K.R.: Obesity. *Silver Spring* 18, 1845–1851 (2010)
7. Borgatti, S.P., Everett, M.G., Freeman, L.C.: Ucinet for windows: Software for social network analysis. Analytic Technologies, Harvard (2002)
8. Brass, D.J., Galaskiewicz, J., Greve, H.R., Tsai, W.: Taking stock of networks and organizations: A multilevel perspective. *Academy of Management Journal* 47(6), 795–817 (2004)

9. Burke, M., Marlow, C., Lento, T.: Feed me: Motivating newcomer contribution in social network sites. In: ACM CHI 2009: Conference on Human Factors in Computing Systems, pp. 945–954 (2009)
10. Burt, R.S.: Structural holes. Harvard University Press, Cambridge (1992)
11. Carr, L.J., Bartee, R.T., Dorozynski, C., Broomfield, J.M., Smith, M.L., Smith, D.T.: Internet-delivered behavior change program increases physical activity and improves cardiometabolic disease risk factors in sedentary adults: results of a randomized controlled trial. *Preventive Medicine* 46, 431–438 (2008)
12. Centola, D.: The spread of behavior in an online social network experiment. *Science* 329(5996), 1194–1197 (2010)
13. Centola, D.: An experimental study of homophily in the adoption of health behavior. *Science* 334(6060), 1269–1272 (2011)
14. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. *New England Journal of Medicine* 357, 370–379 (2007)
15. Courneya, K.S.: Understanding readiness for regular physical activity in older individuals: An application of the theory of planned behavior. *Health Psychology* 14, 80–87
16. Deci, E.L., Ryan, R.M.: Promoting self-determined education. *Scandinavian Journal of Educational Research* 38, 3–14 (1994)
17. Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication* 12(4) (2007)
18. Eyster, A.A., Brownson, R.C., Donatelle, R.J., King, A.C., Brown, D., Sallis, J.F.: Physical activity social support and middle- and older-aged minority women: results from a us survey. *Social Science & Medicine* 49, 781–789 (2001)
19. Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., Stern, A.: Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ* 328, 1166 (2004)
20. Friedkin, N.E.: A Structural Theory of Social Influence. *Structural Analysis in the Social Sciences*. Cambridge University Press (2006)
21. Haynie, D.L.: Delinquent Peers Revisited: Does Network Structure Matter? *The American Journal of Sociology* 106(4), 1013–1057 (2001)
22. Martin, A.J., Dowson, M.: Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research* 79, 327–365 (2009)
23. McNeill, L.H., Kreuter, M.W., Subramanian, S.V.: Social environment and physical activity: A review of concepts and evidence. *Social Science & Medicine* 63, 1011–1022 (2006)
24. McPherson, M., Smith-Lovin, L., Cook, J.M.: Bird of a feather: Homophily in social networks. *Annual Review of Sociology* 27, 415–444 (2001)
25. U.S. Department of Health and Human Services. Physical activity and health: A report of the Surgeon General. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic disease Prevention and Health Promotion
26. Pate, R.R., Pratt, M., Blair, S.N., et al.: Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine. *JAMA* 273(5), 402–407 (1995)
27. Plasqui, G., Westerterp, K.R.: Physical Activity Assessment With Accelerometers: An Evaluation Against Doubly Labeled Water. *Obesity* 15(10), 2371–2379 (2007)
28. Podolny, J.M., Baron, J.N.: Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review* 62, 673–693 (1997)

29. Provan, K.G., Fish, A., Sydow, J.: Interorganizational networks at the network level: A review of the empirical literature on whole networks. *Journal of Management* 33, 479–516 (2007)
30. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 68–78 (2000)
31. Scott, J.: *Social Network Analysis: A Handbook*. Sage, Thousand Oaks (2000)
32. Sheeran, P.: Intention behavior relations: A conceptual and empirical review. *European Review of Social Psychology* 12(1), 1–36 (2002)
33. Sparrowe, R.T., Liden, R.C., Wayne, S.J., Kraimer, M.L.: Social networks and the performance of individuals and groups. *Academy of Management Journal* 44(2), 316–325 (2001)
34. Strecher, V.J., Seijts, G.H.: Goal setting as a strategy for health behavior change. *Health Education & Behavior* 22, 190–200 (1995)
35. Terry, D.J., Hogg, M.A.: Group norms and the attitude-behavior relationship: A role for group identification. *Personality and Social Psychology Bulletin* 22, 776–793 (1996)
36. Vallerand, R.J.: *Intrinsic and Extrinsic Motivation in Sport and Physical Activity: A Review and a Look at the Future*. John Wiley & Sons, Inc., Hoboken (2007)
37. Van Ryzin, M.J., Gravely, A.A., Roseth, C.J.: Autonomy, belongingness, and engagement in school as contributors to adolescent psychological well-being. *Journal of Youth Adolescence* 38, 1–12 (2009)
38. Vandelanotte, C., Spathonis, K.M., Eakin, E.G., Owen, N.: Website-delivered physical activity interventions: A review of the literature. *American Journal of Preventive Medicine*. 33(1), 54–64 (2007)
39. Voorhees, C.C., Murray, D., Welk, G., et al.: The role of peer social network factors and physical activity in adolescent girls. 29, 183–190 (2005)

Detecting Overlapping Communities in Location-Based Social Networks

Zhu Wang¹, Daqing Zhang², Dingqi Yang², Zhiyong Yu², and Xingshe Zhou¹

¹ Northwestern Polytechnical University, 710072 Xi'an, China

zhu.wang@mail.nwpu.edu.cn, zhouxs@nwpu.edu.cn

² Institut TELECOM SudParis, 91000 Evry, France

{daqing.zhang,dingqi.yang,zhiyong.yu}@it-sudparis.eu

Abstract. With the recent surge of location-based social networks (LBSNs, e.g., Foursquare, Facebook Places), huge amount of digital footprints about users' locations, profiles as well as their online social connections become accessible to service providers. Different from social networks (e.g., Flickr, Facebook) which have explicit groups for users to subscribe or join, LBSNs usually have no explicit community structure. In order to capitalize on the large number of potential users, quality community detection approach is needed so as to enable applications such as direct marketing, group tracking, etc. The diversity of people's interests and behaviors when using LBSNs suggests that their community structures overlap. In this paper, based on the user-venue check-in relationship and user/venue attributes, we come out with a novel multi-mode multi-attribute edge-centric co-clustering (*M² Clustering*) framework to discover the overlapping communities of LBSNs users. By employing inter-mode/intra-mode features, the proposed framework is able to group like-minded users from different social perspectives. The efficacy of our approach is validated by intensive empirical evaluations using the collected Foursquare dataset of 266,838 users with 9,803,764 check-ins over 2,477,122 venues worldwide.

Keywords: Community Detection, Overlapping Community, Edge-Clustering, Location-Based Social Networks.

1 Introduction

With the wide adoption of GPS-enabled smart phones, location-based social networks (LBSNs) have been experiencing increasing popularity, attracting millions of users. In LBSNs, users can explore places, write reviews, upload photos, and share location and experiences with others. Check-ins are performed at physical locations (i.e., venues), such as universities, monuments, or bars. The soaring popularity of LBSNs has created opportunities for understanding collective user behaviors on a large scale, which are capable of enabling many applications, such as direct marketing, trend analysis, group search and tracking.

One fundamental task in social network analysis is to identify social subgroups (communities) for users. A *community* is typically thought of as a group of users

with more and/or better interactions amongst its members than between its members and the remainder of the network [1,2]. However, unlike social networks (e.g., Flickr, Facebook) which provide explicit groups for users to subscribe or join, the notion of community in LBSNs is not well defined. In order to capitalize on the huge number of potential users, quality community detection approach is needed.

It has been well understood that people in a real social network are naturally characterized by multiple community memberships. For example, a person usually belongs to several social groups like family, friends and colleges; a researcher may be active in several areas. Thus, it is more reasonable to cluster users into overlapping communities rather than disjoint ones. Most of the community detection approaches proposed so far are based on structural features (e.g., links) [3], but the structural information of online social networks is often sparse and weak, thus it is difficult to detect interpretable overlapping communities by considering only network structural information [4]. Fortunately, LBSNs provide rich information about the user and venue through check-ins. Those information makes it possible to cluster users with different preferences and interests into different communities in LBSNs. In reality, for some applications (e.g., advertising and marketing) it is important to group users based on both their interests as well as their social links with others.

By leveraging both network structural information (inter-mode) as well as node attributes (intra-mode) to detect communities, we can naturally obtain communities with richer and interpretable information, even though it is a highly challenging task. Classical co-clustering is one way to conduct this kind of community partitioning [5]. However, the identified communities are disjoint which contradicts with the actual social setting, where each user can belong to several communities. *Edge-Clustering* has been proposed to detect communities in an overlapping manner [6], but it did not take intra-mode features into consideration and therefore cannot be directly applied to LBSNs.

In summary, the main contributions of this work are:

- We formulate the overlapping community detection problem in LBSNs as a co-clustering issue which considers both the user-venue check-in network structure as well as attributes of users and venues. To the best of our knowledge, this work is the first attempt addressing the overlapping community detection problem in LBSNs. Specifically, we detect overlapping communities from an edge-centric perspective.
- We represent users and venues in LBSNs as two types of modes (nodes), and select both inter-mode and intra-mode features for co-clustering, while existing multi-mode clustering methods mainly concern the inter-mode features. We show that different perspectives of social communities can be revealed by introducing different intra-mode features.

The rest of this paper is structured as follows. Section 2 presents the related work. Section 3 formally defines the multi-mode multi-attribute overlapping community detection problem. The proposed community co-clustering framework is presented in Sections 4, followed by empirical study and experimental evaluation

in Section 5 and 6. We conclude our work and discuss possible future directions in Section 7.

2 Related Work

In this section, we briefly review the related work which can be classified into three categories.

The first category contains the research on understanding the collective user behaviors based on LBSNs dataset. Scellato et al. [7] analyzed the social, geographic and geo-social properties of four social networks (BrightKite, Foursquare, LiveJournal and Twitter). Noulas et al. [8] investigated the user check-in dynamics and the presence of spatio-temporal patterns in Foursquare. Cheng et al. [9] studied the mobility patterns of Foursquare users and revealed the factors affecting people’s mobility. Vasconcelos et al. [10] analyzed how Foursquare users exploited three features (i.e., tips, dones and to-dos) to uncover different behavior profiles. Only two studies aimed at uncovering community structures in LBSNs. Li et al. [11] proposed two different clustering approaches to identify user behavior patterns on BrightKite. One approach exploited the update (i.e., check-ins, photos and notes) of users to classify them into four disjoint groups according to their mobility. The second approach clustered users based on attributes such as total number of updates, social features and mobility characteristics, and led to the identification of five disjoint groups. The second study was performed on Foursquare. Noulas et al. [12] used a spectral clustering algorithm to group users based on the categories of venues they had checked in, aiming at identifying communities and characterizing the type of activity in each region of a city. Although the aforementioned studies offer important insights into properties of user interactions in LBSNs, none of them worked on *overlapping community* detection using network links and node attributes. Our work aims to fill in this gap by discovering overlapping communities.

The second category involves the work on community detection which is a classical task in complex network analysis [12][13][14]. A *community* is typically thought of as a group of nodes with more and/or better interactions amongst its members than between its members and the remainder of the network [12]. To extract such sets of nodes, one typically chooses an objective function that captures the intuition of a network cluster as set of nodes with better internal connectivity than external connectivity, and then one applies approximation or heuristics algorithms to extract node clusters by optimizing the objective function. In general, community detection methods can be classified into two types: overlapping and non-overlapping methods. Some popular methods are modularity maximization [13][14], Girvan-Newman algorithm [1], Louvain algorithm [15], clique percolation [16], link communities [17], etc. As users in LBSNs have rather weak and sparse relations [18], one cannot naively apply community detection based solely on the links found in these social networks and expect to generate interpretable communities.

The third category focuses on community detection by considering both link and node attributes for social networks, which are the closest to our work. Several

existing works on attributed graph clustering fall into this category. The main idea is to design a distance/similarity measure for vertex pairs that combines both structural and attribute information of the nodes. Based on this measure, standard clustering algorithms like K-Medoids and spectral clustering are then applied to cluster the nodes. A weighted adjacency matrix is used as the similarity measure in [19], where the weight of each edge is defined as the number of attribute values shared by the two end nodes. The authors applied graph clustering algorithms on the constructed adjacency matrix to perform clustering. The state-of-the-art distance-based approach is the SA-Cluster proposed by Zhou et al [20] which defined a unified distance measure to combine structural and attribute similarities. Attribute nodes and edges are added to the original graph to connect nodes which share attribute values, and a neighborhood random walk model is used to measure the node closeness on the augmented graph. Afterwards, a clustering algorithm SA-Cluster is proposed based on the K-Medoids method. However, all these works in the last category attempted to optimize two contradictory objective functions and intended to identify disjoint communities, thus the communities detected were not optimal and had no clear semantic meanings.

In this work, we propose to leverage both the structure links between users and venues as well as their attributes to discover the overlapping community structure. Specifically, we formulate the overlapping community detection problem into an multi-mode multi-attribute edge-centric co-clustering issue, viewing both inter-mode links and intra-mode attributes as unified features for clustering. With this novel representation, users, venues together with their attributes are grouped in a natural way.

3 Problem Statement

In this paper, a community is defined as a cluster of edges (check-ins) with user and venue as two modes, where the common attributes of users and venues characterize the properties of the community. We use $U = (u_1, u_2, \dots, u_m)$ to represent the user set, and $V = (v_1, v_2, \dots, v_n)$ to denote the venue category set, a community $C_i (1 \leq i \leq k)$ is a subset of users and venue categories, where k is the number of communities. On one hand, the check-in relationship between users and venue categories form a matrix M , where each entry $M_{ij} \in [0, \infty)$ corresponds to the number of check-ins that u_i has performed over v_j . Therefore, each user can be represented as a vector of venue categories, and each venue category can be denoted as a vector of users. On the other hand, users and venue categories might have several independent attributes, denoted as $(a_{i1}, a_{i2}, \dots, a_{ix})$, and $(b_{j1}, b_{j2}, \dots, b_{jy})$ respectively. Normally, every attribute reveals a certain social aspect of users or venue categories. For instance, a user has a certain number of followers and followings, and a venue category has a common operating time. Therefore, both user mode and venue mode have two types of representations: an inter-mode representation as well as an intra-mode representation. Based on the above notations, the overlapping community detection problem in LBSNs can be formulated as a multi-mode multi-attribute edge-centric co-clustering issue, which will be fully exploited in the next section.

4 Multi-mode Multi-attribute Edge-Centric Co-clustering Framework

The observation that a check-in on LBSNs reflects a certain aspect of the user’s preferences or interests enlightens us to cluster edges instead of nodes, as the detected clusters of check-ins will naturally assign users into overlapping communities with connections to venues. Specifically, after obtaining edge clusters, overlapping communities of users can be recovered by replacing each edge with its vertices, i.e., a user is involved in a community as long as any of her check-ins falls into the community. In such a way, the obtained communities are usually highly overlapped. The key idea of the proposed framework is shown in Fig. 1.

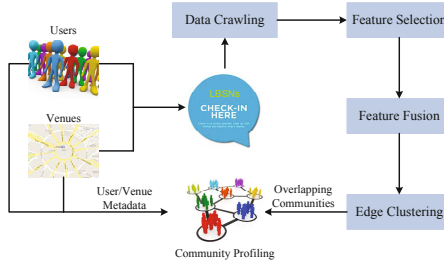


Fig. 1. Community discovering framework

As indicated in Fig. 1, we first select features based on the characteristics of the collected LBSNs dataset and then perform feature normalization and fusion. Second, the overlapping community structure is detected by using the proposed edge-centric co-clustering algorithm M^2 Clustering. Due to space limitation, in this paper we mainly focus on the feature selection and community detection, leaving the detailed elaboration of community profiling as a future work.

4.1 Edge-Centric Co-clustering

As stated in the introduction section, we define a community in LBSNs as a group of users who are more similar with users within the group than users outside the group. Therefore, communities that aggregate similar users and venues together should be detected by maximizing intra-cluster similarity rather than maximizing modularity. This objective function is formulated as:

$$Obj = \arg \max_C \sum_{j=1}^k \sum_{e_c \in C_j} sim(e_c, C_j), \quad (1)$$

where k is the number of communities, $C = \{C_1, C_2, \dots, C_k\}$ is the detected community set, e_c denotes an edge of community C_j , and $sim(e_c, C_j)$ is the

similarity between e_c and C_j . With this formulation, the key is to characterize the similarity between an edge and a community. To this end, we first introduce the definition of edge similarity.

In a user-venue check-in network, each edge is associated with a user vertex and a venue vertex. By taking an edge-centric view, each edge is treated as an instance with its two vertices as features. In other words, the similarity between a pair of edges can be defined as the similarity between the corresponding user pair and venue pair as:

$$sim_{edge}(e_i, e_j) = F(sim_u(u_i, u_j), sim_v(v_i, v_j)), \quad (2)$$

where $sim_u(u_i, u_j)$ is the similarity between two users, $sim_v(v_i, v_j)$ is the similarity between two venues, and F represents the function used to combine these two similarities, by balancing the weights of the user mode and the venue mode. The formalism of F depends on the characteristics of the expected communities as well as the targeted applications. Considering the similarity trade-off between user mode and venue mode, two widely used formalisms of F are $(sim_u + sim_v)/2$ and $\sqrt{sim_u \times sim_v}$. In this work, we adopt the second notion to ensure that a pair of edges are of high similarity *if and only if* they are of high similarity in both user-mode and venue-mode.

Meanwhile, since each community contains a set of edges, based on Equation 2, the similarity between an edge and a community can be calculated.

Inter/Intra-mode Features

The inter-mode feature describes the structure similarity between a pair of edges based on the physical links between users and venues, and the intra-mode feature depicts attributes similarity where each attribute corresponds to a certain social aspect of users or venues. As we have mentioned, in many real applications, both inter-mode links and intra-mode attributes are important.

Considering the characteristic of user-venue links (mainly check-ins) in LBSNs, we study two inter-mode features: 1) characterizing a user based on a vector of venue categories, namely *user-venue similarity*; 2) characterizing a venue category by using a vector of users, which is defined as *venue-user similarity*.

Meanwhile, by analyzing the available user/venue related metadata in LBSNs, we identify three intra-mode features which are *user social-status similarity*, *user geo-span similarity* and *venue temporal similarity*. All the above mentioned features will be presented in detail in the empirical study section.

Feature Fusion

Due to the characteristic of various similarity features, different calculation methods might be used which lead to different value ranges. Therefore, the absolute values of different features must be normalized. To this end, we simply normalize each similarity measure sim_x into the interval $[0, 1]$ as $sim'_x = \{sim_x - \min(sim_x)\} / \{\max(sim_x) - \min(sim_x)\}$, where sim'_x is the normalized format of measure sim_x .

Afterwards, another issue is to fuse different features. Considering that each edge consists of two nodes, we first defined user similarity and venue similarity as:

$$sim_{u/v} = \frac{1}{|f_{u/v}|} \sum sim'_{u/v*}, \quad (3)$$

where $|f_u|$ and $|f_v|$ represent the number of selected features for user-mode and venue-mode, respectively; sim'_{u*} and sim'_{v*} refer to the normalized similarity. Then, based on Equation 2, the edge similarity is calculated as $sim_{edge} = \sqrt{sim_u \times sim_v}$.

Clustering Algorithm

Based on the above formulation, the multi-mode multi-attribute edge clustering problem is converted into an ordinary clustering issue, which can be handled by adjusting k-means as follows:

- While k-means selects the mean (i.e., geometric center) of all the instances (i.e., edges) in a cluster as its centroid, we represent each centroid by using the whole set of instances within the cluster. According to the definition of the similarity between an edge and a cluster in Equation 2, if a set of multi-mode multi-attribute edges are denoted by a single vector, the obtained similarity will be significantly different.
- The similarity between a given pair of instances is not directly calculated but based on the similarity between the corresponding pair of vertices. As each edge includes two vertices and each vertex consists of multiple attributes which are usually represented as feature vectors of different dimensions (i.e., length), it is difficult to define a unified distance measure to characterize the similarity between a pair of multi-mode multi-attribute edges.
- While representing each centroid as a set of instances ensures the precision of the obtained similarity, the computation complexity increases from $O(k \times N)$ to $O(N^2)$. To improve the time efficiency, each centroid C_j is denoted as a structure which consists of four components: a list of current instances within the centroid (E_{C_j}), a list of instances that are assigned to the centroid during last iteration (E_{A,C_j}), a list of instances that are removed from the centroid during last iteration (E_{R,C_j}), and the similarity array between the previous centroid and the whole set of instances ($sim(E_{P,C_j}, E)$).

Based on the above adjustments, the proposed k-means based clustering method is presented in Algorithm 1. At the beginning, k edges are randomly selected (line 1) based on which a set of initial centroids are constructed (lines 2-4). Afterwards, during the iteration, given a centroid C_j we compute the similarity that each edge e_i has obtained and the similarity it has lost (line 10) during the last reassignment, based on which the current similarity between e_j and C_j is calculated (line 11). Edge e_i is assigned to the centroid that is most similar to itself, and the corresponding similarity is marked as $maxsim_i$ (lines 12-14). Centroid updating is performed based on the reassignment of edges (line 17).

At the end of each iteration, the current value of the objective function is calculated (Obj_{cur} , line 24) to compare with the previous value Obj_{pre} (line 19). The iteration terminates if and only if the absolute difference between these two values is smaller than the predefined threshold ϵ (line 20). Experiments based on our dataset show that the algorithm usually converges within 100 iterations.

Algorithm 1. M^2 Clustering Algorithm

Input: E , an edge list $\{e_i | 1 \leq i \leq n\}$; k , the number of communities; M_u , the user-user similarity matrix; M_v , the venue-venue similarity matrix;

Output: C , a set of detected communities;

- 1 k edges are randomly selected $\{e_j | 1 \leq j \leq k\}$;
- 2 **for** each e_j **do**
- 3 $E_{C_j} = \{e_j\}$; $E_{A,C_j} = E_{C_j}$; $E_{R,C_j} = \emptyset$; $sim(E_{P,C_j}, E) = \text{zeros}(|E|)$;
- 4 **end**
- 5 $\{maxsim_i | 1 \leq i \leq n\} = 0$;
- 6 **repeat**
- 7 $Obj_{pre} = \sum maxsim_i$; reset $\{maxsim_i\}$;
- 8 **for** each C_j **do**
- 9 **for** each e_i in E **do**
- 10 calculate $sim(E_{A,C_j}, e_i)$; calculate $sim(E_{R,C_j}, e_i)$;
- 11 $sim(E_{C_j}, e_i) = sim(E_{P,C_j}, e_i) + sim(E_{A,C_j}, e_i) - sim(E_{R,C_j}, e_i)$;
- 12 **if** $sim(E_{C_j}, e_i) > maxsim_i$ **then**
- 13 $maxsim_i = sim(E_{C_j}, e_i)$; assign e_i to C_j ;
- 14 **end**
- 15 **end**
- 16 **end**
- 17 update the centroids;
- 18 $Obj_{cur} = \sum maxsim_i$;
- 19 $\Delta = \text{abs}(Obj_{cur} - Obj_{pre})$;
- 20 **until** $\Delta < \epsilon$;

5 Empirical Study Based on Foursquare

5.1 Data Collection

Foursquare API provides limited authorized access for retrieving check-in information, therefore we resort to Twitter streaming API [2] to get the publicly shared check-ins within Tweets. The data collection started from October 24th, 2011 and lasted for 8 weeks, which results in a dataset of more than 12 million check-ins performed by 720,000 users over 3 million venues. Meanwhile, we also crawled metadata related to users and venues, including every user’s Twitter profile and every venue’s Foursquare profile.

Before community detection, we pre-process the collected dataset as follows. First of all, we excluded check-ins that are performed over invalid venues. In this paper, invalid venues refer to those that cannot be resolved by Foursquare API,

and thus the detailed information of these venues is not available. Consequently, about 7.52% of the check-ins are removed from the dataset. Secondly, we only keep users who have performed at least one check-in per week on the average (referred as active users), which means inactive users together with their check-ins are excluded. Finally, users who used agent software conducting remote and large scale automatic check-ins (with a check-in speed faster than than 1,200 km/h, which is the common airplane speed) are defined as *sudden move* users [9], and check-ins from these users are eliminated. We observed a total number of 9,276 *sudden move* users, which occupy about 3.36% of the active users.

After the above data cleansing, the remained dataset includes 266,838 users and 9,803,764 check-ins which were performed over 2,477,122 venues.

5.2 Feature Description

Inter-mode Features: User-Venue Similarity

Foursquare classifies venues into 400 categories under 9 parent categories. We identify 274 venue categories by merging those similar ones, and consequently based on a user’s check-in venues, each user can be represented as a vector of 274 dimensions. We build a $266,838 \times 274$ matrix to represent all the active users within the collected dataset. Afterwards, this matrix is refined based on principal component analysis, which is able to convert a set of observation of correlated variable into a set of value of linearly uncorrelated variable under a latent space. By applying principal component analysis on the raw matrix, we obtain a $266,838 \times 100$ matrix which covers 95.62% of the total variance. After the conversion, each user is represented as a vector of 100 dimensions in the latent space.

Based on the above matrix, the *user-venue similarity* for a pair of users u_m and u_n is calculated based on cosine similarity.

Inter-mode Features: Venue-User Similarity

As we have mentioned, each venue category of Foursquare can be denoted as a vector by treating users as features as well. Following the same approach as the above section, we obtain a 274×100 matrix by performing principal component analysis on the original $274 \times 266,838$ matrix, which covers 95.34% of the total variance. As a result, each venue category corresponds to a vector of 100 dimensions in the latent space. Similarly, the *venue-user similarity* is also defined using cosine similarity.

Intra-mode Features: User Social-Status Similarity

There are two lists in each user’s Twitter profile, a follower list and a following list. In this paper, we define a user’s social status as the ratio of her number of followers to her number of followings. Specifically, the social status of a user u_m is formalized as $s_s = n_{followers}(u_m) / n_{followings}(u_m)$.

According to the above definition, users with high social status are those who have many followers and fewer followings. To some extent, these users act as *hubs* of the social network.

We introduce the first intra-user similarity feature namely *user social-status similarity* based on the user social status metric. Given a pair of users u_m and u_n , this feature is defined as $sim_{us}(u_m, u_n) = \min(s_{s,m}, s_{s,n}) / \max(s_{s,m}, s_{s,n})$, where $s_{s,m}$ and $s_{s,n}$ represent the social status of u_m and u_n respectively. Apparently, the value of user social-status similarity falls into the interval $[0, 1]$.

Intra-mode Features: User Geo-Span Similarity

The user geo-span (a.k.a. radius of gyration) is another metric that can be used to distinguish the life style of different users, which is defined as the standard deviation of distances between a user’s check-ins and her home location. In LBSNs, a user’s home location is defined as the centroid position of her most popular check-in region [21]. The user geo-span metric is able to indicate not only how frequently but also how far a user moves. Generally, a user with low radius of gyration mainly travels locally (with few long-distance check-ins), while a user with high radius of gyration has many long-distance check-ins. The formal definition for radius of gyration is as follows:

$$r_g = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i - r_h)^2}, \quad (4)$$

where n is the number of check-ins made by a user, and $r_i - r_h$ is the distance between a particular check-in location r_i and the user’s home location r_h .

By using the radius of gyration metric, we introduce the second intra-user similarity feature named *user geo-span similarity*. Specifically, for a pair of users u_m and u_n , the calculation of this feature is similar as the user social-status.

Intra-mode Features: Venue Temporal Similarity

Generally, people visit and check in different kind of venues at different time, such that different venue categories can be distinguished according to their temporal check-in patterns [22]. In this paper, we divide a week into 168 (7×24) time slots and each time slot corresponds to one hour in a certain day of the week, reflecting the temporal characteristic of each user check-in. build a weekly temporal check-in band for each venue category at the hour granularity, which means each temporal band corresponds to a vector of 168 dimensions (7×24). Since we have identified 274 venue categories, a 274×168 matrix is constructed and then principal component analysis is performed on this matrix, producing a new matrix of 274×20 which covers 99.92% of the total variance. Consequently, *the venue temporal similarity* between a pair of venues can also be defined based on cosine similarity.

6 Performance Evaluation

6.1 Experiment Setup and Overall Design

To evaluate the performance of the proposed framework, we chose three big cities (i.e., Paris, New York and Tokyo) as the target societies. We first calculated the

home location of all the active users, then a set of users for each city are selected based on the distance between their home locations and the geometric center of the corresponding city. Specifically, we set the distance threshold as 10km, yielding 1,432, 3,503, and 2,674 users for Paris, New York and Tokyo, respectively. Afterwards, all the check-ins produced by these users during the data collection period are extracted, resulting 49,160, 108,451 and 120,494 check-ins, respectively. Meanwhile, all the inter-mode and intra-mode features used in the experiments are calculated. Based on the dataset of these three cities, we mainly conducted experiments to evaluate the quality of the detected communities indirectly by calculating the intra-community tip similarity.

6.2 Benchmark

In this work, we conduct a series of experiments to evaluate the performance of the proposed community detection mechanism *M²Clustering*. Specifically, we adopt *Edge-Clustering* [6] as the baseline, which is a state-of-the-art overlapping community detection method.

Table 1. Different community detection methods evaluated in the experiments

| Method | Description |
|-------------------------------|--|
| Edge-Clustering | Used as the baseline method. |
| M ² Clustering-I | The first format of M ² Clustering, which uses not only two inter-mode features (i.e., User-Venue Similarity, and Venue-User Similarity) but also the venue-mode feature (i.e., Venue Temporal Similarity). |
| M ² Clustering-II | The second format of M ² Clustering, which uses not only two inter-mode features but also two user-mode features (i.e., User Geo-Span Similarity, and User Social-Status Similarity). |
| M ² Clustering-III | The last format of M ² Clustering, which uses all the two inter-mode and three intra-mode features introduced in this paper. |

6.3 Co-clustering Results

Since the Foursquare data we use does not have the ground truth [23] about the real communities, we resort to indirectly evaluating the proposed framework. Intuitively, users belonging to the same community tend to share similar interests, hopefully they also share more common topics in their tips. Therefore, we attempt to evaluate the proposed community detection framework by testing whether the tips that posted by the same community are also of high similarity, indirectly showing the effectiveness of the proposed community detection mechanism.

In this paper, we define the average similarity among tips within a community as *community tip similarity*. Intuitively, a quality community detection method should achieve high *community tip similarity*, even though the tip information has not been leveraged when clustering communities. Particularly, a tip t_k , which is left by user u_m at venue category v_n , falls into community C_j if and only if there is an edge e_{u_m, v_n} that belongs to C_j .

To compute the similarity between a pair of tips, we first project each tip to a latent topic space by using Latent Dirichlet Allocation (LDA), which is able to mine higher level representations (i.e., topics) from a collection of documents [24]. Specifically, LDA helps to explain the similarity of tips by grouping tips into topics. A mixture of these topics then constitute the observed tips. We use MALLET [25] to obtain the topic representation of each tip. Suppose that tips are grouped into N_T topics, then a tip t_k can be formally represented as a topic vector $\langle tv_1, tv_2, \dots, tv_i, \dots, tv_{N_T} \rangle$, where tv_i is equal to the number of words in t_k that are projected to the i^{th} topic. Consequently, the *community tip similarity* can be defined by using cosine similarity.

In order to conduct the experiments, we first retrieve the tips that are left at the 2,477,122 venues in our dataset, and get a collection of more than 6 million tips. Afterwards, non-English tips are filtered out which leads to 369,083 tips in English contributed by 66,843 users over 228,514 venues. Without loss of generality, we set the number of topics as 100 in the experiment.

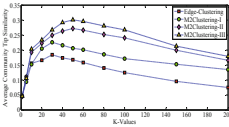


Fig. 2. Average community tip similarity of different methods

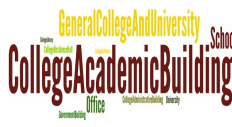


Fig. 3. College Community



Fig. 4. Nightlife Community

Consequently, each tip can be represented as a 100 dimension topic vector. We first perform community detection using the proposed framework on the Paris dataset, and repeated experiments 10 times for each methods listed in Tab. 1. Then for each of the detected communities, we calculate its *community tip similarity*. The average community tip similarity of different methods is shown in Fig. 2.

According to Fig. 2, we have the following observation. Firstly, all the three formats of M^2 Clustering achieve higher community tip similarity than the baseline method *Edge-Clustering*. The reason should be that intra-mode features are able to introduce useful information for community clustering. Secondly, M^2 Clustering-III is the most competitive method while M^2 Clustering-II is the next most competitive one, where the two *user-mode* features have been leveraged. This indicates that users who have similar geo-spans and social statues are most likely to discuss similar topics. Similar results can be obtained based on the Tokyo and New York dataset.

Based on this finding, we present the tag clouds of two Paris Foursquare user communities as an example, where k is set as 50 and M^2 Clustering - III is adopted. Accordingly, Fig. 3 shows a *College* community which is formed by college students or staff, and Fig. 4 presents a *Nightlife* community.

7 Conclusion

In this paper, by leveraging the user-venue check-in network and user/venue attributes, we propose a multi-mode multi-attribute edge-centric co-clustering ($M^2Clustering$) framework to detect overlapping communities for LBSNs users. Experimental results show that the proposed framework is able to better group like-minded users into communities than the state-of-the-art approach *EdgeClustering*, and the detected communities have explicit semantic meanings.

The preliminary study suggests several interesting problems that are worth exploring further. Characterizing and profiling the detected communities in a systematic manner is one direction. How to use the proposed community detection framework helping the study of friend and place recommendation mechanism is another direction.

Acknowledgment. This work is partially supported by the EU FP7 Project SOCIETIES (No. 257493), the National Basic Research Program of China (No. 2012CB316400), the Natural Science Foundation of China (No. 61103185), the Scholarship Award for Excellent Doctoral Student Granted by Ministry of Education, and the Doctorate Foundation of Northwestern Polytechnical University (No. CX201018). This work was done when Zhu Wang was with Telecom Sud-Paris, France.

References

1. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Physical Review E* 69, 26113–26127 (2004)
2. Fortunato, S.: Community detection in graphs. *Physics Reports* 486(3-5), 75–174 (2010)
3. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. *Nature* 466(7307), 761–764 (2010)
4. Cruz, J.D., Bothorel, C., Poulet, F.: Entropy based community detection in augmented social networks. In: *CASoN*, pp. 163–168. IEEE (2011)
5. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: *Proc. of KDD 2001*, pp. 269–274. ACM, New York (2001)
6. Wang, X., Tang, L., Gao, H., Liu, H.: Discovering overlapping groups in social media. In: *Proc. of ICDM 2010*, pp. 569–578 (2010)
7. Scellato, S., Mascolo, C., Musolesi, M., Latora, V.: Distance matters: geo-social metrics for online social networks. In: *Proc. of WOSN 2010*, pp. 8–8. USENIX Association, Berkeley (2010)
8. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An empirical study of geographic user activity patterns in foursquare. In: *Proc. of ICWSM 2011*, pp. 570–573. The AAAI Press (2011)
9. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring millions of footprints in location sharing services. In: *Proc. of ICWSM 2011*, pp. 81–88. The AAAI Press (2011)
10. Vasconcelos, M.A., Ricci, S., Almeida, J., Benevenuto, F., Almeida, V.: Tips, dones and todos: uncovering user profiles in foursquare. In: *Proc. of WSDM 2012*, pp. 653–662. ACM, New York (2012)

11. Li, N., Chen, G.: Analysis of a location-based social network. In: Proc. of CSE 2009, pp. 263–270. IEEE Computer Society, Washington, DC (2009)
12. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In: Proc. of ICWSM 2011, pp. 32–35. The AAAI Press (2011)
13. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Physical Review E* 70, 66111–66116 (2004)
14. Wakita, K., Tsurumi, T.: Finding community structure in mega-scale social networks. In: Proc. of WWW 2007, pp. 1275–1276. ACM, New York (2007)
15. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (10), P10008 (2008)
16. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818 (2005)
17. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. *Nature* 466(7307), 761–764 (2010)
18. Tang, L., Liu, H.: Community detection and mining in social media. *Synthesis Lectures on Data Mining and Knowledge Discovery* 2, 1–137 (2010)
19. Steinhaeuser, K., Chawla, N.V.: Community detection in a large real-world social network. In: Liu, H., Salerno, J.J., Young, M.J. (eds.) *Social Computing, Behavioral Modeling, and Prediction*, pp. 168–175. Springer US (2008)
20. Zhou, Y., Cheng, H., Yu, J.X.: Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.* 2(1), 718–729 (2009)
21. Scellato, S., Noulas, A., Lambiotte, R., Mascolo, C.: Socio-spatial properties of online location-based social networks. In: Proc. of ICWSM 2011. The AAAI Press (2011)
22. Ye, M., Janowicz, K., Mülligann, C., Lee, W.C.: What you are is when you are: the temporal dimension of feature types in location-based social networks. In: Proc. of GIS 2011, pp. 102–111. ACM, New York (2011)
23. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *PNAS* 99(12), 7821–7826 (2002)
24. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (2003)
25. McCallum, A.K.: MALLET: A Machine Learning for Language Toolkit (2002)

CrowdLang: A Programming Language for the Systematic Exploration of Human Computation Systems

Patrick Minder and Abraham Bernstein

Dynamic and Distributed Information Systems Group
University of Zurich, Switzerland
{lastname}@ifi.uzh.ch

Abstract. Human computation systems are often the result of extensive lengthy trial-and-error refinements. What we lack is an approach to systematically engineer solutions based on past successful patterns.

In this paper we present the *CrowdLang*¹ programming framework for engineering complex computation systems incorporating large crowds of networked humans and machines with a library of known interaction patterns. We evaluate *CrowdLang* by programming a German-to-English translation program incorporating machine translation and a monolingual crowd. The evaluation shows that *CrowdLang* is able to simply explore a large design space of possible problem-solving programs with the simple variation of the used abstractions. In an experiment involving 1918 different human actors, we show that the resulting translation program significantly outperforms a pure machine translation in terms of adequacy and fluency whilst translating more than 30 pages per hour and approximates the human-translated gold standard to 75%.

Keywords: *CrowdLang*, Programming Language, Human Computation, Collective Intelligence, Crowdsourcing, Translation Software.

1 Introduction

Much of the prosperity gained by the industrialization of the economy in the 18th century was the result of increased productivity after dividing work into smaller tasks performed by more specialized workers. Wikipedia, Google, and other stunning success stories show that with the rapid growth of the World Wide Web and the advancements in communication technology, this concept of Division of Labor can also be applied to knowledge work [1, 2]. These new modes of collaboration—whether they are called collective intelligence, human computation, crowdsourcing, or social computing²—are now able to routinely

¹ This work was supported in part by the Swiss National Science Foundation (SNSF-Project: 200021-143411/1). A short research note summarizing a part of the evaluations in this paper was published at the ACM WebSci Conference 2012 [12].

² A clear distinction between these concepts is an ongoing debate in the community [3, 4, 2]. Relying on [4] this paper considers *human computation* as computation that is carried out by humans and *human computation systems* as “paradigms for utilizing human processing power to solve problems that computers cannot yet solve”.

solve problems that would have been unthinkable only a few years ago by interweaving the creativity and cognitive capabilities of networked humans and the efficiency and scalability of networked machines in processing large amounts of data [5]. The advent of crowdsourcing markets such as Amazon’s Mechanical Turk (MTurk) further fosters this development. Hence, Bernstein et al. suggest that we can view these systems as constituting a kind of a “global brain” [5].

Even though a plethora of human computation systems (HCS) exists, our understanding of how to “program” these systems is still poor: human computers are profoundly different from traditional computers due to the huge motivational, error and cognitive diversity within and between humans [5]. Hence, HCSs are mostly used for parallel information processing (e.g., image labeling). These tasks share in common that they are massively (or embarrassingly) parallelizable, have a low interdependence between single assignments, and use relatively little cognitive effort. Many tasks, however, cannot be captured in this paradigm. Consider, e.g., the joint editing of lengthy texts as accomplished on Wikipedia. Here, a large number of actors work on highly interdependent tasks that would be very difficult to cast into a bulk parallelization with low interdependence. Hence, to harness the full potential of HCSs, we need powerful new programming metaphors and infrastructures that support the design, implementation, and execution of human computation. Specifically, we need a programming language that supports the whole range of possible dependencies between single tasks, allows for the seamless reuse of known human computation patterns incorporating both humans and machines to exploit prior experience, and integrates multiple possible execution platforms (e.g., micro-task markets, games with a purpose) to leverage a large ecosystem of participants. To move from a culture of “*Wizard of Oz*” techniques, in which applications are the result of extensive trial-and-error refinements, a programming language has to support the recombination [6] of interaction patterns to systematically explore the design space of possible solutions. Recent research only partially addresses these challenges by providing programming frameworks and models [7–9] for massive parallel human computation, concepts for planning and controlling dependencies [10, 11], and theoretical deductive analysis of emergent collective intelligence [2].

In this article, we present the *CrowdLang* human computation programming language and framework for interweaving networked humans and computers. *CrowdLang* supports cross-platform workforce integration, the management of human computer latency, and incorporates abstractions for group decisions, contests, and collaborative interaction patterns as proposed by Malone et al. [2]. *CrowdLang* also supports the management of arbitrary dependencies among tasks and workers, and not only asynchronous parallelization. We show *CrowdLang*’s feasibility and strength by programming a collection of text translation programs. The resulting translation programs are able to speedily translate non-trivial texts from German to English achieving a significantly better quality than pure machine translation approaches. Also, given the simple recombination of patterns supported by *CrowdLang*, we were able to unearth a novel human

computation pattern, which we call “*Staged-Contest with Pruning*,” that outperforms all other known patterns in the translation task.

2 Background and Related Work

Relevant to this paper is research about frameworks for supporting the design of HCS and the analysis of emergent collective intelligence.

A number of programming frameworks and concepts addressing the distinct challenges in engineering human computation systems have been proposed recently. Little et al. [7, 13] proposed the use of the imperative programming framework *TurKit*. Investigating workflows composed by iterative and parallel traditional programming constructs, they explored basic technical problems caused by the high latency associated with waiting for a response from a human worker when writing and debugging human computation code. They support the idea of a “crash-and-rerun” programming model, which allows a programmer to repeatedly rerun and debug processes without republishing costly previously completed human computation.

Several programming frameworks inspired by the MapReduce [14] programming metaphor have been proposed to coordinate arbitrary dependencies between interdependent tasks. These frameworks model complex problems as a sequence of partitioning, mapping, and reducing subtasks. For example, Kittur et al.’s *CrowdForge* programming framework [8] starts by breaking down large problems into discrete subtasks either by using human or machine computers. Then human or machine agents are used to collect a set of solutions. Finally, the results of multiple workers are merged and aggregated into the solution of the larger problem. Similarly, Ahmad et al.’s *Jabberwocky* programming environment [9] extended this idea by providing an additional human and resource management system for integrating workforces from different markets, as well as a high-level procedural programming language. Finally, Noronha et al. [15] suggest a “divide-and-conquer” management framework inspired by corporate hierarchies.

These studies highlight the importance of designing new environments for programming human computation systems but are restricted by the structural and synchronous rigidity of the MapReduce programming metaphor when modeling workflows with arbitrary dependencies [16]. Further, they do not provide any explicit treatment of cognitive diversity in and between human actors [5] or abstractions for complex coordination patterns such as group decision procedures [2]. Finally, they assume that computation can be fully specified ex-ante. In many complex problem-solving tasks, however, processes are difficult to specify ex-ante and only gain more specific definitions during execution or may start out as well-defined tasks and then lose their specific definition due to some unexpected exceptions. Thus, it was proposed that processes move along a *specificity frontier* from well defined and static to loosely defined and dynamic [10]. Zhang et al. [11], for example, propose a system that exploits a self-organizing crowd to solve a planning under constraints problem. This system illustrates the crowd-based solution

of a process somewhere in the middle of the specificity frontier. To harness the full potential of human computation systems, we believe that programming languages designed for this purpose should exhibit all these features.

Complementing these (empirical) explorations of possible patterns, several studies [3, 4] taxonomize various aspects of HCSs. Malone et al. [1] examined about 250 different HCSs and identified in the *Collective Intelligence Genome* the characteristics (“genes”) that can be recombined to the basic building blocks (“genome”) of human computation systems. Their conceptual classification framework suggests characterizing each building block by answering two pairs of questions. First, they considered staffing (*Who is performing the task?*) and different kinds of incentives (*Why are they doing it?*). Second, they analyzed a specific system by defining the goal of a task (*What is being done?*) and problem-solving process (*How is it being done?*). We believe that this framework is not only suitable for analyzing existing applications but also for designing new ones by recombining the basic building blocks as Bernstein et al. [6] also proposed in the context of business processes.

3 The *CrowdLang* Programming Framework

Conventional programming languages are developed to interoperate with deterministic machines. When moving from programming pure machine computation to hybrid machine-human or pure human computation systems, these languages are not a good match as they lack abstraction for dealing with the cognitive, error, and motivational diversity within and between humans [5] and the varying degrees of detail in many human task definitions.

The objective of *CrowdLang* is to build a general-purpose programming language and framework for interweaving human and machine computation within complex problem solving. *CrowdLang* intends to incorporate explicit methods for handling (cognitive, error, and motivational) diversity, complex coordination mechanisms (and not only batch processing) and abstractions for human computation tasks such as group decision processes. In a future version, the framework will also support the specificity frontier to allow unstructured, constraint-restricted computation and for run-time task decomposition as well as the modeling of non-functional constraints such as budget, completion time, or quality. Last but not least, the *CrowdLang* engine has to address the technical challenges associated with crowd worker latency (waiting for human response) [7].

The framework consists of three major components: (1) The *CrowdLang Library* simplifies the design of new human computation systems. It supports the seamless reuse of existing interaction patterns by providing an extensible programming library. The integrated *intelligent discovery assistant* supports the exploration of the whole design space through simple pattern recombination. (2) The *CrowdLang Engine* addresses the technical challenges of executing human computation algorithms by managing the crowd latency (waiting for human response), debugging human computation code, and the re-executing of human computation after exceptions. The *CrowdLang Integrator* integrates different execution platforms such as micro-task markets and games with a purpose.

4 The *CrowdLang* Programming Language

In accordance with Malone et al.’s empirical exploration [2], *CrowdLang* supports operators for task decomposition and group decision processes.

4.1 Basic Operators, Task Decomposition and Aggregation

CrowdLang provides language constructs for defining basic operators, data items, and control flow constructs (see Figure 1). A *Task* represents the transformation of a given problem statement into a solution. The transformation is performed either by humans (*Human Computation*) or machines (*Machine Computation*). A *Problem Statement* defines a task in terms of a question and the required input data. A *Solution* represents the computed results for a *Problem Statement*. A *Sequence Flow* defines the execution order of single tasks and manages therefore classical producer/consumer relationships, where the produced output of a previous task is consumed as an input in the next task.

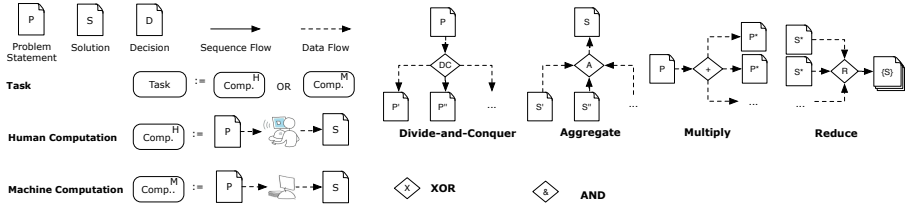


Fig. 1. Basic *CrowdLang* Operators, Routing, Aggregation, and Task Decomposition

CrowdLang provides a set of routing operators to distribute computation and aggregate results (see Figure 1). The *Divide-and-Conquer* operator decomposes a problem statement into multiple parallelizable, distinct subproblems. The *Aggregate* operator, in contrast, aggregates the results of several subtasks to a solution of the initial problem statement.

A given problem can be distributed to actors in three different ways. The *Multiply* control flow operator indicates that a given problem gets transformed multiple times in parallel. Hence, copies of the original problem statement get allocated to multiple independent actors potentially leading to different solutions (in particular when performed by human actors). Hence, the result of such an execution is a set of solutions. The *Reduce* operator takes a set of solutions and determines the “best” solution candidate employing a decision procedure. Together, the *Multiply* and *Reduce* are the building block for many parallelizing crowd computing patterns. *CrowdLang* provides the established *exclusive (XOR)* and *parallel (AND)* control flow operators. XOR is used to create or synchronize alternative paths; AND can be used to create and synchronize parallel paths.

4.2 Building Blocks of Collective Intelligence

In accordance to [2], *CrowdLang* defines a set of basic building blocks classified as *Create* and *Decide* interaction patterns.

Create Interaction Patterns. The framework defines two variations of the create interaction pattern: Collection and Collaboration.

A *Collection* occurs when actors independently contribute to a task. Malone et al. [2] illustrated the Collection in terms of posting videos on YouTube. In *CrowdLang* a Collection is defined as a multiplied independent transformation of a problem statement into a proposed solution using the *Multiply* operator. *CrowdLang* defines two variations of the Collection gene. First, A *Job* (see Figure 2a) is a simple *Multiply-AND* combination resulting in a set of solutions.

A *Contest* (see Figure 2b) is a *Job* followed by a *Reduce* selecting the *Job*'s best solution based on a decision.

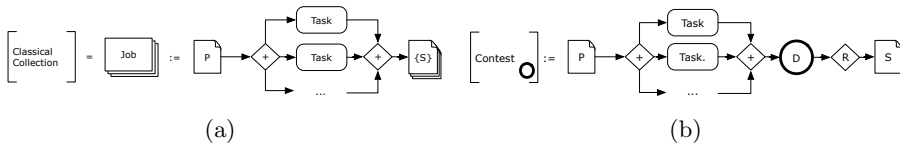


Fig. 2. (a) Classical Collection and (b) Contest

A *Collaboration* occurs when actors cooperate either by contributing iteratively or by solving different parts of a problem. *CrowdLang* supports two variations of the collaboration gene (see Figure 3a). First, an *Iterative Collaboration* models problem solving as an *iterative process of interdependent solution improvement* whereas the submitted contributions are strongly interdependent on previous ones. It can be likened to the `repeat ... until <condition>` construct of a typical programming language. A typical example of this process is article writing in Wikipedia, iterative labeling, or OCR. Based on a problem statement, a crowd worker builds an initial version of the solution followed by a decision process where either the crowd or a machine decide whether the proposed solution needs further refinement. This procedure will be repeated until the decision procedure accepts the solution.

Second, a *Parallelized Interdependent Subproblem Solving* represents the combination between a divide-and-conquer of the initial problem, the parallel execution of the partial problems, and the aggregation of the results to the final solution. The main advantage of this pattern is that it makes it possible to first split a problem into a set of independent subproblems that can then be solved in parallel. Open-source programming is an example of this pattern, where an overall problem specification is divided into subsystems, each of which are programmed and then linked together to build the resulting system.

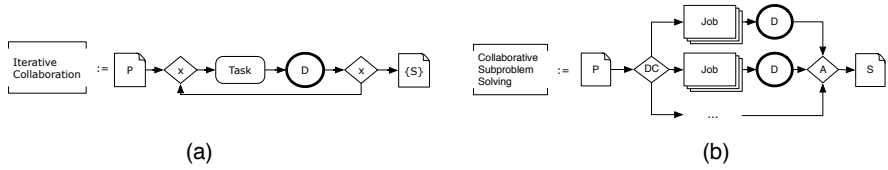


Fig. 3. (a) Collaboration and (b) Parallelized Interdependent Subproblem Solving

Decide Interaction Patterns. A *Group Decision* is defined as a mechanism that determines the best solution by using multiple crowd workers in an independent manner. Examples of group decisions are the evaluation of different solutions by voting, forced agreement, or parallel guessing with aggregation. An *Individual Decision* is a decision that is the result of an evaluation by a single human or machine agent. Note that these specifications depart from Malone et al.’s framework, under which a group decision is defined as a decision that a group makes that subsequently holds for all participants (e.g., elections, ballot questions for new laws, etc.). Our operationalization allows for group-based decisions that affect only individual actors or affect all individuals differently. This can be exemplified by the use of a recommendation system to aid movie selection.

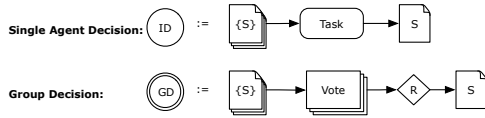


Fig. 4. The Decide Gene: Single Agent and Group Decisions

5 Design a New Application with *CrowdLang*

Using *CrowdLang*, we developed a family of 9 non-trivial text translation programs incorporating human crowd worker and machine translation.

5.1 Translating Text with *CrowdLang*

The development process—incorporating the *CrowdLang Library* and *Intelligent Discovery Assistant (IDA)* for recombining different workflow refinements — included the following five steps:

1. *Identify the Core Activities:* A programmer starts with the definition of an abstract problem-solving algorithm by identifying abstract core activities (operations) and Producer-Consumer dependencies [16] among them.
2. *Define the Design Space:* Then, (s)he selects a set of suitable interaction patterns from the *CrowdLang Library* that can be applied as operators for the abstract core activities.

3. *Generate the Recombinations*: Then the *IDA* systematically generates a set of alternative refinements by recombining the selected patterns.
4. *Execution*: The programmer executes the alternative refinements.
5. *Evaluation*: Finally, (s)he evaluates the generated variations and selects the best algorithm among the set of alternative refinements.

1. Identify the Core Activities. We started by defining an abstract problem-solving workflow for the translation task and modeled the core activities and producer-consumer dependencies among them in Figure 5. This workflow starts by first iteratively splitting the input—an article—into paragraphs and then sentences (**Task Decomposition**); then processes the resulting sentences in parallel by sequentially applying machine translation (**MT**) and crowd-based rewriting (**Rewrite**); Then, using an aggregate operator (**A**), the sentences are combined into paragraphs that are then assigned to crowd workers to improve the language quality by enhancing paragraph transitions and enforcing a consistent wording (**Improve Language Quality**). Finally, the grammatical correctness is improved (**Check Syntax**) by eliminating syntactical and grammatical errors.

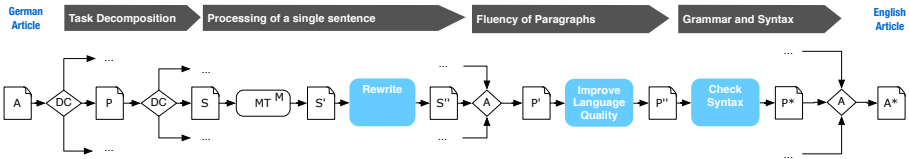


Fig. 5. Abstract translation algorithm

2. Define the Design Space. Then, we selected the following set of suitable interaction patterns for the abstract core activities identified in the previous step.

Contest with Six Sigma Pruning (CP) uses an adapted contest pattern to generate semantically correct sentences and improve text quality (see Figure 2). First, 3 different workers generate solutions. Then, these proposed solutions are pruned using the Six Sigma rule [17, p. 320 - 330]. The Six Sigma rule—a method originally used in operations research—intends to improve the output quality of a process by minimizing variability. Specifically, we compared the crowd workers’ working time on a task compared to a previously collected average. Defining the average work time as \hat{w} we hypothesize that tasks should be accomplished within the interval $\hat{w} \pm 3\sigma$ with σ as the standard deviation of the normal distribution. We minimize the number of “lazy turkers” (someone who tries to maximize his earnings by cheating) by rejecting results of workers when the working time is shorter than the lower bound $\hat{w} - 3\sigma$. We also eliminate so-called “eager beavers” (people who are going beyond the task requirements) with the upper bound $\hat{w} + 3\sigma$. We select the best solution among this remaining using a group decision. In particular, we ask 5 workers to rank the proposed solutions and then apply the Borda rule [18] to determine the winning solution.

Iterative Improvement (II) uses an iterative collaboration interaction pattern to generate semantically correct sentences and improve text quality. We define three termination conditions: (1) two out of three crowd workers assess a sentence as semantically correct, (2) the result of an iteration step is equivalent to the previous one, or (3) we exceed the number of three iterations.

Iterative Dual Pathway Structure (DP) is an adaptation of [19]. We assign the same problem (e.g., an initial translation) to two different paths. In each of the two paths, a worker is asked to improve the translation $Comp^{H1}$ and $Comp^{H2}$. At the end of this step the solutions of the two paths are compared. If the two solutions are equivalent based on an individual decision by a third crowd worker then we have a final result. If not, we iterate by sending each of the results back along its path for additional improvements until they are judged equivalent.

Find-Fix-Verify (FFV) [20] checks the grammatical and syntactical correctness of a text fragment by first asking crowd worker to find misspellings and grammatical errors. Then a group of crowd workers is asked to propose a solution for the identified problems. Finally, the solutions are verified by three independent crowd workers. Additionally, we adapted this pattern slightly by also introducing also a spell-checking software $Comp^M$.

3. Generate the Recombinations. We systematically generated a set of 9 alternative refinements for the algorithm by recombining the previously selected interaction patterns (see Table 5.1). For each refinement we chose 3 patterns for both Rewrite and Improve Language Quality.

Table 1. Resulting pattern recombinations

| | CPxCP | CPxII | CPxDD | IIxCP | IIxII | IIxDD | DPxCP | DPxII | DPxDD |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Rewrite | CP | CP | CP | II | II | II | DP | DP | DP |
| Improve L. Quality | CP | II | DP | CO | II | DP | CP | II | DP |
| Check Syntax | FFV | FFV | FFV | FFV | FFV | FFV | FFV | FFV | FFV |

5.2 Evaluation

We evaluated the different translation algorithms implemented with *CrowdLang* along a number of dimensions. We compare the results of the 9 runs as well as a pure machine translation with a gold standard human translation using an automatic text analysis measure. The best two program combinations additionally get compared to the gold standard by the crowd as well as professional translators.

Experimental Setup. The evaluation was conducted on a standard German to English translation task. Specifically, we generated translations for 15 different articles from Project Syndicate³—a Web source of op-ed commentaries—totaling

³ <http://www.project-syndicate.org/>

153 paragraphs with 558 sentences and 10'814 words translated from German to English. As a baseline, we considered Google Translate.

Evaluation Aspects. First, we considered different performance metrics such as average work time, throughput time (including waiting time), and cost per sentence. Second, based on literature research [21], we judged the translation quality along three different dimensions:

1. *Adequacy:* The meaning of the reference translation is also conveyed by the output of a translation algorithm
2. *Fluency:* The translation being evaluated is judged according to how fluent it is without comparing it against a reference translation.
3. *Grammar:* A translation segment is being evaluated according to its grammatical correctness without comparing it against a reference translation.

Evaluation Methodology. The crowd-based translation processes were evaluated using automatic machine, non-professional, and professional human evaluation.

First, we approximated the translation quality with the METEOR [22] score, which automatically estimates human judgment of quality using unigram matching between a candidate and reference translation. We considered one reference translation for each evaluation segment. Hence, a translation attains a score of 1 if it is identical to the reference translation.

Second, the translated text went through three stages of human evaluation. A monolingual group consisting of 89 native and 194 non-native speakers of English recruited on MTurk judged a set of 3 randomly extracted sentences with respect to adequacy on an ordinal scale from 1 (None) to 5 (All Meaning) [21]. A monolingual group consisting of 283 participants (140 native and 143 non-native speakers), was asked to judge a randomly extracted sentence with respect to fluency on an ordinal scale from 1 (Incomprehensible) to 5 (Flawless English) [21]. Finally, a bilingual group of 8 professional translators from the Swiss company 24translate (<https://www.24translate.ch/>) evaluated the translations by comparing each version of a translation to the German source text.

5.3 Results

Automatic Evaluation. First, we compared the resulting quality of all 8 recombinations against the baseline. In direct comparison, two algorithms—CPxCP and CPxII—outperformed the baseline (0.29) by reaching a METEOR score of 0.38 and 0.36 respectively as shown in Table 2. Note that we view the awful performance of most other recombinations not as a failure of our approach but as a desired result of a systematic exploration of the design space. Just like in biologic gene recombination, many possible solutions are not viable. Nonetheless, an approach that explores all combinations (or if computationally infeasible most combinations using some optimization approach) is more likely to uncover good solutions such as the CPxCP algorithm than one that tries to apply some kind of heuristic to immediately hone in on good ones.

Table 2. Summary of METEOR evaluation

| | CPxCP | CPxII | CPxDD | IIxCP | IIxII | IIxDD | DPxCP | DPxII | DPxDD |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| METEOR | 0.389 | 0.369 | 0.335 | 0.290 | 0.290 | 0.290 | 0.309 | 0.298 | 0.285 |
| Precision | 0.76 | 0.74 | 0.72 | 0.68 | 0.68 | 0.68 | 0.70 | 0.68 | 0.69 |
| Recall | 0.71 | 0.68 | 0.65 | 0.64 | 0.64 | 0.64 | 0.64 | 0.63 | 0.65 |

Quantitative Human Evaluation. 283 human non-professional evaluators rated the crowd-based translations in respect to adequacy and fluency on average as 3.16 and 3.37 on the ordinal scale from 1 (Incomprehensible) to 5 (Flawless English). In comparison, the professional reference translation scored on average 4.24 and 3.58. As such, the crowd-based algorithms outperform the baseline machine translation and are outperformed by the reference translation. All differences are significant at the 95% level using the non-parametric Friedman test [23]. Furthermore, the 8 professional translators evaluated CPxCP as the best of all non-professional translation algorithms (see Figure 6).

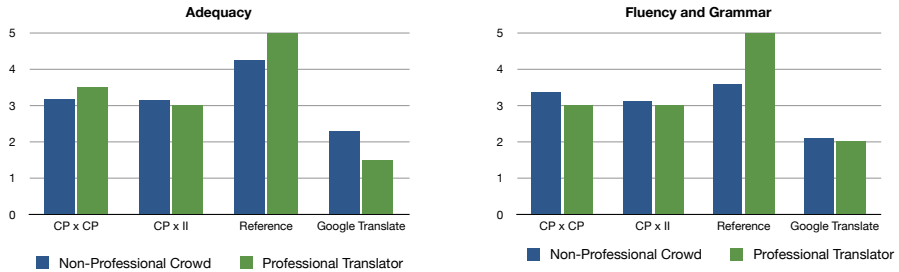


Fig. 6. Mean evaluation scores for the evaluation of adequacy, fluency and grammar by 283 English native speakers and 8 professional translators

Qualitative Evaluation. While these results show that the resulting translations are far from perfect, they still make useful translations available at a fraction of the time and cost of traditional solutions. In particular, the analysis of the follow-up interviews with the professional translators and an in-depth analysis of the adequacy, fluency and grammar score distribution (see Figure 7)⁴ show that the differences in quality are mostly caused by a few challenges in the German language morphology. For example, Translator-1 judges one of the translations as a “*Good solid translation that reflects exactly what the original says,*” whereas the pure machine translation failed, which was expressed by Translator-2 “*Non-sensical. [...]*” However, in some cases the CPxCP algorithm failed totally.

Using the professional translators’ reviews, we were able to identify several types of problem that had occurred in our experiments:

1. Word order and punctuation often lead to problems when the word order provided by the machine translation reflects the morphology of the German

⁴ The question as to whether the Likert scale should be considered equi-distant or ordinal is under debate in the social sciences. Here, we interconnected the data points, for illustration purposes only without trying to take a stance on this issue.

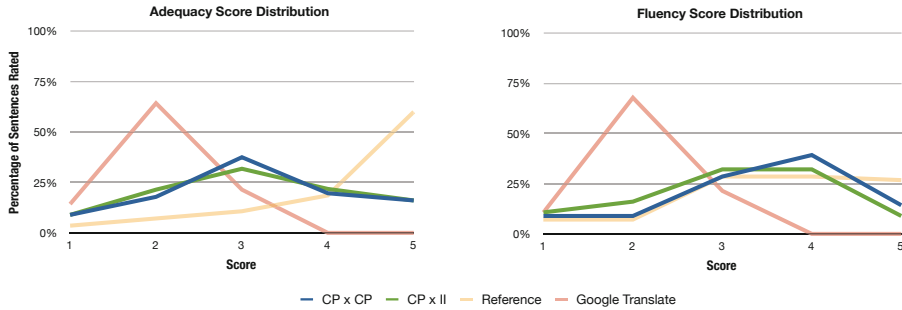


Fig. 7. Proportional score distribution per paragraph for the different translation programs in regard to adequacy and fluency

language. Translator-5 elucidates this in detail: “[...] reflects the German original [...] Adverbs come after the verb ‘to be’ in English. [...]”

2. Some translations struggle in using appropriate tenses, as expressed by Translator-1 “This would be fine except for two places that incorrectly use a relative clause with ‘which’ [...] A reader could still understand it [...]”
3. In very few instances, problems were observed that should only occur when non-native speakers or machines are editing a translation.

We subsequently found that installing text-improvement “subroutines” in the program to address these specific challenges can significantly improve the results while still keeping the throughput time and costs low. An empirical evaluation of these subroutines is forthcoming.

On average, an article translation was completed within 24 minutes for CPxCP or 35 minutes for CPxII. In terms of cost, the translation of a sentence cost 0.09\$ with CPxCP and 0.12\$ with CPxII.

6 Discussion, Findings, and Limitations

Our evaluation highlights a number of interesting findings.

(1) The translation programs illustrate that *CrowdLang* lends itself to the simple exploration of a large design space of possible program alternatives. Whilst we cannot provide empirical proof that this feature generalizes to a large number of other applications, it does, however, indicate that a systematic exploration of the design space of possible human computation programs based on known and novel patterns may help to find good solutions. This technique promises to help the transition from an era of “*Wizard of Oz* techniques,” where well-functioning programs are the result of lengthy trial-and-error processes, to a more engineering-oriented era - a goal first postulated by Bernstein et al. [20].

(2) The empirical evaluation shows that it is indeed possible to significantly improve the quality of generated translations employing monolingual crowd workers at astonishing speeds. Whilst the translations are far from perfect, they make useful translations available at a fraction of the time and cost of traditional solutions.

We are confident that the incorporation of further text improvement “subroutines” in the program— such as the use of bilingual crowd workers for the most complex German sentence structures only — can solve these kind of problems.

(3) Our adaptation of the Six Sigma rule to human computation allows us to run the processes without any sophisticated pruning techniques. We could forgo any use of “control questions”— a considerable saving in terms of effort. On the downside, our evaluation is limited in that the usage of such quality control measures may have led to better results. An evaluation of this question is forthcoming.

(4) Our pairing of the systematic exploration of the design space with the empirical evaluation helped us to find the novel human computation pattern *Staged Contest with Pruning* (CPxCP). This best-performing pattern combined contests over several stages by pruning the intermediate results using the Six Sigma rule and automatic comparison with the input to uncover cut-and-pastes.

A major limitation is that our programs have so far only been evaluated in German to English translation tasks. An evaluation using standard machine translation tasks (e.g., EU Parliament dataset), exploring the sensitivity of our programs to different machine translation tools, and other language pairs is forthcoming.

7 Conclusion and Future Work

In this paper we introduced *CrowdLang* – a general-purpose framework and programming language for interweaving human and machine computation. Using the practical task of text translation, we illustrated that *CrowdLang* allows the “programming” of complex human computation tasks that entail non-trivial dependencies and the systematic exploration of the design space of possible solutions via the recombination of known human computation patterns.

Our empirical evaluation showed that some of the resulting programs generate “good” translations indicating that the combination of human and machine translation could provide a fruitful area of human computation. Finally, it unearthed a novel human computation pattern: the “Staged Contest with pruning.”

We hope that *CrowdLang* will be used by others to implement their human computation programs, as it will allow them to easily compare different solutions.

References

1. Malone, T., Laubacher, R., Johns, T.: General management: The age of hyperspecialization. *Harvard Business Review* 89(7-8), 56–65 (2011)
2. Malone, T., Laubacher, R., Dellarocas, C.: The collective intelligence genome. *MIT Sloan Management Review* 51(3), 21–31 (2010)
3. Quinn, A., Bederson, B.: Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, pp. 1403–1412. ACM (2011)
4. Law, E., Ahn, L.: Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(3), 1–121 (2011)
5. Bernstein, A., Klein, M., Malone, T.: Programming the global brain. *Communications of the ACM* 55(5), 1–4 (2012)

6. Bernstein, A., Klein, M., Malone, T.: The process recombinator: a tool for generating new business process ideas. In: Proceedings of the 20th International Conference on Information Systems, pp. 178–192. Association for Information Systems (1999)
7. Little, G., Chilton, L., Goldman, M., Miller, R.: TurkIt: human computation algorithms on mechanical turk. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, pp. 57–66. ACM (2010)
8. Kittur, A., Smus, B., Khamkar, S., Kraut, R.: Crowdforge: Crowdsourcing complex work. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 43–52. ACM (2011)
9. Ahmad, S., Battle, A., Malkani, Z., Kamvar, S.: The jabberwocky programming environment for structured social computing. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 53–64. ACM (2011)
10. Bernstein, A.: How can cooperative work tools support dynamic group process? bridging the specificity frontier. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 279–288. ACM (2000)
11. Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D., Horvitz, E.: Human computation tasks with global constraints. In: CHI (2012)
12. Minder, P., Bernstein, A.: How to translate a book within an hour - towards general purpose programmable human computers with crowdlang. In: ACM Web Science 2012, New York, NY, USA (2012)
13. Little, G., Chilton, L., Goldman, M., Miller, R.: Exploring iterative and parallel human computation processes. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 68–76. ACM (2010)
14. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. *Communications of the ACM* 51(1), 107–113 (2008)
15. Noronha, J., Hysen, E., Zhang, H., Gajos, K.: Platemate: crowdsourcing nutritional analysis from food photographs. In: Proc. of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 1–12. ACM (2011)
16. Malone, T., Crowston, K.: The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)* 26(1), 87–119 (1994)
17. Chase, R., Aquilano, N., Jacobs, F.: Operations management for competitive advantage. McGraw-Hill/Irwin, New York (2006)
18. Young, H.: An axiomatization of borda’s rule. *Journal of Economic Theory* 9(1), 43–52 (1974)
19. Chen, Y., Liem, B., Zhang, H.: An iterative dual pathway structure for speech-to-text transcription. In: Human Computation: Papers from the AAAI Workshop (WS 2011), San Francisco, CA (August 2011)
20. Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., Panovich, K.: Soylent: a word processor with a crowd inside. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, pp. 313–322. ACM (2010)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation (2002)
22. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65 (2005)
23. Iman, R., Davenport, J.: Approximations of the critical region of the friedman statistic. Technical report, Sandia Labs, Albuquerque, NM, USA, Texas Tech Univ., Lubbock, USA (1979)

Experiments in Cross-Lingual Sentiment Analysis in Discussion Forums

Hatem Ghorbel

Information and Communication Systems Lab (ISIC),
HES-SO, HE-Arc Ingénierie, St-Imier, Switzerland
`hatem.ghorbel@he-arc.ch`

Abstract. One of the objectives of sentiment analysis is to classify the polarity of conveyed opinions from the perspective of textual evidence. Most of the work in the field has been intensively applied to the English language and only few experiments have explored other languages. In this paper, we present a supervised classification of posts in French online forums where sentiment analysis is based on shallow linguistic features such as POS tagging, chunking and common negation forms. Furthermore, we incorporate word semantic orientation extracted from the English lexical resource SentiWordNet as an additional feature. Since SentiWordNet is an English resource, lexical entries in the studied French corpus should be translated into English. For this purpose, we propose a number of French to English translation experiments such as machine translation and WordNet synset translation using EuroWordNet. Obtained results show that WordNet synset translation have not significantly improved the classification performance with respect to the bag of words baseline due to the shortage in coverage. Automatic translation haven't either significantly improved the results due to its insufficient quality. Propositions of improving the classification performance are given by the end of the article.

Keywords: Cross-Lingual Sentiment Analysis, Machine Translation, Supervised Classification.

1 Introduction

Sentiment analysis is an emerging discipline whose goal is to analyze textual content from the perspective of the opinions and viewpoints they hold. A large number of studies have focused on the task of defining the polarity of a document which is by far considered as a classification problem: decide to which class a document should be attributed; class of positive, negative or neutral polarity.

Most of the work in the field has been intensively applied to the English language. For this purpose, a large number of English corpora and resources (such as MPQA [1], Movie Review Data [2], SentiWordNet [3] and WordNet-Affect [4]) have been constructed to aid in supervised and unsupervised polarity classification of textual data. We find however relatively very few works that have explored sentiment analysis in a multilingual framework [5]. Basically, supported

by the advance of machine translation systems, researchers tend to translate the target language into English at different levels of the analysis in order to reuse existing English corpora, resources and tools.

In this context, we address in this paper the issue of polarity classification applied to French online discussions forums. We used a supervised learning approach where we trained the classifier with manually annotated French forum posts extracted from the web. As classification features, beyond the word unigrams feature taken as the baseline in our experiments, we extracted further linguistic features including lemmatized unigrams, POS tags, simple negation forms and semantic orientation of selected POS tags. The latter is extracted from the English lexical resource SentiWordNet after translating from French into English.

The main goal of our experiments is to address the problem of loss of precision in defining the semantic orientation of French words using English lexical resources, mainly due to the intermediate process of French into English translation augmented with further issues such as word sense disambiguation.

In the rest of the paper, we commence by briefly describing the previous work in the field of sentiment analysis and polarity classification in a multilingual framework. Then we discuss the issue of sentiment analysis when applied to spontaneous posts of discussion forums, moreover, we describe the extracted data and the sentiment annotation process. Afterwards, we describe the set up of extracted features and the process of French to English translation. Finally we provide and discuss the obtained experiment results and end up by drawing some conclusions and ideas for future work.

2 Previous Work

Much of the previous work focuses on defining the characteristics of conveyed opinions on the basis of textual data with processing granularity ranging from words, to expressions, sentences and documents. We mainly discern two types of research approaches that aim at solving this problem: statistical and semantic approaches. Statistical approaches make use of learning techniques to classify the semantic polarity of conveyed opinion into positive and negative classes and approximate the value of their intensity. These techniques vary from supervised to unsupervised learning, typically probabilistic methods (such as Naive Bayes, Maximum Entropy), linear discrimination (such as Support Vector machine) and non-parametric classifiers (such as K-Nearest Neighborhood) as well as similarity scores methods (such as phrase pattern matching, distance vector, frequency counts and statistical weight measures).

Generally, semantic approaches improve sentiment classification by integrating features from common sense ontologies, sentiment and lexical-semantic resources. For instance, [6-9] classify polarity using emotion words and semantic relations from WordNet, WordNet Gloss, WordNet-Affect and SentiWordNet respectively. [6] approach is based on acquiring synonyms and antonyms of a set of seed sentiment words in WordNet. Similarly [10] use WordNet to obtain a synonym set of the unseen word to determine how it interacts with the elaborated

sentiment seed lists using Naive Bayes method. On the other hand Taboada et al. [11] have developed a lexicon-based semantic orientation calculator (SO-CAL), taking into account valence shifters (intensifiers, downtoners, negation, and irrealis markers). They have created dictionaries of words annotated with their semantic orientation (polarity and strength) and made use of Mechanical Turk to check their consistency and reliability.

In this context, a large number of annotated corpora and sentiment oriented resources have been constructed among which we mention the MPQA [1], Movie Review Data [2], SentiWordNet [3], WordNet-Affect [4]), Product Review [12], Book Review [13], the Whissell's Dictionary of Affect Language [14], Linguistic Inquiry and Word Count Dictionary (LIWC2001) [15], in addition to the huge amount of available raw sentiment oriented data found in forums, blogs, chat rooms, review, debates and E-opinion web sites.

Most of the work in sentiment analysis was devoted to the English language, an important number of resources and tools have been elaborated accordingly. When addressing the same issue to other target languages, the reuse of such existing English resources and tools came out as a plausible approach. In this context, Banea et al. [16] have shown that automatic translation is a viable alternative for the construction of resources and tools for subjectivity analysis in a new target language. They have set up three experiments (i) machine translation of manually annotated English corpus for subjectivity (MPQA) to the target language (ii) machine translation of SemCor, a raw English corpus being automatically annotated using OpinonFinder [17] tool to the target language and finally (iii) automatic annotation of a machine translation target language corpus to the source language to be annotated using OpinonFinder. Annotations here by are projected back into the target language. The obtained corpora in the three experiments are then used for training a subjectivity classifier. All of the experiments have shown interesting performance of the subjectivity classifier with an F-measure of 71.83% for Romanian and 73.41% for Spanish.

Other approaches [9, 18] have used machine translation from target language to English in order to reuse existing English polarity classifier tools already trained with English resources. Balahur and Turchi [19] have set up polarity classification experiments by translating the English data provided in the NTCIR 8 Multilingual Opinion Analysis Task(MOAT)¹ to French, German and Spanish using three different machine translation systems; Google, Bing and Moses translator. Each obtained corpus was used for training a Support Vector Machine (SVM) classifier. Additionally, a test corpus (called Golden Standard) have been constructed using a machine translation with Yahoo system and then manually corrected for all the languages. In the first set of experiments, different combinations of training set and test set are taken (sets translated using the same translation system or varying training sets and limiting test sets to the Golden Standard). In the second set of experiments, all translated data are combined together to form a unique vector model based on unigrams and bigrams features for training. Tests are performed using the Golden Standard. Results have

¹ <http://research.nii.ac.jp/ntcir/ntcir-ws8/data-en.html>

shown that machine translation systems are mature enough to produce reliable training data for languages other than English, however, noise and sparseness in the translated data should be reduced to improve classification performance.

Further cross-lingual sentiment analysis work motivated by the existing of bilingual corpora has been proposed by Génèreux and Poibeau [20] in the framework of DEFT'09 workshop who used a parallel English/French corpus² already aligned at the sentence level. Once, English version of texts were annotated for subjectivity using OpinonFinder tools, sentiment (objective or subjective) tags are projected to the French parallel version and used as a training corpus for subjectivity classifier. Authors reported an F-measure of 83%.

Similarly, Kim and Hovy [21] applied word alignment on bilingual corpora (The European Parliament corpus) in order to extract translation pairs of opinion bearing words used as clues to determine the sentiment of a whole email. Mihalcea et al. [22] proposed a cross-language projection method by translating the lexicon in OpinonFinder using two bilingual dictionaries (an authoritative English-Romanian dictionary and the free the Universal Dictionary). They chose the first sense in the bilingual dictionary in the case of ambiguity. The obtained translated lexicon was used in a rule-based subjective classifier with a comparable performance to the English OpinonFinder. Such a classifier was later on exploited to create a Romanian subjective corpus used for training an automatic classifier.

Apart from the translation of training corpora and the use of parallel corpora, researchers have translated sentiment dictionaries such as the Linguistic Inquiry and Word Count (LIWC) dictionary [15] translated into several other language (French [23] for instance) and used to boost a rule-based multilingual sentiment analysis software.

3 Data Description

In this work we address the issue of sentiment analysis of posts in online discussion forums aiming to provide a platform that initiates debates on general issues. The data is driven from the discussion forum of Infrarouge³, a TV French program on Swiss TV (RTS) discussing different political, social and economic issues in Switzerland. Posts are presumed to contain already opinionated text holding the user's viewpoint on the discussing subject.

Initially, and in order to evaluate the task of classifying posts in a positive, negative or a neutral classes, we go through a manual sentiment annotation task⁴

² Excerpts of Hansard corpus composed of Canadian parliament debates available at <http://www.cse.unt.edu/~rada/wpt/data/English-French.training.tar.gz> aligned by Ulrich Germann.

³ Infrarouge <http://www.infrarouge.ch> is a Swiss TV program (in French) that weekly presents a direct debate on hot social, economic and political issues such as issues related to popular votes, elections and national referendums. The debate is initiated by invited participants related to the subjected and continued offline on a dedicated web forum by different Infrarouge TV program viewers.

⁴ We use the term of sentiment annotation to denote the human task of classifying a text as positive, negative or neutral according to the conveyed sentiment.

as already performed in movie review classification [1, 24]. Movie review annotation was a simple task in the sense that positive opinion means an opinion that appreciates the movie and hence we talk about a positive position. Negative position reveals a negative opinion that is seen as a form of depreciation and devaluation of the movie. Positions where no final decision of appreciation or depreciation was retained are said to be neutral. Nevertheless when it comes to evaluate participant position towards a conflicting issue posted in online debates for instance, things are not as simple as movie reviews. Conflicting issues are often open questions dealing with everyday social, economic and political problems. Posts are elements of a conversation between participants on a such issues, therefore, opinions are not quite explicit and mostly conveyed by the means of strong argumentations, rhetorics and supports rather than merely as a *for* or an *against* position.

For example, in Infrarouge forum , we have experienced some difficulties when annotating debates such as "Succession of Calmy-Rey : the four social candidates come out !" or "redundancies: the hemorrhaging until when?" or "Criminality : is-it the end of the suisse exception ?". Indeed, it is hard in many cases to detect a unique position since participants post may be mitigated and so doesn't necessarily converge to a clear and unique position. Participants may give mixed arguments *for* and *against* without explicitly revealing their position or argue someone else's position or raise a further related issues in their posts. From such puzzle of arguments it is often hard even to a human annotator to project the post content to a single positive or negative position and therefore to settle on a global polarity of the post. In order to evaluate the difficulty of the human annotation task, we have calculated the kappa ratio between two separate human annotations.

In a first experiment, we considered a three class annotation where two different annotators should annotate each post with one of the three classes (positive, negative and neutral). Results showed that only an agreement measured as a kappa ratio of 0.59 is obtained, which reflects a certain disagreement between the annotators. A simple analysis of the data shows that the conflict comes mainly from the neutral class annotation, which is apparently perceived differently by the annotators. Indeed, the neutrality is often a quite subjective point which seems difficult to be agreed on. For these reasons and in order to resolve the problem of human annotators' disagreement, we decided in a second experiment to eliminate the class of neutral positions and we assumed that all the posts are opinionated. Once neutral class is eliminated from annotation guidelines (posts which are classified as neutral by at least one of the annotators are deleted from the corpus) we obtain an acceptable kappa ratio of 0.77.

From a discursive viewpoint, an issue can be considered as a yes/no question that generally prefers an affirmative or a negative answer over other alternatives. For conversation analysts, this is known as the preference structure where sequences such as adjacency pairs are forwarded as organization units consisting of a first turn (the question) followed by a second turn (the answer) [25]. According to Schegloff [25], each yes/no question is followed by an affirmative

or a negative answer, but with a certain preference according not only to the grammatical design of the question but also ”to the actions which the questions are being used to perform and on the displayed knowledge state or epistemic strength from which the questions are asked” [2: 10] [26].

At this conversation analysis level, the context of polarity is determined from a grammatical perspective (the existing of a grammatical negation form) not from a semantic perspective (a positive or negative sentiment is conveyed). However, it is typical that grammatical negation affects the overall sentiment polarity of the assertion but not the opposite. That is why in online discussions, sentiment analysis should take into account the conversation structure since the sentiment conveyed in a post could not be independent either of the post replied to or of the discussion (or thread) topic.

4 Methodology

4.1 Features Design

Similarly to previous sentiment analysis studies, we have defined three categories of features: lexical, morpho-syntactic and semantic (word semantic orientation or polarity). Lexical and morpho-syntactic features have been formulated at the word level, whereas semantic features have been formulated at the sentence and post level. As a baseline of our experiments, only features composed of word unigrams are included. Each unigram feature formulates a binary value indicating the presence or the absence of the corresponding word at the review level⁵.

Lemmatization is argued to be relevant in sentiment analysis because it aids in grouping all inflected forms of a word in a single term feature. For example the words *aimé*, *aimait* and *aimer* share the same polarity but will be considered as three separate features for the classification process. Some studies showed that restricting features to specific part-of-speech (POS) categories would improve performance (for instance Hatzivassiloglou and McKeown [28] have restricted features to adjectives). In our approach, POS tags are proposed to be used to enrich unigram features with additional morpho-syntactic information so as to disambiguate words that have similar spellings but different polarity. For example, it would distinguish the different usages of the word *negative* that can either be a neutral noun *a piece of photography film* or an adjective *a negative opinion*. Moreover POS tags are useful to handle negation and to aid word sense disambiguation before polarity extraction in SentiWordNet as it will be detailed hereafter.

Negation is handled at the shallow level of morpho-syntactic constituency of sentences so as to avoid the heavy processing of its deep syntactic structure. The detection of negated expressions is performed by searching specific patterns formed from common negation forms. The scope of the negation is definite by a

⁵ This choice is only technical since the Simple Vector Machine algorithm requires the normalization of features for a better classification performance [27].

window of words that follows particular patterns of POS categories. We defined two simple patterns that cope with the negation form (1) at the verb level for example *le scénario ne brille pas* (*the scenario is not outstanding*) and (2) at the adjective and noun level for example *sans histoire originale* (*there is no originality in the story*). Within the scope of the negation, polarity of the *nouns*, *verbs*, *nouns* and *adjectives* is being inverted. The entailments of such a polarity inversion are first situated at the lexical level; unigrams features are inverted during features vector construction: that is if we consider the previous example, in stead of having in the feature *original*, we would have a different feature *!original* in the vector. Second, at the semantic level, polarity value is inverted (from positive to negative and vice-versa) in the calculation of the overall polarity of a post as we will detail in the following section⁶.

As argued in previous works [6–9], the incorporation of corpus and dictionary based resources such as WordNetAffect, SentiWordNet and Whissell’s Dictionary of Affect Language contributes in improving the sentiment classification. Based on such results, we use the lexical resource SentiWordNet⁷ to extract word polarity and calculate the overall polarity score of the post for each POS tag. SentiWordNet is a corpus-based lexical resource constructed from the perspective of WordNet [29]. It focuses on describing sentiment attributes of lexical entries describe by their POS tag and assigns to each synset of WordNet three sentiment scores: positivity, negativity and objectivity.

More specifically, to each post, we added two features holding a binary value of positivity and negativity for each POS category (namely adjective, adverb, noun and verb POS tags). This binary value is defined according to the sum of extracted scores from SentiWordNet: if we find a greater value of positivity, then the positive feature is set to one and the negative feature to null, and vice versa. The same process is applied to all POS categories, we obtain hence 8 additional binary features.

4.2 Does Translation Preserve Sentiments?

Since SentiWordNet describes English lexical resources, we need to translate terms from French into English before polarity extraction. Words are lemmatized before being passed through the bilingual dictionary. We use POS information as well as the most frequently⁸ used sense selection to disambiguate senses and predict the right synset. We only considered the positivity and the negativity features for the four POS tags noun, adjective, verb and adverb for this task.

An important issue to consider is whether sentiments expressed in a language A are preserved when translated into a language B? A first and an immediate

⁶ Although it is shown in [11] that when a word is negated, it does not necessarily entail a sign inversion. Consider for example the adjective “excellent” with a polarity score of +5, then it would be incoherent to attribute the score of –5 to “not excellent”.

⁷ SentiWordNet versions 1 and 3.

⁸ This choice is based on the assumption that reviewers spontaneously use an everyday language.

answer is yes, and yet this is that aim of the translation in principle. However, when giving the task of sentiment analysis to a computer program, the answer is not straight forward; a number of parameters should be taken into account before giving such an answer. First, how does the sentiment analyzer process, is-it language independent or built upon linguistic resources? Second, how the translation is performed, manually or automatic, and what is its quality. On the lights of what previously discussed, let's consider some illustrating examples to show the different sides of the issue:

1. On ne peut pas laisser au simple⁻ fait du hasard ce qui se passe au dessus de nos ttes.
2. We cannot leave with the simple⁻ matter of chance⁺ what takes place above our heads.

The French sentence (1) conveys a negative opinion expressed by the negation of the verb *laisser*, the adjective *simple* used in a negative context (according to the English WordNet, it corresponds to the 4th synset entailing a sense of mere, bare and apart of anything else. In SentiWordNet, this synset has a negative score of 0.375 and a null positive score). In the English translation (2), *hasard* is translated into *chance*. In French, *hasard* is a neutral word having the sense of the measure of how likely it is that something occurs, however, the first synset in the English WordNet of its translation *chance* has the sense of opportunity which is rather positive. Indeed, we should go to the fourth synset of *chance* in WordNet to find its just sense in that context. In SentiWordNet, the polarity of this synset is found to be neutral, similarly to the French *hasard*.

As a matter of fact, as it is shown in the above example, the translation of a word does preserve its polarity since this is an intrinsic element of the word sense; nevertheless, this cannot be found unless senses are correctly disambiguated according to the context.

The process of disambiguation enables the definition of the right sense according to the context of the word, if we refer to WordNet taxonomy; each entry should have a list of senses or synsets that are ranked according to their frequencies in already tagged data provided for the database construction. Obviously, in a multilingual context, the synset ranking is not necessarily equivalent since different tagged data is used across languages and so different sense frequencies are observed.

Now, if we chose the most simple word sense disambiguation method based on the first listed sense selection in WordNet (for a given POS it achieves a 57% of recall and precision according to McCarthy et al. [30]), and in case of translation performs well, we will get inappropriate matching of synsets across languages, which will surely have an impact on polarity calculation.

When coping with this issue, some researchers have solved the problem from the perspective of polarity calculation by assigning a unique scoring to each word which corresponds to the mean of synset scores [11]. Others have suggested to develop more efficient word sense disambiguation tools capable of selecting the right synset in each time before polarity calculation [31], this is more challenging since this is generally a language dependent task.

In this work, besides going through a complete translation, we applied a word synset translation but only of relevant lexicon that is likely to hold sentiment information. We make use of EuroWordNet [32] where multilingual synsets are already aligned: each French sense is matched to the corresponding English sense; hence the first synset of the French *hazard* is matched to the English synset of *coincidence* and *happenstance*. Doing so, it would be sufficient to disambiguate the French sense in order to obtain the matching English sense whose polarity would be evaluated using SentiWordNet.

5 Experiments

We have used in our experiments the Infrarouge corpus collected from Swiss TV program website⁹ and consists of 650 posts. As described in section 3, two human annotators have agreed on 318 positive posts and 332 negative posts. Table 1 describes in more details the constructed data.

Table 1. Characteristics of the training and test data

| Corpus | Posts | Words | Nouns | Verbs | Adjectives |
|----------|-------|--------|--------|--------|------------|
| Negative | 332 | 43'309 | 8'665 | 7'787 | 2'927 |
| Positive | 318 | 41'171 | 8'560 | 7'376 | 2'578 |
| Total | 650 | 84'480 | 17'225 | 15'163 | 5'505 |

The data is preprocessed with the TreeTagger [33], a French POS tagger and lemmatization tool in order to define POS tags and base form of the words. As a machine learning algorithm, we applied Support Vector Machine (SVM) classification method and utilized SVM^{Light} [34] classification tool with its standard configuration (linear kernel) to implement a series of experiments where each time we define a set of combined features and evaluate the accuracy of the approach. Evaluation is calculated using 5-folder cross validation.

Table 2. Results in terms of global accuracy using French and English translated data

| Experiments | French | English Tr |
|-------------------|--------|------------|
| (1) Baseline | 63.08% | 64.12% |
| (2) Negation | 63.88% | – |
| (3) SWN1 Polarity | 66.15% | 66.41% |
| (4) SWN3 Polarity | 66.05% | 61.07% |

As sketched in table 2, we have set features as simple bag of words after POS tagging and lemmatization in baseline experiments (1) applied first to original French corpus and then to respective English machine translation.¹⁰ The next

⁹ <http://www.infrarouge.ch>

¹⁰ We used Google Translator <http://translate.google.com>

experiments (2) have shown that French negation didn't improve much the results as expected (less than +1%), this is mainly due to the difficulty of capturing the negation scope defined as a window of words around the negated verb. In fact, such a limitation to the verbal phrase level makes it hard to capture negation at the other levels such as noun and adjective phrases which needs deeper syntactic dependency analysis.

A further problem is when negation is coupled with word polarity feature. In deed, words found to be negated would have their polarity score inverted, which is not necessarily coherent as pointed out by Taboada et al. [11], take for instance the adjective *excellent* has a positive score of 1, nevertheless attributing a negative score of -1 to *not excellent* wouldn't be relevant since the latter is not as negative as *worst* for example. Thus, the policy of polarity inversion should be revised.

When we have integrated word polarity in experiments (3) and (4) as additional features in the French data, only very limited improvement is shown (+0.7%). In fact, in these experiments, we first go through a word synset French to English translation using EuroWordNet (EWN) before extracting the polarity scores from the English sentiWordNet (SWN1 then SWN3). These tentative results are mainly due to the following three reasons.

1. EWN recall. As shown in table 3, the recall of EWN is weak; only 50% of the words synset are translated. As a matter of fact, the coverage of French EWN dictionary is not very high (there exists only 18'777 entries) compared to the English WordNet 1.5 that contains 126'617 (7 times) entries and to the English WordNet 3 that contains 155'287 entries (8 times). Consequently, if only half of the words are translated, the recall issue would affect the quality of the calculated polarity at the sentence or the post level.
2. SWN precision. Unlike EWN, SWN recall is found to be relatively high (more than 90% for SWN3) because the latter is based on WordNet 3 which has a high lexical coverage. However, recall conclusions should be carefully drawn since all SWN entries are already English translations and hence are successful outputs of EWN. Therefore, the measured recall remains tentative and relative to EWN results.

As a matter of fact, recall of SWN3 shouldn't be a source of difficulties since the latter is based on WordNet 3 known as having a high coverage. Nevertheless, it is the quality of the polarity information stored in SWN3 that should be checked. Further experiments with SWN3 have shown that polarity scores of all words of the corpus (nouns, verbs, adverbs and adjectives) are distributed as the following: 5.09% positive, 3.29% negative and 91.62% neutral which gives evidence of the positive bias degree of the resource. This explains also the difficulties in classifying negative posts (only 60% of negative posts are correctly classified whereas 72% of positive posts are correctly classified). Detailed experiments using SWN1 have shown that even if SWN1 is smaller in size compared to SWN3 (a recall of about 75.5%), it is found to be less positively biased than SWN3 since about 70% of the negative posts are correctly

classified. These findings confirm Taboada [11] experiments set up for affective dictionary comparison using the SO-CAL tool.

3. WSD: First synset choice. As already discussed in section 4.2, this issue is the weak point of our research, since we have not yet developed a Word Sense Disambiguation algorithm for French, the first synset choice for EWN translation or for SWN polarity scoring entails a lack of polarity precision and explains partly the weak classification results obtained in experiments (3) and (4) compared to previous similarly set up experiments with French movie reviews [35].

Similar conclusions could be drawn when interpreting English machine translated data where SWN1 polarity scores yield significantly better classification results (66.41%) than those of SWN3 (61.07%) despite the coverage difference between the two resources. This confirms the polarity bias hold within SWN3.

Table 3. EuroWordNet and SentiWordNet evaluation using French data

| Experiments | Results |
|-----------------------------|---------|
| Size (words) | 42'956 |
| Features | 12'234 |
| EWN recall | 47.01% |
| EWN recall (N, V, Adj, Adv) | 54.32% |
| SWN1 recall | 75.5% |
| SWN3 recall | 91.5% |

6 Conclusions

In this article, we addressed the problem of sentiment analysis in French online discussion forums. We have shown that this is a difficult task and requires deeper processing than movie review sentiment analysis since even human annotators are found in situations of doubt when distinguishing between positive and negative posts. We have argued that posts are elements of a conversation between participants on a predefined issue, hence, opinions are not quite explicit and mostly conveyed by the means of strong argumentations, rhetorics and supports rather than merely considered as a *for* or an *against* position.

For these reasons, we attempted to include polarity scores at different levels as additional features in the training process. Since no existing polarity resources are constructed for the French language, we went through a translation process so as to use the available English resources. The main drawbacks of our approach concern first the process of translation and second the polarity score extraction.

First, for the translation issue, two methods of translations are evaluated, full machine translation and EuroWorNet synset translation. Even if performance is found to be similar, the problems are different. On the one hand, the quality of machine translation is not sufficient to cope with spontaneous forum posts that

often violate linguistic assumptions. Much of the textual evidence and clues used in sentiment analysis may be translated incorrectly and therefore opinion polarity may be altered in the translated text despite of the relevant advance of current systems. On the other hand, EuroWorNet synset translation requires a preliminary task of word sense disambiguation to chose the correct context dependent synset.

Second, when extracting the polarity scores of words, the absence of a word sense disambiguation algorithm doesn't allow for the choice of the adequate polarity according to the context. Moreover, there still remain some pending questions about the quality of the SentiWordNet resource since it is figured to be positively biased.

As a future work, we plan to integrate a word sense disambiguation algorithm for French in the core of our sentiment analysis tool and develop deeper linguistic and discourse analysis such as elaborated forms of negation, conditionals, intensifiers (such as very, quite, extremely, etc), discourse connectors (such as although, in spite of, even, etc) as well as the discourse structure of conversations. We intend also to detail the analysis at the sentence level before scaling at the post level. A further challenging work we intend to perform is to construct a French sentiment dictionary to get rid of the translation task and eliminate all the related problems. In this sense, there exists already worth pointing out work [23, 36] that could be an interesting starting point.

References

1. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2005), Vancouver, B.C., Canada, pp. 347–354 (October 2005)
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, pp. 79–86 (July 2002)
3. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation, LREC, vol. 6 (2006)
4. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, pp. 1083–1086 (May 2004)
5. Banea, C., Mihalcea, R., Wiebe, J.: Multilingual sentiment and subjectivity analysis. In: Zitouni, I., Bikel, D. (eds.) Multilingual Natural Language Processing, Prentice-Hall (2011)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of Knowledge Discovery and Data Mining (KDD 2004), Seattle (2004)
7. Esuli, A., Sebastiani, F.: Determining the semantic orientation of terms through gloss classification. In: Proceedings of CIKM 2005, pp. 617–624 (2005)
8. Nastase, V., Sokolova, M., Shirabad, J.S.: Do happy words sound happy? a study of the relation between form and meaning for english words expressing emotions. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2007), pp. 406–410 (2007)

9. Denecke, K.: Using sentiwordnet for multilingual sentiment analysis. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE 2008), Cancun, Mexico, pp. 507–512 (2008)
10. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, pp. 1367–1373 (August 2004)
11. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2), 267–307 (2011)
12. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: The Third IEEE International Conference on Data Mining (2003)
13. Gamon, M., Aue, A.: Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In: Proceedings of the ACL 2005 Workshop on Feature Engineering for Machine Learning in Natural Language Processing. Association for Computational Linguistics, Ann Arbor, US (July 2005)
14. Whissell, C.M.: The dictionary of affect in language. In: Lutchik, R., Kellerman, H. (eds.) *Emotion: Theory, Research, and Experience*, pp. 113–131 (1989)
15. Pennebaker, J.W., Francis, M.E., Booth, R.J.: *Linguistic Inquiry and Word Count (LIWC): LIWC 2001*. Erlbaum Publisher, Mahwah (2001)
16. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual subjectivity analysis using machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 127–135. Association for Computational Linguistics, Stroudsburg (2008)
17. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP 2003, pp. 105–112. Association for Computational Linguistics, Stroudsburg (2003)
18. Abbasi, A., Chen, H., Salem, A.: Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)* 26(3), Article 12 (2008)
19. Alexandra, B., Marco, T.: Multilingual sentiment analysis using machine translation? In: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 52–60. Association for Computational Linguistics, Jeju (2012)
20. Génereux, M., Poibeau, T.: Approche mixte utilisant des outils et ressources pour l'anglais pour l'identification de fragments textuels subjectifs français. In: Actes de l'atelier de clôture de la cinquième édition du DÉfi Fouille de Textes (DEFT 2009), Paris (June 2009)
21. Kim, S.M., Hovy, E.H.: Identifying and analyzing judgment opinions. In: Proceedings of the Human Language Technology Conference of the NAACL (HLT-NAACL), New York, USA (2006)
22. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the Association for Computational Linguistics (ACL 2007), Prague (June 2007)
23. Piolat, A., Booth, R.J., Chung, C.K., Davids, M., Pennebaker, J.W.: La version française de liwc: modalités de construction et exemples d'application. *Psychologie Française* 56, 145–159 (2011)
24. Ghorbel, H., Jacot, D.: Further experiments in sentiment analysis of french movie reviews. In: Proceedings of the 7th Atlantic Web Intelligence Conference on Advances in Intelligent Web Mastering 3, AWIC 2011, Fribourg, Switzerland, vol. 86, pp. 19–28 (2011)

25. Scheggloff, E.A.: Sequence organization (2005) (unpublished manuscript)
26. Koshik, I.: Beyond Rhetorical Questions: Assertive Questions in Everyday Interaction. John Benjamins (2005)
27. Ben-Hur, A., Weston, J.: Data Mining Techniques for the Life Sciences. Springer (2009)
28. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. In: Proceedings of the 8th conference on European Chapter of the Association for Computational Linguistics, Madrid, Spain, pp. 174–181 (1997)
29. Fellbaum, C. (ed.): WordNet An Electronic Lexical Database. MIT Press (1998)
30. McCarthy, D., Koeling, R.: JulieWeeds: Eurowordnet general document. Technical Report CSRP 569l, Department of Informatics, University of Sussex, Falmer, Brighton (2004)
31. Rentoumi, V., Giannakopoulos, G., Vouros, G.A.: Sentiment analysis of figurative language using a word sense disambiguation approach. In: Proceedings of the International Conference on RANLP, pp. 370–375 (2009)
32. Vossen, P.: Eurowordnet general document. Technical Report Version 3 Final, University of Amsterdam (2010)
33. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, pp. 44–49 (1994)
34. Joachims, T.: Making large-scale svm learning practical. ACM Transactions on Information Systems, TOIS (1998)
35. Ghorbel, H., Jacot, D.: Sentiment Analysis of French Movie Reviews. In: Pallotta, V., Soro, A., Vargiu, E. (eds.) DART 2011. SCI, vol. 361, pp. 97–108. Springer, Heidelberg (2011)
36. Gala, N., Brun, C.: Propagation de polarités dans des familles de mots: impact de la morphologie dans la construction d’un lexique pour l’analyse d’opinions. In: Actes de Traitement Automatique des Langues Naturelles (TALN 2012), Grenoble (2012)

Quality Assessment of User Comments on Mobile Platforms Considering Channel of Activation and Platform Design

Christopher Fröch¹ and Martin Schumann²

¹ Evolaris Next Level GmbH, Junior Scientist, Graz, Austria
christopher.froech@evolaris.net

² Evolaris Next Level GmbH, Senior Scientist, Graz, Austria
martin.schumann@evolaris.net

Abstract. In this paper we present the results of an experimental three-steps-study concerning quality assessment of product reviews. As reviews and comments on products and services are gaining importance in the context of purchasing decisions providers of review platforms are seeking for ways to improve the quality of their platforms. In the presented study user expectations regarding quality of comments were collected as well as reader perceptions of quality. Additionally a thorough text analysis of experimentally obtained product reviews was conducted. The main results of this research were that quality expectations do not necessarily lead towards good quality comments provided by the same person. Moreover it could be observed that the combination of text and star rating is preferred by people and also will lead to better understandability of resulting comments. The channel of activation, NFC or QR codes did not cause any significant difference considering comment quality or appropriate platform.

Keywords: product reviews, quality assessment, content analysis, experiment.

1 Introduction

Many customers inform themselves and also partly base their purchasing decisions on product reviews or comments on the internet. The number of user generated content has increased in the past few years and so do the comments on products. Not only is the user generated content increasing, but also the number of platforms and providers. Jensen said that “*Technologies that enable websites to support the creation, sharing, and deployment of user-generated mobile services could be key factors in the spread of the mobile Internet*“[5]. Through better performance of mobile internet and the fast development processes of mobile devices the mobile sector is playing a bigger role in people’s lives. Also the connection from physical object to the digital world through 2D QR-Codes or near field communication affects the increase of user generated content positively. The barriers to give feedback and write comments on products or services are getting weaker and the whole process becomes easier. Hoegg states that,

the lower the entry barrier, the more likely is the occurrence of low quality content [4]. Therefore this study tried to find out **which interactivation, NFC or QR-Codes, in combination with different configurations of a feedback platform provides the best formal quality.**

As barriers disappear and possibilities to generate content easily increase it is not only professionals who generate content and comments on platforms, but also regular non-professional people. The question is **which criteria do regular generators of comments consider to be important for high-quality comments?** The approach followed within this research project is different state of the art research that examines intrinsic and/or extrinsic motivators for generating content. Prior studies [6], [8], [10] focused on the assumption that users have different motivators. As these motivations differ from person to person and entry to entry the formal quality could differ. Although people do have clear expectations of how a qualitative good feedback or comment looks like, they do not always follow their own rules. Therefore it is necessary to find out if **there is a difference in formal quality with respect to expectations and the actual comments?**

In 2010 von Reischach found out that people like comments consisting of text in combination with a star-rating most. Furthermore he found out that NFC is the preferred kind of interactivation [9]. In addition Jensen states that with an expansion of the mobile web an increasing number of applications will include geospatial and social components [5]. So another interesting group is the consumers of user generated content. They only read or view the feedback or comments. Although they do not know the background of the author they trust their comments. They have to understand and decode the message behind the comments from the generators. This could be influenced by formal quality and text difficulty, but **do consumers have different expectations?**

This study tries to approach these questions in an explorative way with a technological focus. To answer the questions a questionnaire, an experiment and a consumer-rating were applied. The aim of the questionnaire was to address the first subquestion. In order to find out the formal quality of comments an experiment was conducted. In this experiment user had to generate content on different mobile platform configurations. Afterwards the texts, which were written in German, were analyzed. Later on these texts were rated by two people to find out whether formal quality measures do meet the consumer's quality needs.

The main goal of the experiment was to give recommendations for the design of mobile platforms that are activated by near field communication and 2D QR-Codes. In the second chapter of this paper the current state of the art is outlined to give a brief overview of prior research on mobile platforms and user generated content creation. The following empirical section describes the experimental frame and methods of research. Subsequent the results are presented. The paper concludes with the recommendations and proposing topics for further research.

2 State of the Art

To find out what generators expect from comments Chen & Xu [2] examined the criteria topicality, novelty, reliability understandability and scope. They showed that topicality and novelty are very essential. Reliability and understandability were also found to be significant factors but scope was not. Another survey from Agichtein et. al [1] used formal criteria, like punctuation, semantic and grammar to analyze texts and their quality.

Another method to analyze texts is to evaluate text difficulty. There exist some different formulas for that purpose. One of them is the Flesch Reading Ease formula. It was defined for analyzing official documents. It was primarily set up to evaluate texts in English [3]. Mihm [7] modified the rating tableau in order to use this formula for German texts too. He changed the original scale because the word length in English and German texts differs. Some other formulas are the Amstadt formula and four different versions of the “Wiener Sachtextformel”.

In contrast von Reischach [9] conducted a survey that examined the needs of people who read feedback and product comments. One of his main results was that people prefer feedback and comments in combination with star-ratings. Another observation in this study was that the interactivation by NFC is fastest.

It is not only the preferences of users that influence their expectations but also the motivation of users who generate content. To find out which intrinsic motivation users have an analysis of 385 random start pages of blogs in Polish was done. In the course of this research six different motivations were identified [10]:

- *self-expression*
- *social interaction*
- *entertainment*
- *passing the time*
- *information and*
- *professional advancement*

Similarly Nardi [8] interviewed 23 bloggers from California. He found out that it was important for them to document their life, express emotions, communicate opinions and ideas and be part of forums. These aspects could influence the expectations on quality of their comments.

3 Methods

In March 2012 the participants of a Living Lab focused on mobile technologies were invited to take part in this experiment. They were provided with mobile devices in order to ensure NFC capability of test devices. 60 people participated and were rewarded a 10 € coupon. The sample consisted of 26 women and 34 men and most of them from urban areas. Figure 1 depicts the age distribution of the sample of which a huge part was between 18 and 30 years.

Age distribution among the participants

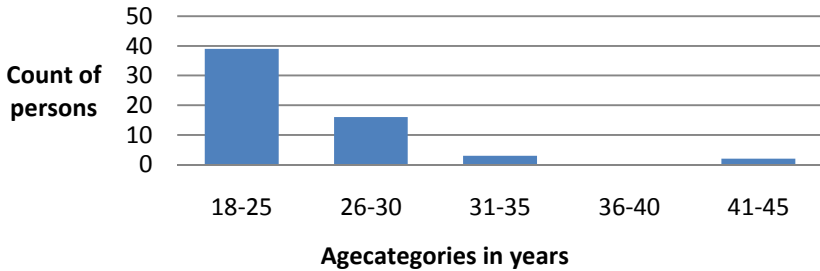


Fig. 1. Age distribution among the participants

In the beginning the participants had time to play four different “*Minigames*”. These “*Minigames*” were simple reflex and action games. Two examples are shown in figure 2. The assignment of the “*Minigames*” was randomly decided by a random generator. After five minutes the participants had the possibility to write a short text using the mobile device. This was just a test so that everybody could get used to the smartphone input. Then the recording started. The attendant chose which “*Minigame*” he wanted to comment first. Then he had to activate the mobile platform either by NFC or QR-Codes. The activation codes and tags were fixed at the bottom of the “*Minigames*” as you can see on figure 2:



Fig. 2. Position of the QR-Code and the NFC-Tag

When activation was successful an acoustic signal sounded and the participants were able to enter their comments. There were five different combinations of input possibilities and each NFC-Tag or QR-Code referred to one combination: “Only Text”, “Text with Like-Button”, “Text with Star rating”, “Text with Photo-upload” and “Text with Like-Button, Star rating and Photo-upload.” (Figure 3 shows a screenshot of the mobile platform with its different possibilities for generating content).

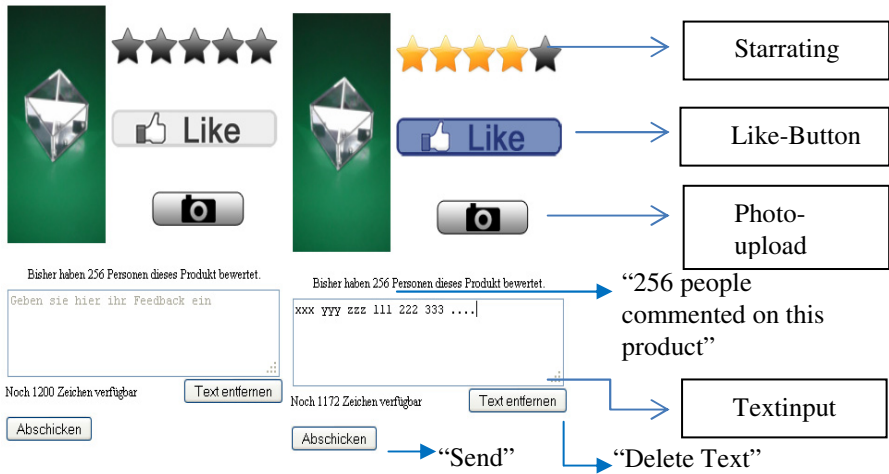


Fig. 3. Mobile Platform before and after the input

Every person had to activate two input combinations by means of NFC and another two by means of QR-Code. When he or she finished one comment they returned the mobile device to the supervisor. When the participant wanted to start the next “*Mini-game*” he got the smartphone back again.

When the four comments were created the person had to answer a questionnaire. The questionnaire consisted of three demographic questions and six specific items regarding quality criteria. The participants had to rate importance of the different criteria for themselves. They rated the priority from 1 very important to 6 not important at all.

After the experiment ended two people defined five criteria for rating the obtained comments. These five criteria are described in chapter 4.5. They then rated all 240 texts from grade 1 – very good to 5 – very poor. The average ratings per platform and interactivation were computed and compared. Interrater reliability was computed in form of Cronbach’s Alpha and is 0,716.

4 Scientific Frame

The texts were analyzed and punctuation and capitalization and text difficulty were evaluated. The criteria punctuation and capitalization follow the new rules of German orthography from 2006. To compare the interactivation by NFC and QR code the duration of the activation was observed. The participants also had to answer a questionnaire in which they had to indicate their priority of the criteria: punctuation, capitalization, text difficulty, grammar, topicality and novelty.

4.1 Punctuation

Each text was analyzed as it would have been a continuous text. Therefore every missing period and comma was counted and summed up. Additionally every period and comma that was excess or not in the correct position was summed up.

The means were computed to compare the different mobile platforms and kind of interactivation.

4.2 Capitalization

Every incorrect use of capitalization which is more important in German language than in English was summed up. The means were computed to compare the different mobile platforms and kind of interactivation.

4.3 Text Difficulty

Every text was evaluated by means of the Flesch Reading Ease formula [3]. Using the modified scale of Mihm [7] enabled application of this formula also on German texts.

4.4 Duration of Activation

The time recording started when the user received the smartphone. When the platform was activated an acoustic signal indicated the start of the commenting process. The duration from the first touch until the acoustic signal was measured.

4.5 Consumer Rating

The texts were rated from 1, very good, to 5, worse. The following criteria were used to analyze the texts:

- Is there a solution for the game stated in the text?
- How long will it take to find a solution for the game?
- Is the comment short and structured?
- Is the comment understandable?
- Is it possible to follow the comment?

5 Results

In this chapter the results of all three parts of the study starting with the questionnaire are presented.

5.1 Questionnaire

60 people answered the questionnaire and rated the priority of the six criteria. (Figure 4) The most important criterion for the participants was that the texts are understandable. The second most important aspect was that they could learn something new

regarding the product. More than half of the participants also considered correct grammar of the comments and that the name of the product is mentioned to be important. It was considered to be unimportant if punctuation and capitalization is correct. **It can be summed up that it is very important for the comment generator that the text content is understandable.**

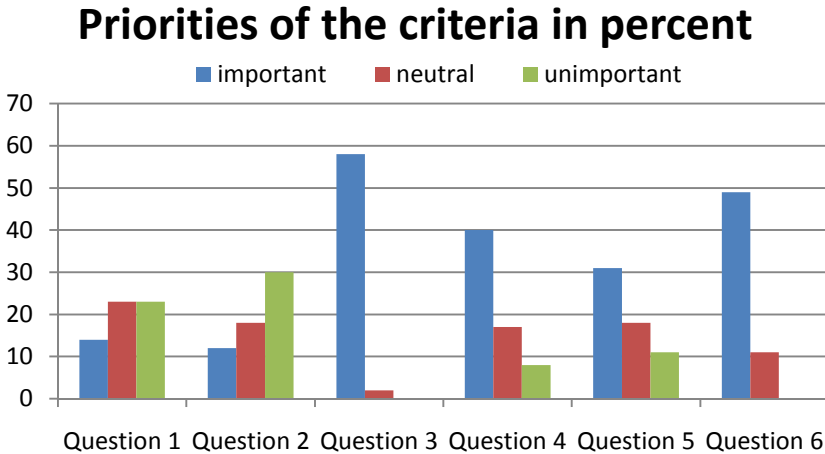


Fig. 4. Priority allocation of the questionnaire in percent

Question 1: No unnecessary use of interpunctuation. (Like tipping three or more periods)

Question 2: The correct usage of capitalization

Question 3: The text difficulty

Question 4: The correct usage of grammar

Question 5: The name of the product appears

Question 6: That something new is added, that you cannot find in the product description.

5.2 Text Analysis

The experiment with the “*Minigames*” showed that there is a difference between expectations and the actual comments provided by the participants because there are different results for the possible platform combinations although there is no significant difference regarding the interactivation by NFC or QR-Code.

As figure 5 depicts the best value in terms of understandability was achieved by both activation technologies NFC and QR codes on the mobile platform “text in combination with rating”. The lowest results were for NFC with “only text” and for QR-Codes with “text with photo-upload”.

5.3 Consumer Rating

The consumer ratings introduced a new point of view. The two raters perceived the comments from the mobile platform “text with rating, like button and photo upload” as bottom quality. The detailed results are shown in figure 6.

Flesch Reading Ease Grades Ø

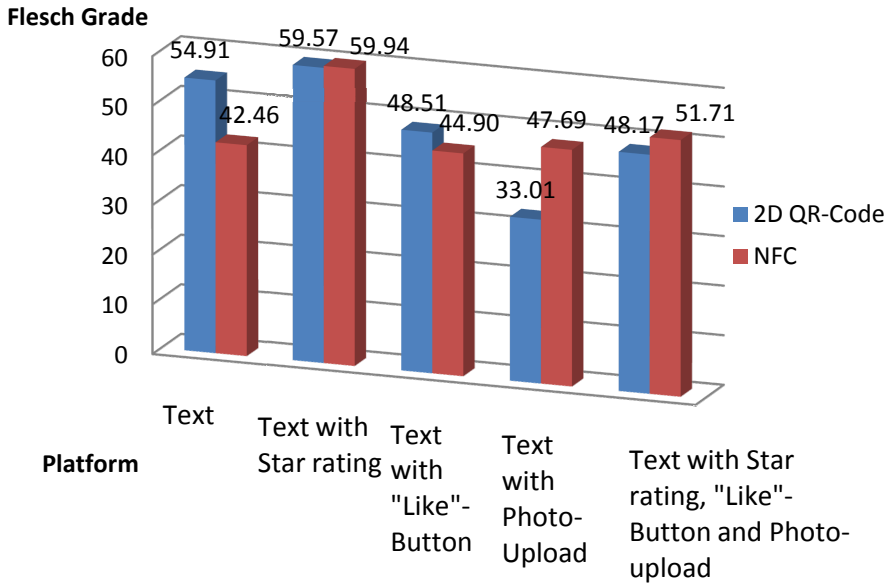


Fig. 5. Average Flesch Reading Ease Grade

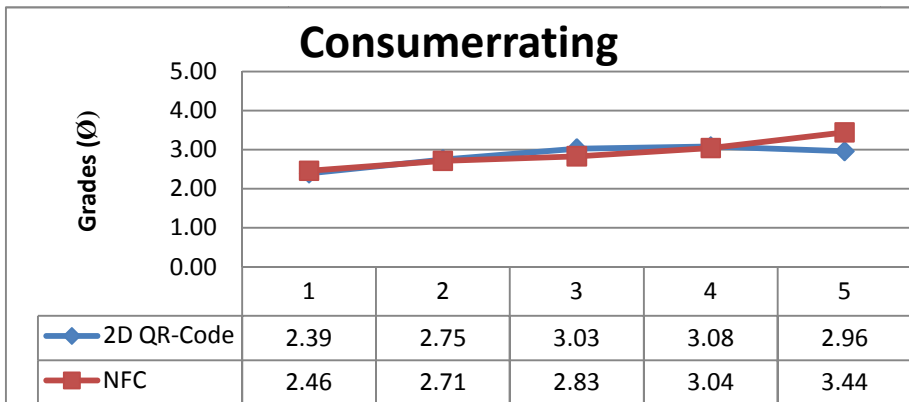


Fig. 6. Average consumer rating

- 1: Text
- 2: Text with Star rating
- 3: Text with "Like"-Button
- 4: Text with Photo-upload
- 5: Text with Star rating, "Like"-Button, Photo-Upload

A correlation test between consumer rating and Flesch value showed no consistent picture. Table 1 lists the correlation coefficients and corresponding statistical significance.

Table 1. Correlation of consumer rating and Flesch Reading Ease Score

| Platform | Correlation (Spearman's Rho) | Significance | Sample size |
|---|---------------------------------|--------------|-------------|
| Text | 0,193 | 0,176 | N=50 |
| Text with star rating | 0,257 | 0,092 | N=44 |
| Text with "Like" button | 0,336 | 0,023 | N=46 |
| Text with photo upload | 0,044 | 0,769 | N=48 |
| Text with star rating, "Like" button and photo upload | -0,288 | 0,038 | N=52 |

6 Conclusions and Limitations

In Table 2 the results are summed up and linked to initial research questions:

Table 2. Overview of research questions and results

| Question | Answer |
|--|---|
| Which criteria do the regular generators of comments consider as important for formal quality of comments? | As the questionnaire shows the most important criterion is understandability, followed by topicality and novelty. |
| Is there a difference in formal quality between expectations and the actual comments? | The evaluation of the texts generated in the experiment shows that there is a difference between the expectations and the actual quality of comments. |
| Do consumers have different expectations regarding actual quality? | The consumer rating shows that the consumers of comments have different expectations than the generators of comments. |

The experiment showed that there are differences between expectations regarding comments and the actual comments that were provided. The expectations of generators and readers of comments are not exactly the same. Platform providers will have to decide if they want to focus more on the generator or the reader of product reviews. As the results indicate it is good to provide a platform including text and star rating as this is the combination preferred by generators. This combination also received the second best rating from readers.

Unlike expected there are no differences between comments on platforms activated by NFC or QR codes. Nevertheless further research is needed in this field as NFC is not as commonly used as QR codes. Actually it seems that the kind of interactivation does not influence formal quality of comments. In accordance with initial expectations it was possible to observe that the interactivation by means of NFC was faster than by QR codes after a learning process.

As comments and reviews on products and services are supposed to effect the consumers decision making process another important question arises in the context of the consumer rating. Which platform design affects the readers most and what kind of comments makes them buy the products? This will be an important topic for further research because it will provide deeper insights in readers' expectations and their actual activities. In this context it will also be interesting to find out which kind of mobile platforms induce users to generate more positive recommendations.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 183–194 (2008)
2. Chen, Z., Xu, Y.: User-Oriented Relevance Judgment: A Conceptual Model. In: Hawaii International Conference on System Sciences, Hawaii, vol. 4, 101 p. (2005)
3. Flesch, R.: A new readability yardstick. *Journal of Applied Psychology* 32(3), 221–233 (1948)
4. Hoegg, R., Martignoni, R., Meckel, M., Stanoevska-Slabeva, K.: Overview of business models for Web 2.0 communities. In: GeNeMe, Dresden, pp. 23–37 (2006)
5. Jensen, C.S., Vicente, C.R., Wind, R.: User-generated content: The case for mobile services. *Computer* 41(12), 116–118 (2008)
6. Leung, L.: User-generated content on the internet: an examination of gratifications, civic engagement and psychological empowerment. *New Media & Society* 11(8), 1327–1347 (2009)
7. Mihm, A.: Sprachstatistische Kriterien zur Tauglichkeit von Lesebüchern. *Linguistik und Didaktik* 4, 117–127 (1973)
8. Nardi, B.A., et al.: Why We Blog. *Communications of the ACM* 47(12), 41 (2004)
9. von Reischach, F., Dubach, E., Michahelles, F., Schmidt, A.: An evaluation of product review modalities for mobile phones. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 199–208 (2010)
10. Trammel, K.D., Tarkowski, A., Hofmokl, J., Sapp, A.M.: Rzeczpospolita blogów [Republic of Blog]: Examining Polish Bloggers Through Content Analysis. *Journal of Computer-Mediated Communication* 11(3) (2006)

A Method Based on Congestion Game Theory for Determining Electoral Tendencies

Guillermo De Ita¹, Luis Altamirano¹,
Aurelio López-López², and Yolanda Moyao^{1,*}

¹ Faculty of Computer Sciences, BUAP, México
{deita, altamirano, ymoyao}@cs.buap.mx

<http://www.cs.buap.mx/>

² Instituto Nacional de Astrofísica, Óptica y Electrónica
allopez@inaoep.mx

<http://ccc.inaoep.mx/>

Abstract. We present a novel method to study the tendencies of vote in sectorial democratic elections. Our method is intended to determine the relevant profiles characterizing the political behavior of voters. Those profiles allow us to model how the voters, in a specific election organized by sectors, make their vote decision. Furthermore, the same set of profiles are used for representing the different strategies applied by the candidates that compete in the election.

We apply congestion games theory to simulate the distribution of the votes among the candidates, describing an automated way to estimate the likely number of votes for each candidate. Therefore, we can determine who will be the winner candidate of the election, according to a specific political scenario. We report the application of our model to simulate the elections of a director in a university setting, obtaining estimations very close to the actual outcomes.

Keywords: Social Behavior Modeling, Social Simulation, Electoral Simulation, Congestion Games, Multi-Agent System.

1 Introduction

In Artificial Intelligence (AI), the application of intelligent agents has brought a great deal of commercial interest, and it has shown useful for decision making. As more and more commercial transactions are performed on networks, there is a growing interest in designing smart autonomous agents performing specific actions. One of the possible applications of intelligent agents is to simulate specific human tasks. For example, an important human task has been the selection of a representative from a population.

The selection of a representative is both an important and common issue in democratic systems, for instance; the candidate of a political party, the head of a

* This research was partially supported by PROMEP. The first and third author was in addition partially supported by SNI, México.

department, a local city mayor, and so on. This process refers to the need to elect the best representative within a group of people, according to their perception expressed as votes.

Different mathematical formalisms have been developed to describe electoral systems and outcomes by modeling both voting rules and human behavior (see e.g. [1, 4, 10-12]). In [5, 11], an analysis on the winning coalition structure of an election system is done as a simple legislative game, considering the importance of relative ideological positions in a legislative decision game, that is, as a non-cooperative game. While in [1, 3, 9], the dynamics of their model is based on applying the search for equilibrium points, which must fulfill the expectations of voters and the optimum policy choices of representatives, assuming stationary environments. In [12], they report an effort to account for political attitudes and beliefs in a computational model, which is based on a psychological theory. This model was compared against another previously proposed model based on Bayesian learning. The new model could reproduce more political attitudes when applied on simulations of a presidential campaign in 2000.

Nowadays, demoscopic studies (opinion polls) are accomplished in order to determine some electoral preferences. Those surveys, as snapshots of a moment, allow us to make predictions for a very short term. An opinion poll, in its traditional elaboration form, usually reflects outlying questions about the candidates, and about the political competition, such as; popularity indexes, perceptions on the nature of the candidates or their images, the impact of their campaigns, mottos, etc. Often the factors that are measured through those surveys point more to the interest of the candidates or their parties, than to the interest or perception of the voters.

Due to this panorama that lacks the necessary analytic tools for studying electoral tendencies, we propose a simulation system of the political behavior for certain voter segments in accordance to the changes of strategies that the candidates perform during their campaign.

A key element to find the winner of an election is to recognize the profiles characterizing the voters, given that all agents form their set of strategies based on promises and actions which try to influence the voters. Those promises and actions are reflected via a set of weights assigned to the profiles characterizing the voters.

We consider each one of the candidates as an independent player with his own strategies. Each agent competes against each other in order to win a political election, i.e., the political election is a competition among all the players. As the profiles used for characterizing the voters, and forming the strategies of the agents have a limited nature, then a congestion game is formulated.

Thus, we can consider our logical model as a formulation of a congestion game [3, 9, 13]. A congestion game consists of a set of players, a set of resources, strategies for each player, and a cost function associated to each resource. A state of the game is defined by the strategies each of the players has selected, where each of the players is assumed to act selfishly, trying to minimize his individual cost. A congestion game can be modelled as a congestion network, i.e. a triple

$(\mathcal{A}, \mathcal{P}f, k)$, where \mathcal{A} is the set of n players, $\mathcal{P}f$ is the set of limited resources, and k is the increasing cost function which depends on the number of players using the same resource. As mentioned, every player $A_i \in \mathcal{A}$ has to choose a strategy which allows to decrease its cost function value. A state $e = (s_1, \dots, s_n)$ is reached when each player selects a strategy s_i . In a congestion network, several players simultaneously aim at allocating sets of resources. The cost of a resource (one edge of this network) is given by a function of the congestion, i.e. the number of agents using the same resource. So, each agent $A_i \in \mathcal{A}$ chooses a strategy forming a state $e = (s_1, \dots, s_n)$ and the cost function is computed for all limited resource and according to the state e .

Building a congestion game for the problem at hand allows us to perform a search for the singular points in the competition system [13], enabling to predict the possible winner of the specific election. In each singular point of the contest, we can determine the winning strategy and the candidate who will obtain the maximum number of sectors, that is, we can determine who will be the winner, as well as the winning strategy.

The paper is organized as follows. Section 2 explains how voters are characterized. We discuss a method for determining preferences in profiles of a sector of voter in Section 3. The process of defining the strategies of competitors is detailed in Section 4. In Section 5, we present how to compute the estimation of number of votes for candidates. Finally, Section 6 includes conclusions and further work.

2 Characterizing a Population of Voters

Our method starts by considering that there exists a population of voters distributed in k sectors, let $Pot = \{Z_1, Z_2, \dots, Z_k\}$ be the k voter sectors. Let $WZ_i = |Z_i|$ be the number of voters within the sector Z_i . We assume that the cardinalities $WZ_i, i = 1, \dots, k$ are known values.

Let $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ be a set of n intelligent agents. $A_i \in \mathcal{A}$ represents one of the competitors contending for a position or, in political terms, a candidate competing in an election. A sector $Z_i, i = 1, \dots, k$ is favorable to the candidate who obtains the majority of the votes, with respect to the number of votes obtained by the other candidates.

We describe herein a novel method to determine the competitor who obtains the maximum number of sectors in a democratic election. A relevant element in our method is to determine the main "profiles" used for characterizing the political behavior of the members of each sector, as well as to determine its relative political importance among them. Usually, analysts can approximate those profiles and their relative positions, after analyzing previous elections and carrying out a deep study on the political behavior of the voter population.

We represent the profiles used to characterize voters via a discrete set $\mathcal{P}f = \{P_1, \dots, P_m\}$. The elements of $\mathcal{P}f$ are called "profiles". Such profiles are the key objects used to characterize how the voters base their vote decisions. The members of each sector are characterized and identified by specific values on the

set of profiles which represents the main political characteristics of the members in that sector.

In our method, the set $\mathcal{P}f$ is also used to represent the set of strategies applied by the competitors to attract the preferences of the voters. The quantity on a profile that an agent believes to have in certain features (profile entry) is represented by a weight. An agent (and sometimes, his campaign team) organizes the profiles and their corresponding weights in different ways, creating in this manner different political programs to be applied, in accordance with the answer of the population to those programs.

We call to each of the agent programs a *strategy*. An agent applies one of its strategies to compete with other agents in order to obtain a maximum number of sectors of the population. A strategy s_i of A_i is a set of pairs: $s_i = \{(P_1, w_{i1}), (P_2, w_{i2}), \dots, (P_m, w_{im})\}$, where each weight $w_{ij}, j = 1, \dots, m$ tries to reflect the importance of the profile P_j that the agent A_i assigns in one of its political programs (s_i).

In fact, each agent (or his advisers) has to determine which profiles (and their corresponding weights) the agent should promote, and he also has to plan how to arrange those profiles in his political program. According to the different scenarios or to the results obtained through the opinion polls during campaign, as well as to the agent knowledge about the preferences of the voters, the agent selects one of his strategies.

Let $S(A_i) = \{s_{i1}, s_{i2}, \dots, s_{in_i}\}$ be the set of different strategies that the agent $A_i \in \mathcal{A}$ can apply for attracting voters. Once all agents have chosen one of their strategies $s_i \in S(A_i), i = 1, \dots, n$, a state (an action now in a multi-agent system) is formed $e = (s_1, \dots, s_n) \in S(A_1)X \dots XS(A_n)$. Let $\mathcal{S} = \{e_1, \dots, e_o\}$ be the set of different states of the multi-agent system.

Then, a state $e_j, j = 1, \dots, o$ is one of the possible configurations of the multi-agent system, and according to the strategies applied by the different agents, they can obtain a certain number of votes in each state. As the agents change their strategies during campaign in order to obtain more votes, they interactively form new states in this multi-agent system.

While more and more agents utilize the same limited profile, such profile tends to saturate, and its influence on the voters will also become smaller and smaller. Thus, a congestion game is an ideal formalism for modeling this kind of resource sharing [3, 9].

In our study, we consider how to distribute the votes among the agents in a specific state $e \in \mathcal{S}$. Thus, we develop a method to estimate the number of votes obtained by each agent, according to the strategy that each one applies.

We applied the method to model the election of the Director of the Faculty of Computer Sciences (FCC) in the State University of Puebla (BUAP), México, realized in March, 2011. During this election, the process was organized in 11 sectors; 5 belonging to faculty members (professors), 5 to students and 1 to administrative workers. We applied the model to recognize the main political preferences of the students when they make their political vote.

In the FCC, there are two undergraduate programs: Computer Sciences and Computing Engineering. For each program, there are two sectors: Basic and Advanced. Then, there are four sectors at undergrad level; Basic_Eng, Advanced_Eng, Basic_Cs, Advanced_Cs, and additionally one sector at graduate level (M. Sc.): Graduate.

There are about 1900 students at the undergraduate level and 46 students at graduate level. There are 117 faculty members and 15 administrative workers. In this election, 1448 of undergrad students and 40 graduate students cast their voted. While 100% of administrative workers and 111 professors voted.

The political preferences of professors and administrative personnel during the elections in the FCC were captured via classical opinion polls, this due to the size of those sectors. In fact, the largest size of any of those sectors was 32 professors. Then, we could collect, for all professors and administrative personnel, their political preferences.

On the other hand, the size of the student population and their vague answers for determining only one preferable candidate, generated the adequate scenario to validate our model. Thus, we simulate the tendencies of the vote just for the five student sectors.

3 A Method for Determining a Hierarchy of Preferences on the Student Profiles

An opinion poll was applied to the students in order to assign a relative importance order on the student profiles, according to what they considered from more to less important for making their vote decision [2].

We found that the following twelve features in profiles were the most important for students to consider when they decide their vote:

P_1 : Opinion of classmates, P_2 : Opinion of academic advisers, P_3 : Opinion of course instructors, P_4 : Opinion of political student groups, P_5 : Opinion of official administration, P_6 : Commitment shown by the candidate, P_7 : Academic background of candidate, P_8 : Political group supporting the candidate, P_9 : Political work during the campaign, P_{10} : Possible contact with the candidate, P_{11} : Image and confidence shown by the candidate, P_{12} : If the students are in favor of reelection of director in post.

Given a sector $Z_i \in Pot$, a weight wz_{ij} for each profile $P_j \in Pf$ is computed based on responses obtained in an opinion poll. We processed the opinion poll and we computed the average of the responses. Thus, a list of values wz_{ij} were obtained as the average of the values assigned to the profile $P_j, j = 1, \dots, m$ by the selected sample of each student sector $Z_i, i = 1, \dots, k$.

An adjustment based on minimum squares was applied to the sample in order to eliminate 'false positives'. The false positive cases are represented as responses of voters who do not want to cooperate with the opinion polls, either when they try to be tricky, submit contradictory answers, or have apathetic responses. And such cases are detected when the responses of one poll is quite different to the average of that sector.

An order of relevance is given on the profiles of each sector $Z_i \in Pot$. Let $AvP_j Z_i$ be the relative value given to the profile P_j with respect to the other profile values of the sector Z_i . We want that the values $AvP_j Z_i, j = 1, \dots, m$ represent relative percentages that determine a relative order on the set of profiles characterizing Z_i , and also that the values $AvP_j Z_i$ build a hierarchy among the profiles of the same sector.

In general, there are different methods for determining a hierarchy of preferences among a list of values (see [6-8]). One of the most simple methods for determining an order on the preferences of the individuals is for example, that the values $AvP_j Z_i$ are taken as the percentages of members from Z_i in which the profile P_j is their main profile.

Other simple method for determining relative percentages $AvP_j Z_i$, is to assign the same importance to all profiles and then, each percentage $AvP_j Z_i$ is equal to 100% divided by the number of relevant profiles in Z_i .

In our model, it is important to distinguish between profiles with a positive impact and those which have a negative impact, when the students make their vote decision. In general, the sum of the percentages of the positive profiles is higher than the sum of the percentages of the negative ones, because the positive profiles had a greater influence than the negative ones to decide the vote, in a proportion according to the voting scenario which is being modeled.

For example, according to the atmosphere that prevailed during the election in the FCC, we detected that positive profiles had more impact than negative ones. The sum of the percentages on positive profiles was 100%, while the sum of the percentages on the negative profiles generally produced values from 40% to 50%. So we detected that the total positive profiles influence was twice or thrice than the negative profiles influence, according to the student sector.

We determine, through an analysis of the responses of the opinion polls, that the first seven more valuable profiles ranked by the students have a positive impact, while we give to the remaining four profiles a negative influence. According to the opinion poll applied to the students, the average values wz_{ij} from 1 to 7 were the most significant profiles, 1 being the most relevant profile value. Then, by computing $Pos_{ij} = 11 - wz_{ij}$ we can determine a relative position in the hierarchy of the positive profiles given in ascending order, that is, 4 is the less important and 10 is the most important value on the positive profiles.

Given a fixed sector $Z_i, i = 1, \dots, k$, in order to build a relative order among positive profiles of the sector, the values Pos_{ij} are summed on all positive profiles: $Sum_Positive_i = \sum Pos_{ij}$ and then the relative percentage of a positive profile is defined as: $AvP_j Z_i = Pos_{ij} / Sum_Positive_i$.

The case of the profile 9; " P_9 : effect of the *Political_Campaign*", was considered at the beginning of the election very conservatively, because we did not know beforehand how intense or effective the campaign of the candidates will be. So a conservative assessment was assigned to P_9 with a relative percentage equal to the maximum relative percentage on the other positive profiles.

The last values obtained from the responses of the student poll represent negative profiles. The negative profiles have obtained average values wz_{ij} between

8 and 11. We assign a relative order among them, according to the formula: $Pos_{ij} = wz_{ij} - 7$, since 7 was the turning value between positive and negative profiles. In a similar way as we did for positive profiles, a relative order was computed for the negative ones. First, we add the values Pos_{ij} on all negative profile: $Negative_i = \sum Pos_{ij}$.

A special profile was P_{12} specified as: "Do you agree with the re-election of the director?." In the opinion poll, P_{12} shown a very important negative character since a high percentages of the students reject the idea of the re-election of the current director. So, we assigned a relative value on P_{12} equal to the addition of the position values of all negative profiles previously considered. Then, the relative value for the profile P_{12} was $Pos_{i12} = Negative_i$. Since P_{12} was dominant in the set of all profiles which have considered negative and has an influence equivalent to the addition of the other negative profiles.

Then, we had a total value of the sum of all negative profiles (including profile 12), defined as: $Sum_Negatives_i = \sum Pos_{ij}$ on the set of the negative profiles. And for each negative profile, we compute its relative percentages as: $AvP_jZ_i = Pos_{ij}/Sum_Negatives_i$.

Table 1. Matrix $MPot$: Relative weights to profiles by sectors

| <i>Profile</i> | Basic_Eng | Adv_Eng | Basic-Cs | Adv-Cs | Graduate |
|----------------|-----------|---------|----------|---------|----------|
| P1 | -1.1062 | -2.789 | -2.957 | -0.9381 | -3.06 |
| P2 | -0.9878 | -0.194 | 9.58 | -0.187 | 9.51 |
| P3 | 10.896 | 10.361 | 10.46 | 9.41 | 8 |
| P4 | -1.3414 | -1476 | -1.7 | -2.44 | -1.66 |
| P5 | -1.15 | -2.0354 | -1.73 | -4.41 | -1.22 |
| P6 | 19.565 | 19.458 | 16.64 | 17.47 | 17.956 |
| P7 | 16.197 | 17.651 | 15.96 | 15.46 | 17.07 |
| P8 | 11.677 | 12.168 | 10.91 | 12.5 | 11.11 |
| P9 | 39.13 | 38.916 | 33.278 | 34.95 | 53.867 |
| P10 | 14.92 | 14.94 | 13.563 | 15.6 | 12.889 |
| P11 | 13.958 | 13.615 | 11.364 | 13.844 | 12.62 |
| P12 | -38.646 | -39.124 | -35.463 | -39.5 | -35.31 |

In Table 1, we present the matrix $MPot$ containing the final relative percentages AvP_jZ_i obtained by our ordering method. Negative values are indicators of profiles with a negative impact on the students. Those values show relative percentages among the set of profiles and give a degree of relevance of each profile in the students, when they make their vote decision. Notice that the relevance of each profile is different according to the sector of the student (the five columns of the matrix).

4 Defining the Strategies of Competitors

Central information to model democratic elections is based on the political campaigns (strategies) applied by the competitors during the contest.

In the election that we have modeled, there were two agents competing for the position of director. Of course, the competitors do not know precisely neither the most important profile nor its relevance in the sector. Although, they intuitively recognize the importance of some profiles and they try to influence the voters through their political programs (strategies).

The Matrix St shown in Table 2 contains weights representing the final strategies considered during the simulation of the election. The strategies of each candidate can be considered as a vector of twelve values, each value represents the intention of the candidate to influence the voters through that corresponding profile. Each weight $w_{ij} \in St$ represents the value on the profile P_i that a competitor determines to apply in its strategy to attract votes.

Since the two competitors applied different programs and promises between undergraduate and graduate student sectors, they really used different strategies according to the student academic level (i.e. undergraduate or graduate). The weights constituting the strategies of the candidates were computed based on: their curriculum vitae, the proposals, the political group supporting the candidates, and in this particular case, the knowledge and perception that some of the authors have about both candidates.

Table 2. Matrix St : Final Candidates Strategies for Student Level

| <i>Profile</i> | Dir_Undergrad | Dir_Grad | Opp_Undergrad | Opp_Grad |
|----------------|---------------|----------|---------------|----------|
| P1 | 9.5 | 5 | 5 | 8 |
| P2 | 8 | 9.5 | 5 | 4 |
| P3 | 7 | 9 | 8 | 5 |
| P4 | 8.5 | 8.5 | 5 | 5 |
| P5 | 6 | 6 | 3 | 3 |
| P6 | 7 | 7 | 4 | 4 |
| P7 | 8 | 8 | 5 | 5 |
| P8 | 8 | 8 | 6.8 | 6.8 |
| P9 | 6 | 9 | 8.5 | 5.5 |
| P10 | 7 | 7 | 7 | 7 |
| P11 | 8 | 8 | 6 | 6 |
| P12 | 10 | 10 | 1 | 1 |

For any other election, opinion polls can be designed to calculate the corresponding weights in order to specify in a vector of weights the agent strategies. Each agent applies one of its strategies creating a state e of the multi-agent system. The agents change their strategies according to the opinion polls that they are aware of.

For each state $e \in \mathcal{S}$, the voters make their vote decision and then every candidate obtains a determined number of votes in concordance with its strategies. In a dynamic way, any candidate could change his strategy with the intention to obtain more votes, forming in this way the different sceneries along the election process.

Given a state $e = (s_1, \dots, s_n) \in S$, an *improvement step* for an agent A_i is a change of strategy from s_i to s'_i going to a new state e' and where his percentages of votes increases with respect to the previous value. The neighborhood of a state e consists of those states that derive from e only in one change of the agent strategy [3].

Given the two tables $MPot$ and St both of m rows (m profiles), and the two different strategies applied by the candidates, one for undergraduate students and the other for graduate students. We show in the following section, how to compute automatically the number of votes expected by each candidate, according with its strategy and the characterization of the sectors.

5 Computing the Number of Votes for Each Candidate

For the first four sectors $Z_i, i = 1, \dots, 4$, corresponding to undergraduate student sectors, an addition on their relative percentages multiplied by the corresponding strategy of the two candidates is done on the set of 12 profiles, that is, $Ug_{i1} = \sum_{j=1}^{12} (w_{j,1} * AvP_j Z_i)$ and $Ug_{i2} = \sum_{j=1}^{12} (w_{j,2} * AvP_j Z_i)$, where $s_1 = (w_{1,1}, \dots, w_{12,1})$ is the undergrad strategy applied by the candidate 1, and $s_2 = (w_{1,2}, \dots, w_{12,2})$ is the undergrad strategy applied by the candidate 2.

Let $StPr_i = Ug_{i1} + Ug_{i2}$ be the total of the set of profiles to be shared among all agents. In general, $StPr_i = \sum_{j=1}^n Ug_{ij}$ expresses the total influence of the profiles on sector Z_i to be shared among the candidates. And, according to congestion game theory, the proportional part that a candidate $A_l \in \mathcal{A}$ has to receive when he applies a strategy on the undergraduate sectors is computed as $S(e, A_l, Z_i) = (Ug_{il}/StPr_i)$ for $l = 1, \dots, n$ and $i = 1, \dots, 4$.

We compute the case for the graduate student sector: Z_5 , in a similar way that the case of undergraduate sectors but now we consider the graduate strategies applied by the candidates. Then, if now $s_1 = (w_{1,1}, \dots, w_{12,1})$ is the graduate strategy applied by the first candidate and $s_2 = (w_{1,2}, \dots, w_{12,2})$ is the graduate strategy applied by the second candidate, and so, $Grad_{51} = \sum_{j=1}^{12} (w_{j,1} * AvP_j Z_5)$ and $Grad_{52} = \sum_{j=1}^{12} (w_{j,2} * AvP_j Z_5)$ will be the contribution of the candidates to the graduate sector.

Additionally $Ps_5 = Grad_{51} + Grad_{52}$ represents the total of the set of profiles to be shared among all agents. And the proportional part that a candidate $A_l \in \mathcal{A}$ has to receive when he applies his strategy on the graduate sector is: $S(e, A_l, Z_5) = (Grad_{5l}/Ps_5)$ for $l = 1, 2$.

Given that we are considering that the cardinality of each sector $WZ_i, i = 1, \dots, 5$ is known and they are constant values, and although the agents promise more than before (his strategies $s_i, i = 1, \dots, n$ have higher values), the values WZ_i do not change, then a congestion game is modeled to distribute a fixed value of votes among the agents [9].

Then, the cardinalities $WZ_i, i = 1, \dots, 5$ have to be divided proportionally among all agents due to $S(e, A_l, Z_i)$. Given a state $e \in S$ and for all sectors $Z_i \in Pot$, for each agent $A_l \in \mathcal{A}$, we denote as the percentages of voters from

the sector Z_i which are potential voters for the agent A_l , as $\#Vote(A_l, Z_i)$, and that value is computed as:

$$\#Vote(A_l, Z_i) = W_{Z_i} * \left(\frac{S(e, A_l, Z_i)}{100} \right) \tag{1}$$

The value $\#Vote(A_l, Z_i)$ represents the percentages of members in the sector Z_i which are potential voters for $A_l, l = 1, \dots, n$. Then, $\#Vote(A_l, Z_i)$ has to be computed for all sector $Z_i \in Pot$ in order to know the percentages of votes that all agent can obtain in each sector.

When the election is by sectors, the candidate who obtains a maximum percentages of votes in that sector is the candidate who wins the total sector. Then, fixing a sector $Z_i \in Pot$, the candidate who wins Z_i is defined as the candidate $A_q, q \in [1, n]$ such that:

$$\#Vote(A_q, Z_i) = \max\{\#Vote(A_l, Z_i), l = 1, \dots, n\} \tag{2}$$

Notice that given a state e , there is an agent who wins the maximum number of sectors, we call such an agent *the candidate in the state e*, and is denoted as the competitor $A_q, q \in [1, n]$ such that the number of sectors $Z_i, i = 1, \dots, k$ that A_q has won is maximum on the set of competitors.

Although to change an agent strategy (even if the *the candidate in the state e* does not change his strategy) represents a change in the state from e to e' , and the candidate who wins a maximum number of sectors could change too. The improvement over the number of votes of an agent A_i is necessary to obtain a higher number of sectors for him. Given a state e , a move of improvement for an agent A_i through local values is done by the search of a neighbor e' where A_i wins a higher number of sectors than in the state e .

In our system, we can analyze the fluctuations of the votes tendencies in order to organize the strategies of a specific agent, either as 'bad' or 'good' strategies, according to the number of sectors that the agent wins. Furthermore, we can find which are the best strategies for a particular candidate, according to a specific electoral scenario.

Assuming that all people really vote, we have a fixed total number of votes and, if we look for an optimal point, the search could turn cyclic, meaning that if an agent reduces his number of votes, then any other agent will increase his own number of votes. So, some agents could always improve their number of votes from one neighbor to another.

An adequate variable to avoid a cyclic search is to consider the percentage of potential voters who abstain in each sector. Although the abstention is a fact in democratic systems, to determine this percentage requires a profound analysis of the traits and behavior of the population in previous elections. In our system, the political campaign is developed during a certain period, in such a way that when a candidate recognizes a new way to improve his strategy, that new strategy is applied and then, the likely number of votes have to be re-computed for all the involved candidates. This continues until no further impact can be produced on the number of votes or when the election campaign is over.

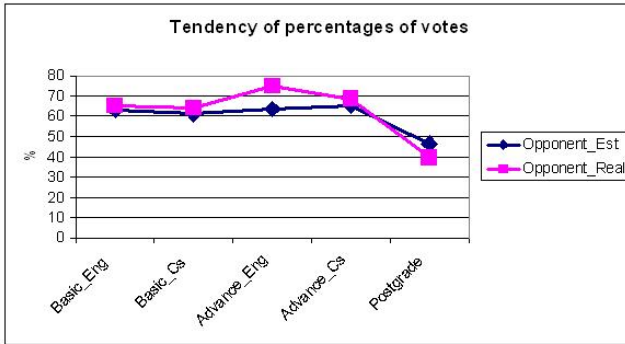


Fig. 1. Percentages of votes obtained by opponent to incumbent

We analyze the percentages of the votes assigned to each candidate according to the strategies applied by the agents. The formulas presented in this section allow us to estimate the number of votes for each agent, and to know who will be *the candidate in the state e* according to the current scenario e .

Comparing the estimated results at the end of the campaign versus the actual percentages of votes in the modeled election, the absolute errors on the percentages of the votes obtained for the candidate labeled as *Opponent to incumbent*, were: 2.5, 2.5, 11.5, 3.6, 7.2, which correspond to the sectors: Basic_Eng, Basic_Cs, Advance_Eng, Advance_Cs and Graduate (labeled as *Postgrade*) sectors, respectively, as depicted in Figure 1.

If we want to model elections at the scale of, for example, city mayor, the key issue in our proposal is the partition of the voters into sectors (e.g. retired people, workers, government employees, housewives, etc.) with common and recognized interests and needs (profiles). That implies that we do not only have to know the sizes of the sectors, but we also have to analyze the political and economical historical behavior of those sectors.

The demoscopic studies can be helpful for recognizing the profiles and their relative importance among them. Of course, this implies a bigger effort than only applying the common opinion polls for analyzing political preferences. However, more precise predictions request deeper studies and our model can serve as a guide for performing those studies.

6 Conclusions

We have designed a model for simulating the process of selecting a representative in a democratic system organized by sectors. Our system can be used to study the tendencies of the vote and for this, it is necessary to determine the relevant profiles that characterize the political behavior of voters. Those profiles model how the voters, in a specific election organized by sectors, make their vote decision.

In our proposal, we assume that each candidate determines a finite set of strategies (political programs). Each one of these strategies is formulated as a set of weights on the profiles characterizing the voters.

The profiles used for characterizing the voters, and for expressing the strategies of the agents, have a limited nature. Consequently, a congestion game is formulated. We present a model to simulate the distribution of the votes among a set of candidates, allowing so, to determine who will be the winner in a specific political scenario.

We have applied our model to simulate the elections of a director in a university setting, obtaining estimations very close to the actual outcomes. Future works includes the extension of the model to consider other election scenarios and the application of our model in other electoral process to confirm its validity.

References

1. Banks, J., Duggan, J.: A dynamic model of democratic elections in multidimensional policy spaces. *Quarterly Journal of Political Science* 3, 269–299 (2008)
2. De Ita, G., Moyao, Y., Contreras, M.: Modeling Democratic Elections via Congestion Networks. In: *First Int. Conf. on Social Eco-Informatics*, vol. 1, pp. 85–90 (2011)
3. Feldmann, A.E., Röglin, H., Vöcking, B.: Computing Approximate Nash Equilibria in Network Congestion Games. In: Shvartsman, A.A., Felber, P. (eds.) *SIROCCO 2008*. LNCS, vol. 5058, pp. 209–220. Springer, Heidelberg (2008)
4. Gill, J., Gainous, J.: Why does voting get so complicated? A review of theories for analyzing democratic participation. *Statistical Science* 17, 1–22 (2002)
5. Jackson, M., Moselle, B.: Coalition and Party Formation in a Legislative Voting Game. *Journal of Economic Theory* 103(1), 49–87 (2002)
6. Morales, P.: Evaluación de los valores: análisis de listas de ordenamiento, CTAN (2011), <http://www.upcomillas.es/personal/peter/otrosdocumentos/ValoresMetodo.pdf>
7. Morales, P.: Cuestionarios y Escalas, Universidad Pontificia Comillas, CTAN (2011), <http://www.upcomillas.es/personal/peter/otrosdocumentos/CuestionariosyEscalas.doc>
8. Pajares, F.M.: Teachers' Beliefs and Educational Research: Cleaning up a Messy Construct. *Review of Educational Research* 16, 307–332 (1992)
9. Quant, M., Borm, P., Reijnierse, H.: Congestion network problems and related games. *European Journal of Operational Research* 172, 919–930 (2006)
10. Quinn, K.M., Martin, D.: An Integrated Computational Model of Multiparty Electoral Competition. *Statistical Science* 17(4), 405–419 (2002)
11. Strom, K.: A Behavioral Theory of Competitive Parties. *American Journal of Political Science* 34, 565–598 (1990)
12. Kim, S., Taber, Ch.S., Lodge, M.: A Computational Model of the Citizen as Motivated Reasoner: Modeling the Dynamics of the 2000 Presidential Election. *Polit Behav.* 32, 1–28 (2010)
13. Fabrikant, A., Papadimitriou, Ch.S., Talwar, K.: The Complexity of Pure Nash Equilibria. In: *Procs. STOC 2004*, Chicago, Illinois, USA, June 13–15, pp. 604–612 (2004)

A Model to Represent Human Social Relationships in Social Network Graphs

Marco Conti, Andrea Passarella, and Fabio Pezzoni

CNR-IIT, via G. Moruzzi, 1 - 56124 Pisa, Italy

{marco.conti, andrea.passarella, fabio.pezzoni}@iit.cnr.it

Abstract. Human social relationships are a key component of emerging complex techno-social systems such as socially-centric platforms based on the interactions between humans and ICT technologies. Therefore, the models of human social relationships are fundamental to characterise these systems and study the performance of socially-centric platforms depending on the social context where they operate. The goal of this paper is presenting a generative model for building synthetic human social network graphs where the properties of social relationships are accurately reproduced. The model goes well beyond a binary approach, whereby edges between nodes, if existing, are all of the same type. It sets the properties of each social link, by incorporating fundamental results from the anthropology literature. The synthetic networks it generates accurately reproduce both the macroscopic structure (e.g., its diameter and clustering coefficient), and the microscopic structure (e.g., the properties of the tie strength of individual social links) of human social networks. We compare generated networks with a large-scale social network data set, validating that the model is able to produce graphs with the same structural properties of human-social-network graphs. Moreover, we characterise the impact of the model parameters on the synthetic graph properties.

Keywords: social networks, human behaviour, modelling, simulations.

1 Introduction

In the last decade the proliferation of personal mobile devices, e.g. the smartphones, led to the emergence of electronic pervasive social networks which are drastically changing the way the information is circulating. In particular there is a convergence between the cyber/virtual and the physical world. Indeed, content generated in the physical space produces outcomes in the cyber/virtual world and, similarly, information generated in the cyber space has immediate influence on the physical world. At the core of this convergence there are humans which, through their devices, transfer the information between the physical and the cyber space in both directions. The analysis of the human social behaviour is therefore becoming fundamental for the development of socially centric platforms [1], whereby the properties of the social relationships between users are taken into account in the core design of the communication algorithms.

In addition, information technologies can also be used as tools to generate simulated social environments where properties of human social relationships can be studied “in vitro”, under controllable parameters. For example, accurate models of human social networks can be used to study information dissemination or opinion spreading at large scale and under a range of parameters’ values.

In this work we present a model for the generation of synthetic social networks whose structure reproduces the main properties of human social networks. It starts from the model presented in [2] which is able to generate single ego networks (a simple form of social network) based on well-known results in the field of anthropology. We extend the original model in order to generate complete social networks formed by interconnecting ego networks. With this purpose, the model relies on well-known properties in the social networks literature, such as the “triadic closure”, the presence of bridges and geographical constraints [3, 4]. The parameters of the model permit to generate different social networks tuning the geographical constraints and changing the criteria the individuals use to create new social relationships. Experimental results demonstrate that generated networks accurately match the properties of human social networks. Specifically, we show that our model is able to reproduce both macroscopic properties of the network, such as its diameter and its clustering coefficient, but also microscopic properties, such as the strength of the tie of individual social links, and the correlation between the tie strength of different social links.

The use of this model for generating synthetic social network has several practical applications. On the one hand, it is a tool for accurately studying processes of social interaction via simulations. For example, it is possible to analyse variations of the information diffusion process using different settings of the model parameters. On the other hand, the model permits the development and the performance evaluation of algorithms and protocols for socially centric platforms and systems.

The remainder of this paper is organised as follows: in Section 2 we give an overview of the results regarding human social networks; in Section 3 we summarise the model for the generation of single ego networks and then we introduce the new model for the generation of complete social networks; in Section 4 we validate our model comparing different generated networks with a real human network; finally, in Section 5 we draw the main conclusions of our work.

2 Background and Related Work

The study concerning the composition and the structure of human social networks are arousing the interest of an increasing number of researchers in many different fields [2, 3, 5–12]. Significant attention has been devoted to ego networks, which are social networks between an individual (*ego*) and the other people (*alters*) the ego has a social relationship with [5]. Despite being small-size networks, ego networks are important as they permit to fully characterise the properties of social links between individuals.

One of the main results about ego networks is that their structure consists of a series of concentric layers of acquaintanceship with increasing size [6]. Based

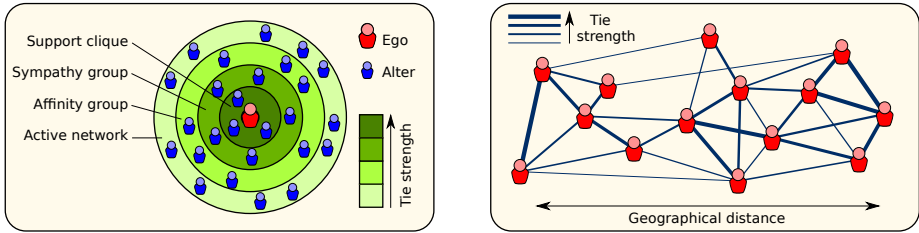


Fig. 1. (a) Ego network structure; (b) Complete social network

on data collected on real human networks, Dunbar et al. [12] identify four layers: “*support clique*”, “*sympathy group*”, “*affinity group*” and “*active network*” (the whole ego network) with average sizes of ~ 5 , ~ 12 , ~ 35 and ~ 150 respectively. The sizes are evaluated considering that layers are inclusive. Sometimes in this paper, we use the term *external part* of a layer in order to refer to the part of the layer not overlapped with its inner levels (e.g., for the sympathy group it is the part of the layer not overlapped with the support clique). Going from inner to outer layers, while the number of alters increases, the strength of the social tie between the ego and the alters diminishes. This means that, typically, an ego has few very strong social relationships in the support clique and a lot of weak ties in the active network (external part). The hierarchical structure of an ego network is depicted in Figure II(a).

It has been shown that this hierarchical structure and the typical sizes of the layers are related to the level of *emotional closeness* (the strength of a social tie) and the cognitive resources humans allocate to social relationships. Intuitively, maintaining a social relationship has a cognitive cost (e.g., due to spending time together, remembering facts about the alter, etc). As the total “cognitive capacity” humans devote to social relationships is limited, the sizes of the layers are also limited [8, 9]. Other results regard the composition of each layer of the ego network with respect to the gender of the alters and to the family relationships [8, 10]. The authors of [2] define a model which allows to generate synthetic ego network graphs that satisfy all these properties. The procedure for the generation of these graphs is summarised in Section 3.1.

While being a very important model for studying certain properties of social relationships, ego networks alone cannot provide a complete representation of human social networks. Indeed they do not capture the mutual relationships between the alters or, in other words, the correlation between different ego networks. This gap can be filled connecting ego networks together in order to form a complete social network as shown in Figure II(b). Due to the high complexity of complete social networks, the characterisation of their properties is way less advanced than that of ego networks. To the best of our knowledge, three main properties have been experimentally characterised in the literature, i.e. (i) triadic closure, (ii) the presence of bridges, and (iii) the dependence of social links on geographical distance.

Properties (i) and (ii) were investigated by Granovetter in [3]. In the paper, the author defines the *triadic closure* as a property of the social networks for which, if a strong social tie exists between two pairs of nodes $A-B$ and $B-C$, there is, with a high probability, a tie between the nodes $C-A$ which closes the triangle. The links in social networks that do not take part in triangles are called “bridges” and, according to the study in [3], they are mainly weak ties. Bridges have an important role in the social network structure as they connect socially distant parts of the network enabling to reach people and information not accessible via strong ties [3]. The presence of bridges leads the diameter of the network to be short, as in the results of the Milgram experiment [11]. At the same time, the triadic closure property guarantees a high level of clusterisation. For these reasons, human social networks can be classified as *small-world networks*, according to the definition given by Watts and Strogatz [13].

The presence of *geographical constraints* (iii) is another key factor in the formation of human social networks. Indeed, for each person, it is more likely to have a social relationship with an individual who lives close to him, than to have a tie with a person who lives far away. This hypothesis is verified experimentally by Onnella et al. in [4]. They analysed a huge data set of social interactions based on mobile phone calls in which each user is tagged with the geographical position where she probably lives. Plotting the frequencies of social ties between users which live at different distances, it emerges that the decay of the tie probability follows a power-law of the form $P(d) \sim d^{-\alpha}$, where d is the geographical distance and α is the power-law exponent. Using the maximum-likelihood method, the authors estimate $\alpha = 1.5$ [4].

In the last five years, thanks to the advent of online social networks (OSNs), the analysis of large social network graphs became more affordable. Indeed most of the recent work in social network analysis focuses on the characterisation of the global properties of a specific OSN, such as Facebook [14–16] and Twitter [17, 18]. Some important results were obtained, e.g. the validation of the “small-world property” [14], the evidence of the Dunbar’s number [17] and the discovery of the power-law distribution of the degree [15]. However, these results are relevant only for the virtual environment since they are strictly related to the particular graph considered. In addition, these analyses and the resulting network models typically do not pay sufficient attention to microscopic features of social links, such as the associated tie strength, but use a binary model where links either exist or not exist (i.e., unweighted graphs).

In this work we define an original approach to social network analysis, by developing a model for the generation of human social networks which, to the best of our knowledge, reproduces the key properties of human social network highlighted in the anthropology literature. In contrast with legacy studies on OSNs we take into account the social aspects which characterise the human social networks, such as the strength of the ties, the cognitive resource consumption of the individuals and the correlation between the strength of ties between different users.

3 The Model

The model described in this section is defined by an iterative procedure able to generate synthetic social network graphs which exhibit the typical features of human social networks described in Section 2.

The procedure operates on two distinct levels of the network structure: the *local level*, in which the ego networks are generated, and the *global level*, in which the ego networks are opportunely connected to form a complete social network. Based on these distinct levels, we can consider our model as the union of two different models: a *single-ego model* and a *multi-ego model* respectively.

The single-ego model is based on the work in [2] which we summarise in Section 3.1. The multi-ego model, which relies on the concepts of triadic closures, bridges and geographical constraints, is described in detail in Section 3.2.

3.1 Single-Ego Model

The model assumes that each ego has a finite budget of cognitive resources for social relationships, expressed as the total time the ego devotes to social interactions. The algorithm adds social links to an ego network, associated with the time devoted by the ego to that particular relationship. The ego network is completed when the ego's total budget is over. The model considers a three-level structure in which layers are called “*support clique*”, “*sympathy group*” and “*active network*” with average size respectively 4.6, 14.3 and 132.5 (reference values are given in [6]). This structure differs from ego network structure defined in Section 2 by the absence of the “*affinity group*” layer. This is justified in [2] by the lack of results about its properties currently available in literature.

The algorithm initialises each ego i with a budget of time bdg , the size of the sympathy group s_{sym} and the size of the support clique s_{sup} . Each of these values is drawn from a carefully defined density function (f_B , f_S and f_W respectively). After the initialisation, the algorithm starts creating new social ties which are characterised by a certain level of emotional closeness, extracted from a density function f_E . The level of emotional closeness is subsequently converted into time by a conversion function h , and then subtracted from the residual time budget bdg . New social relationships are first included in the support clique layer until it reaches the target size, subsequently, in a similar fashion, they are included in the sympathy group (external part). For the external part of the outermost layer, the algorithm adds new social ties until the budget of time is totally exhausted.

Definitions of the density functions f_B , f_S , f_W , f_E and of the conversion function h , summarised in Table 1, are directly obtained from [2].

¹ Note that the model associates a level of emotional closeness to social ties, instead of directly associating a time budget, as the former is the typical way of characterising the strength of social ties in the anthropology literature [8, 9].

Table 1. Functions definition

| Function | Description | Definition |
|----------|---|--------------------------|
| f_B | Time spent by egos in social activity | $Gamma(205, 8.5264)$ |
| f_S | Sympathy group size | $Gamma(4.1, 3.49)$ |
| f_W | Ratio between sympathy gr. and support cl. sizes | $Normal(0.3217, 0.1608)$ |
| f_E | Emotional closeness level ^a | $Normal(0.419, 0.237)$ |
| h | Emotional closeness \rightarrow Time conversion function ^b | $h(e) = 117.18^e$ |

^a We merged together the functions defined in [2] for kin and non-kin. The limits of the intervals of emotional closeness are: $\text{low}_{\text{sup}} = 0.8337$ and $\text{low}_{\text{sym}} = 0.71$.

^b Calculated with the method described in [2] considering f_E .

3.2 Multi-ego Model

The multi-ego model is designed in order generate complete human social networks, in which each node represents an individual whose ego network follows the model described in Section 3.1. In the multi-ego model a node is part of several ego networks with different roles. In this section we first present the high-level strategies the model follows, then we describe the algorithm in detail.

The model considers a human social network as a large group of individuals which are interconnected by social links. Intuitively, the procedure defined by the single-ego model can be applied to each of these individuals in order to generate its ego network. However, applying the single-ego procedure, we have to take into account that each new social link an individual adds to its ego network, also alters the ego network of the other individual involved in the relationship. This means checking, upon creation of a new link, that the properties of the involved ego networks are preserved. In detail, we have to check that (i) the size of the support clique, (ii) the size of the sympathy group, and (iii) the total budget of time remain consistent. Moreover, in order to generate complete ego networks we have to take into account the additional properties described in Section 2, i.e. triadic closure, presence of bridges and geographical constraints.

A new social link can be established either exploiting the triadic closure property or creating a bridge. The strategy to be used is randomly selected based on a given probability. In case the triadic closure strategy is selected, the procedure tries to close a triangle, that is, given an origin node, it selects a node at a distance of 2 hops as link's destination, favouring strong tie hops. On the contrary, in case the procedure follows the bridge creation strategy, the destination node is chosen randomly. In both cases geographical constraints have to be respected. In order to do this, we incorporate geographical information into the nodes, associating to them random locations in a virtual space. Whatever strategy to create links is selected, the model guarantees that the probability to have a social link between two nodes is proportional to a power law of the distance between them. Remember this is consistent with empirical results in the literature [4].

```

1: procedure CREATESOCIALNETWORK( $n, p, f_D, f_B, f_S, f_W, f_E, h$ )
2:   for  $i \leftarrow 1, n$  do
3:      $i \leftarrow \text{CREATEEGO}(f_B, f_S, f_W)$ 
4:      $i.pos \leftarrow \text{EXTRACTFROM}(\text{Uniform}(-1, 1))$ 
5:      $V \leftarrow V + i$ 
6:   end for
7:   for all layer  $l \in \{\text{sup, sym, net}\}$  do ▷ maintaining the ordering
8:     while  $\text{OPEN}(V, l)$  is not empty do
9:        $i \leftarrow \text{random select in } \text{OPEN}(V, l)$ 
10:      if  $\text{RAND}() < p$  then
11:         $j \leftarrow \text{CLOSURESELECT}(i, f_D, \text{OPEN}(V, l))$ 
12:      else
13:         $j \leftarrow \text{BRIDGESELECT}(i, f_D, \text{OPEN}(V, l))$ 
14:      end if
15:       $r \leftarrow \text{NEWSOCIALLINK}(i, j)$ 
16:       $r.e \leftarrow \text{EXTRACTFROM}(f_E \text{ in } (\text{low}_l, \text{up}_l))$ 
17:      update  $E, i.size, j.size, i.dbg$  and  $j.bdg$ 
18:    end while
19:  end for
20:  return  $V, E$ 
21: end procedure

```

Fig. 2. Multi-ego model's algorithm

Algorithm. The pseudo-code of the algorithm used for generating synthetic human social network graphs is shown in Figure 2. The input required by the algorithm consists of: (i) the number of nodes in the network n ; (ii) the probability p to create a new social link using the triadic closure property rather than creating a bridge; (iii) the power-law distribution function f_D which gives the probability to establish a social link between nodes at a specific distance; (iv) the parameters used to define the structure of the single ego networks f_B, f_S, f_W, f_E, h , as required by the single-ego model (see Section 3.1).

In the first part of the algorithm we create and initialise each node i in the network as an ego (lines 2-6). For each node we first call the procedure `CREATEEGO` which sets the size of the sympathy group $i.s_{\text{sym}}$ and the size of the support clique $i.s_{\text{sup}}$. It also assigns the budget of time $i.bdg$ and initialises the counter $i.size$ which is then used to keep track of the total size of the ego network (line 3). We also assign a geographical position of the ego ($i.pos$) which is randomly selected in a given space which, without loss of generality, we assume mono-dimensional, circular and included in the interval between -1 and 1 . This definition guarantees that the distance between any pair of nodes is between 0 and 1 (line 4). Finally, each generated ego is included in the set V (line 5).

After the initialisation of the egos, we start adding social links to the network. First, we create all the social links belonging to all the support cliques, then we continue with the sympathy groups (external part), and finally we add the links of the active networks (external part) (line 7-17). Given the layer l we are populating, the creation of a new social link between two nodes i and j starts

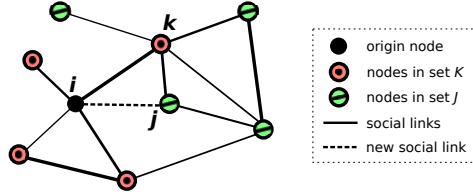


Fig. 3. Triadic closure strategy

with the selection of the node i , drawn randomly from the nodes labelled as *open* (line 9). An “open” node is an ego whose population of the current layer l is not yet completed². The selection of the nodes involved in a new social link from the open node set $\text{OPEN}(V, l)$ guarantees the preservation of the ego network properties. The fundamental part of the algorithm is the selection of node j . We use two different strategies: (i) the *triadic closure* mechanism (procedure CLOSURESELECT) and (ii) the *bridging* (procedure BRIDGESELECT). The former strategy is chosen with a probability given by the parameter p , while the latter with probability $1 - p$ (lines 10-14).

The *bridging*, i.e. the creation of a bridge, is the simplest strategy. We extract a node j from the open egos in the network for the current layer l , excluding the nodes already connected to i , taking into account the geographical constraints. The probability to select a node j is thus proportional to the value of the power-law function f_D (discussed in detail at the end of the section), given the distance $\text{dist}(i, j)$ between i and j . Formally,

$$P(j) \propto f_D(\text{dist}(i, j)) \quad j \in \text{OPEN}(V, l) - \text{Nei}(i) - i \quad (1)$$

where $\text{Nei}(i)$ is the set of one-hop neighbours of node i .

If each node in the network, not connected to node i , is *closed* (not open), node j can not be selected. In this case node i is forced to be closed. We have experimentally checked that this circumstance occurs just in a negligible number of cases and that the overall results are not affected.

Using the *triadic closure* strategy, represented in Figure 3, we first select the set K of the neighbours of i . From this set, we extract an intermediate node k with a probability that is proportional to the tie strength e_{ik} between i and k multiplied, in order to satisfy the geographical constraints, by a function of the distance $\text{dist}(i, k)$ (Equation 2). Given the intermediate node k and the current layer l , we define the set J as the set of open neighbours of k , with respect to l , excluded node i and its neighbours. From the set J we extract node j using the same method used for the selection of node k , considering the social relationship between k and j (Equation 3).

² In case the current layer l is the support clique or the sympathy group, an ego i is open if its ego network size $i.size$ has not reached the thresholds $i.s_{\text{sup}}$ or $i.s_{\text{sym}}$ respectively. In case l is the active network, i is open if it has not exhausted its time budget $i.bdg$.

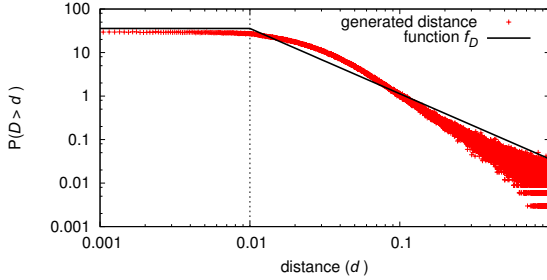


Fig. 4. PDF of the generated distance and the function f_D ($\alpha = 1.5$, $d_{min} = 0.01$)

$$P(k) \propto e_{ik} \cdot \sqrt{f_D(\text{dist}(i, k))} \quad k \in K = Nei(i) \tag{2}$$

$$P(j) \propto e_{kj} \cdot \sqrt{f_D(\text{dist}(k, j))} \quad j \in J = Nei(k) \cap \text{OPEN}(V, l) - i - K \tag{3}$$

If the set J is empty we go a step backward and we select a different node k . If, for each k chosen, it is not possible to define a non-empty set J , the procedure fails and the algorithm recovers selecting j using the bridging. Bridging is also used in case node i has not neighbours, i.e. the set K is empty.

The function of the distance we use in Equations 2 and 3 is defined as the square root of the function f_D . This definition guarantees that the geographical distance between connected nodes in the final network follows the power-law rule defined in f_D . In Figure 4 we show a comparison between a given function f_D and the geographical distances obtained using this algorithm.

After the selection of node j , a new social link r between nodes i and j is created (line 15). Its emotional closeness $r.e$ is extracted from the density function f_E in the same manner as in the single-ego model (line 16). Then, we update the network adding the new social relationship r to the set of links E . We also update the egos i and j , in terms of the ego network sizes ($i.size$ and $j.size$ respectively) and of the residual budget of time ($i.dbg$ and $j.dbg$ respectively) (line 17). It is worth noting that this update can determine the transition of a node from the open to the closed state, with respect to the current layer l .

For each layer l , we generate and add new social links until there are open nodes available. When the set of the open nodes is empty, the procedure switches to the next layer until all the three layers are completed.

Function f_D . According to the results presented in 4 and summarised in section 2 the probability of contact between two users at a certain distance follows a power-law of the form $P(d) \propto d^{-\alpha}$. In order to obtain a related probability density function f_D we have to introduce a thresholds d_{min} from which the power-law holds. Moreover it has to be defined for the range of values of d , which is the interval $(0, 1)$. The function, shown in Figure 4, is thus defined as:

$$f_D(d) \propto \begin{cases} d_{min}^{-\alpha} & \text{for } 0 < d < d_{min} \\ d^{-\alpha} & \text{for } d_{min} < d < 1 \end{cases} \tag{4}$$

Experimental results in [4] suggest that $\alpha = 1.5$. On the contrary, a value for d_{min} cannot be set in general since it strongly depends on the geographical space we consider and on the geographical distribution of the sampled population. Note that, given the number n of nodes in the network, since they are equally distributed in the space, $n \cdot d_{min}$ is the average number of nodes within the distance d_{min} from any given position. Thus, given a node in the network, the closest $n \cdot d_{min}$ nodes (on average) have the same highest-probability to be selected as destination of a social link. This parameter impacts on the clustering coefficient of the network, as we highlight in section 4.2.

4 Model Validation and Properties of Generated Graphs

In this section we validate our model comparing the synthetic social networks it generates with a real social network. In Section 4.1 we describe the real social network we consider for the validation. In Section 4.2 we compare the results with the properties of the reference network and we highlight how key properties of the generated networks depend on the model parameters.

4.1 Reference Network

The reference network we use for the validation of our model is obtained from a large data set crawled from a Facebook regional network on April 2008³. As we discuss in [19], the analysis of this data set, opportunely processed, shows that it shares similar properties with respect to those observed in other types of human social networks, and thus it can be used as a representative network to validate our model. Note that the network resulting from this data set is of a much larger scale with respect to the ones typically analysed in the anthropology literature. It contains more than 23 million social links (Facebook friendships), involving more than 3 million users. For each social link, the data set provides the number of social interactions occurred between the users. A social interaction can be either a wall post or a photo comment. The complete analysis of this data set is available in [19]. Hereafter, we summarise the key outcomes of this analysis that are then used to validate our model.

From the original data set, some users have been dismissed since they were not considered relevant, either for having too few interactions, or because they had joined Facebook just before the beginning of the data collection period. As discussed in [19] both cases can lead to biased representations of ego networks. The new data set obtained from the selection of relevant egos and the social links between them contains 90,925 users and 1,264,658 social links.

As described in [19], it is possible to extract from the data set the frequency of interaction between users. Since there are evidences of a strong correlation between the interaction frequency and the strength of the social tie [8], we can

³ This data set is publicly available for research at

<http://current.cs.ucsb.edu/facebook/>, referred as “Anonymous regional network A”.

Table 2. Structural properties of the reference and generated networks

| | reference network | $p = 0.8$ $d_{min} = \frac{250}{n}$ | $p = 0.8$ $d_{min} = \frac{500}{n}$ | $p = 0.8$ $d_{min} = \frac{1,000}{n}$ | $p = 0.5$ $d_{min} = \frac{500}{n}$ |
|------------------------|----------------------|--|--|--|--|
| mean degree | 27.82 | 133.91 | 133.94 | 134.00 | 133.86 |
| avg. shortest path | 4.06 | 3.40 | 3.26 | 3.11 | 3.12 |
| clustering coefficient | .109 | .152 | .108 | .085 | .079 |
| Jaccard (global) | .038 [.001] | .060 [.001] | .040 [.001] | .030 [.001] | .030 [.000] |
| Jaccard (support cl.) | .069 [.001] | .084 [.001] | .071 [.001] | .064 [.001] | .042 [.001] |
| Jaccard (symp. gr.) | .056 [.001] | .073 [.001] | .059 [.001] | .053 [.001] | .036 [.000] |
| Jaccard (affinity gr.) | .042 [.001] | - | - | - | - |
| Jaccard (active net.) | .031 [.001] | .059 [.001] | .037 [.000] | .025 [.000] | .030 [.000] |

consider these frequencies define the hierarchical structure of ego networks. Authors in [19] show that 4 clusters, corresponding to the typical layers of ego networks highlighted by Dunbar [12], can be identified also in Facebook ego networks.

Relevant properties of the reference network are reported in the second column of Table 2. The high clustering coefficient (with respect to random networks) and the short average path length prove that the reference network is “small-world”. Analysing the properties summarised in the table we have to take into account that, for technical reasons (e.g. the discard of not relevant nodes), the data set captures just a random sub-sample of the social links on the crawled Facebook networks and some of the indexes are influenced by the sampling, i.e. the average degree and the average path length. If we had the complete network, we would most likely find a higher average degree and a shorter path length. On the contrary, the clustering coefficient of a network preserves its value independently of the considered random sub-sample [20].

We use the Jaccard coefficient to estimate the similarity of the neighbourhoods of two adjacent nodes, that is to say the ego networks of two socially tied individuals. This is a very important index, as it describes the correlation between different ego networks. Capturing this aspect is one of the key goals of our model. The Jaccard coefficient for two sets A and B is defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ and it is also not biased by random sub sampling⁴. Since computing the Jaccard coefficient between the end-points of each social link in the network requires huge computational efforts, we estimate its average value considering the pairs of end-points of a sample of 10,000 edges randomly extracted from the network. The estimated average Jaccard coefficient (global) is reported in Table 2 (computed with 95% confidence level). According to this result, considering two socially connected individuals, their common acquaintances are, on average, 4% of the union of their acquaintances. Intuitively, individuals connected by strong ties

⁴ This can be easily seen observing that random sampling proportionally affects both the union and the intersection sets.

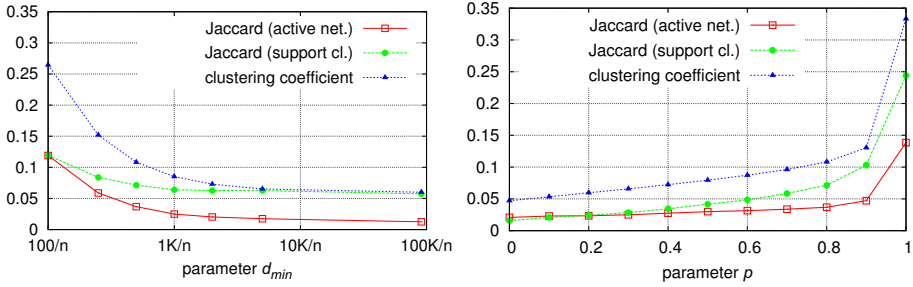


Fig. 5. Clustering coefficient and Jaccard indexes for different (a) d_{min} (with $p = .8$) and (b) p (with $d_{min} = 500/n$). 95% confidence intervals < 0.001 .

should have a higher ego network similarity than individuals connected by weak ties. In order to verify this intuition, we sampled 10,000 edges for each layer of the ego networks (external part) and computed the Jaccard coefficient between the ego networks of the nodes at the endpoints of the links. As expected, results, reported in Table 2, confirm that the similarity is higher for inner layers and lower for outer layers. Specifically, it drops from about 7% for the support clique to about 3% for the active network.

4.2 Results

The majority of the parameters for the model described in Section 3 are directly inferred from the social-anthropological literature as discussed in Section 2. The only parameters we can set in order to conduct experiments are: (i) the number of nodes in the network n ; (ii) the probability of selecting the “triadic closure” strategy, and (iii) the minimum distance d_{min} for f_D . In our experiments we choose to set $n = 90, 925$, which is the number of nodes in the reference network, while we use different values for the parameters p and d_{min} . The main properties of the generated network are reported in Table 2. Note that generated networks do not consider the presence of the “affinity group” layer (see Section 3.1) which we can assume to be merged with the “active network” layer.

The values of the parameters that allow us to best match the properties of the reference networks are $p = .8$ and $d_{min} = 500/n$ (fourth column of the table). These values mean that 80% of the social relationships are established through the triadic closure mechanism, rather than creating a bridge, and that, given a node, the 500 closest nodes (on average) have the same highest-probability to be selected as link’s destination. Results show a strikingly similarity of the social structures between the reference network and the graph generated through the model. Indeed, both networks have the same clustering coefficient and similar Jaccard indexes for the different ego network layers. Note that discrepancies in the mean degrees and in the average shortest path length are due to the sub-sampling of the reference network. Remember that, as shown in 2, apart from these results for the global network, the use of the single-ego model (see Section 3.1) guarantees that well-known ego network properties are also satisfied. They are the size distribution of the network and of the single layers, the

correlation between the layer dimensions and the distribution of the emotional closeness level.

In Table 2 we report the properties of the networks obtained with $d_{min} = 250/n$ (third column of the table) and $d_{min} = 1,000/n$ (fifth column of the table), maintaining $p = .8$. Moreover, Figure 5 (a) shows the clustering coefficient and the Jaccard index computed between pairs of strongly-tied egos (i.e. belonging to each other support clique) and weekly-tied egos (belonging to each other active network). Results show that reducing d_{min} the clustering coefficient and the similarity indexes increase for all layers of the network. Intuitively, this is because with smaller d_{min} the set of nodes selected with highest probability by an ego (those at a maximum distance of d_{min}) is smaller, and geographically very close to the ego. This leads to higher clustering (and similarity).

Similarly to the geographical constraints, also the variation of the parameter p influences the structure of the network. As shown in the last column of the table and in Figure 5 (b), if we diminish the value of p , the clustering coefficient and the similarity indexes decrease. This is expected as the number of links established through the bridging increases, and the bridging mechanism alone leads to the generation of random networks without clusters of socially connected nodes. Note in particular that when $p = 0$ (corresponding to a network without triadic closures) the Jaccard indices in Figure 5 (b) are the same, as in a network without triadic closures the correlation between social links do not depend on the strength of the links anymore.

5 Conclusions

In this work we define a new model for the generation of social network graphs, significantly extending the ego network model presented in [2]. We introduce different strategies to combine ego networks in order to form complete social network graphs, based on well-known properties in the field of social networks analysis i.e. (i) the "triadic closure", (ii) the presence of bridges and (iii) the geographical constraints.

In order to validate our model, we tune the model parameters obtaining a graph with the same structural properties of a real large scale human network obtained from Facebook. Then, we analyse the effect of key parameters on the properties of the generated graphs, highlighting the impact of both geographical constraints and social constraints.

The results presented in the paper confirm that our model leads to the generation of network graphs socially consistent. This model can thus be used for analysing through large scale simulation key properties of human social networks and for the development and the validation of protocols for socially-centric platforms.

Acknowledgments. This work was partially funded by the European Commission under the SCAMPI (FP7-FIRE 258414), RECOGNITION (FP7 FET-AWARENESS 257756), and EINS (FP7-FIRE 288021) projects.

References

1. Conti, M., et al.: Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber-physical convergence. *Pervasive and Mobile Computing* 8(1), 2–21 (2012)
2. Conti, M., Passarella, A., Pezzoni, F.: A model for the generation of social network graphs. In: 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–6 (June 2011)
3. Granovetter, M.: The strength of weak ties. *American Journal of Sociology*, 1360–1380 (1973)
4. Onnela, J.P., Arbesman, S., González, M.C., Barabási, A.L., Christakis, N.A.: Geographic constraints on social network groups. *PLoS ONE* 6(4), e16939 (2011)
5. Roberts, S.G., et al.: Exploring variation in active network size: Constraints and ego characteristics. *Social Networks* 31(2), 138–146 (2009)
6. Zhou, W.X., Sornette, D., Hill, R.A., Dunbar, R.I.M.: Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences* 272(1561), 439–444 (2005)
7. Dunbar, R.I.M.: The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews* 6(5), 178–190 (1998)
8. Hill, R.A., Dunbar, R.I.M.: Social network size in humans. *Human Nature* 14(1), 53–72 (2003)
9. Roberts, S., Dunbar, R.: Communication in social networks: Effects of kinship, network size, and emotional closeness. *Personal Relationships* 18(3), 439–452 (2011)
10. Dunbar, R.I.M., Spoor, M.: Social networks, support cliques, and kinship. *Human Nature* 6(3), 273–290 (1995)
11. Travers, J., Milgram, S.: An Experimental Study of the Small World Problem. *Sociometry* 32(4), 425 (1969)
12. Sutcliffe, A., Dunbar, R., Binder, J., Arrow, H.: Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology* 103(2), 149–168 (2012)
13. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* 393(6684), 440–442 (1998)
14. Ugander, J., Karrer, B., Backstrom, L., Marlow, C.: The Anatomy of the Facebook Social Graph. *CoRR* abs/1111.4 (2011)
15. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC 2007*, vol. 40(6), p. 29 (2007)
16. Wilson, et al.: User interactions in social networks and their implications. In: *Proceedings of the 4th ACM European conference on Computer systems, EuroSys 2009*, pp. 205–218. ACM, New York (2009)
17. Gonçalves, B., Perra, N., Vespignani, A.: Validation of dunbar’s number in twitter conversations. *CoRR* abs/1105.5170 (2011)
18. Teutle, A.: Twitter: Network properties analysis. In: 2010 20th International Conference on Electronics, Communications and Computer (CONIELECOMP), pp. 180–186 (February 2010)
19. Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F.: Analysis of ego network structure in online social networks. Technical Report IIT-CNR TR-10-2012 (2012), <http://www.iit.cnr.it/sites/default/files/TR-10-2012.pdf>
20. Lee, S.H., Kim, P.J., Jeong, H.: Statistical properties of sampled networks (2009)

C4PS - Helping Facebookers Manage Their Privacy Settings

Thomas Paul, Martin Stopczynski, Daniel Puscher,
Melanie Volkamer, and Thorsten Strufe

CASED, Technische Universität Darmstadt

Abstract. The ever increasing popularity of Online Social Networks has left a wealth of personal data on the web, accessible for broad and automatic retrieval. Protection from undesired recipients and harvesting by crawlers is implemented by access control, manually configured by the user in his privacy settings. Privacy unfriendly default settings and the user unfriendly privacy setting interfaces cause an unnoticed over-sharing. We propose *C4PS - Colors for Privacy Settings*, a concept for future privacy setting interfaces. We developed a mockup for privacy settings in Facebook as a proof of concept, applying color coding for different privacy visibilities, providing easy access to the privacy settings, and generally following common, well known practices. We evaluated this mockup in a lab study and show in the results that the new approach increases the usability significantly. Based on the results we provide a Firefox plug-in implementing C4PS for the new Facebook interface.

1 Introduction

Over 850 million users allegedly share personal information, private photos, videos, opinions and discussions on Facebook. The shared personal information include their age, gender, sexual preferences, taste and hobbies. All this data stored in Facebook or any other Online Social Network (OSN) can be linked to the relating individual by their real names published in their profiles.

Access to all this information is controlled by the OSN service provider, based on the user's privacy settings. Several studies have shown that despite increasing awareness [17], users due to the intricacy of the task are incapable of configuring their intended settings, and indeed do not understand their activities' implications [23]. However, the fact that Facebook and other OSNs have modified the default privacy settings to be more and more open with each update, makes it very important that users can easily grasp and change their privacy settings.

Consequences of this situation span unintended over sharing, and more serious threats, arising as scraping and harvesting [21,25], automated social engineering [5,6], social phishing [15] as well as various further attacks. In face of this perilous incomprehensibility, [17,10] go as far as proposing to abandon access control entirely and applying usage control and data ownership instead. However, this approach is not feasible with current technology, and the reasoning is in stark contrast to several other studies [3,22].

Previous research concordantly argues that privacy enhancing technologies, including distributed and secure data storage are important for OSN. Yet, it can only improve the situation if the users are actually able to properly configure their privacy settings. Furthermore, there is consent that this can only be ensured by increasing intelligibility of current privacy controls.

To this end we propose *C4PS - Colors for Privacy Settings*, a novel concept for privacy settings and their representation. *C4PS* aims at minimizing the cognitive overhead of the authorization task, based on three foundations:

- Color coding of authorization settings with immediate feedback upon change,
- one-click configuration based on proximity of data and respective controls,
- group-based access control through aggregated configuration, and easy group management based on drag-and-drop.

While we implemented and tested *C4PS* as a proof of concept for Facebook, the idea is generally applicable to any OSN, or other web pages with privacy settings. We started with a *C4PS* mockup for the Facebook interface early 2011 to evaluate, if *C4PS* indeed simplifies the authorization task and performed a lab user study. The results

- indicate that modifying and inspecting the privacy settings is significantly easier and more efficient when applying *C4PS* and
- confirm previous studies showing that even users who consider themselves proficient with the Facebook site are unable to correctly perform precise privacy settings.

Based on the results of the study we provide a Firefox plug-in applying *C4PS* to the modified Facebook interface after the introduction of the *Timeline* for download.

The rest of this paper is organized as follows: Putting *C4PS* into perspective, we give an overview of related work in Section 2. We present the rationale concept and design of *C4PS* in Section 3. The methodology of our user study is described in Section 4 and its results in Section 5. We conclude the paper with a summary and future work in Section 6.

2 Related Work

Improving privacy in OSNs is a very widely discussed issue in current literature [19,18]. One research area covers the confidentiality concerns towards OSN providers as one single entity that needs to be trusted. Approaches to resolve this vulnerability include several proposals to apply encryption and/or decentralized storage of user data. The range starts with cutting the profile in centralized OSNs into atomic parts, encrypting each part separately and distributing keys to authorized recipients only [14]. It ends with completely distributed peer to peer (p2p) OSNs like PeerSoN [9], DESCENT [16] or Safebook [11]. These approaches help to assure users' privacy needs with technical support by architectural means

or applying crypto, assuming that users are aware of the consequences of publishing personal data, as well as able and willing to commit themselves in subject of privacy. These approaches still require the data owners to grant access to authorized users to selected data.

Several studies and experiences have shown that the ability to understand and modify privacy settings is generally missing [1,7,3,2]. One class of proposals attempts to decrease the frequency of explicit acts of authorization by applying methods from machine learning to pre-configure the overall settings [13]. To “detect and report unintended information loss” [4] supports users, too. Explicit authorization however is still needed to train the recommender and to fine tune the settings.

Further approaches have tried to make it easier for users to manually grasp their current privacy settings. [12] use an interface, based on Venn diagrams. But they don’t meet our design Principle 2 to use well known pattern and don’t help users in managing their groups. [22] present a privacy setting interface which helps users of Facebook to understand the effect of their changes by providing an audience view. Users are presented their own profile in the way that a single potential other recipient would see it. The limitation of this approach is that users are not efficiently able to figure out the visibility of profile items to whole groups of friends, nor does it aid the users in granting authorization. Mazzia et al. addressed this limitation in [24] by creating the “PViz Comprehension Tool” which is able to illustrate privacy settings by color (from light to dark), “based on the user’s privacy selection for a selected profile item”. These improvements alleviate to build and verify the user’s mental model of the interface by showing the effects of the conducted adjustments.

Our approach in contrast leads to a new mental model in terms of OSN access control. It is based on color coding, which is well known from other areas of the user’s environment. Based on daily experience, users understand the effects of their adjustments at our privacy setting interface with a minimum amount of effort. Combined with single click changes we seriously reduced the obstacle of configuring access control rule sets.

3 C4PS - Improved Interface

To improve the usability of privacy settings, we developed a corresponding *C4PS* - overlay for Facebook.

3.1 Design Principles

The concept of *C4PS* is based on four main principles. The first three cover usability aspects according to ISO 9241, and the last one the applicability of the interface.

P1 - Little Effort: To ensure high accuracy when working with the interface, the user shall be able to check or change his privacy setting with as little effort (easy and fast) as possible (inspired by ISO 9241-11 – effectiveness and efficiency; and [20]).

P2 - Applying Common Practices: To minimize the learning effort while becoming accustomed to our interface, commonly accepted and well-known usability patterns shall be used to support users – like colors, drag and drop, tooltips or graying out inactive elements (inspired by ISO 9241-10 – conformity with user expectations).

P3 - Direct Success Control: To avoid gaps between intended and actually performed adjustments (as shown in [23]), results of modifications to the privacy settings shall be displayed and visible instantly (inspired by ISO 9241-10 – self descriptiveness).

P4 - Applicability: To cause the least possible cognitive overhead for accustomed users and to stay independent of Facebook, *C4PS* needs to allow for direct integration into the existing web pages.

Based on these four principles, we developed concepts for *C4PS*, identifying a need for new functionality for both the main privacy settings as well as the group management.

3.2 C4PS Privacy Settings

Regarding the main privacy setting functionality we highlight each attribute in the profile by a particular color, depending on the group of people who are granted access. We also enable the user to change the accessibility with just one click, support the group selection with tooltips, make this privacy settings mode easily accessible, and provide very brief instructions. In addition, the privacy settings mode provides a button to check how others see the profile. These concepts are explained in detail in this subsection.

Color Coding: The colors used are guided by the well-known traffic light colors (*P2*). Blue was added to represent custom settings. The corresponding color definition is:

- *Red:* Visible to nobody
- *Blue:* Visible to selected friends
- *Yellow:* Visible to all friends
- *Green:* Visible to everyone

All privacy settings are visualized by our color scheme in the *C4PS* privacy setting mode (*P4*), so that an attribute’s visibility can be directly derived from its coloring (cmp. Fig. [1]).

Easy To Modify Setting for Single Attributes: The user can change the privacy setting for a specific attribute by simply clicking the buttons on the edge of the row on the right side (*P1*). The color of the buttons shows the visibility that will be set for the entry by clicking on it (e.g. in Fig. [1]). The settings are changed immediately (*P3*), which is reflected directly by a color change of the attribute’s cell. If the user chooses “selected friends” (blue), a window opens in which friends or groups are granted access to the mentioned attribute.

Tooltip: To further increase the usability, tooltips indicate the setting corresponding to the color for each button ($P2$). Tooltips are shown when the mouse hovers over the button (cmp. Fig. 1).

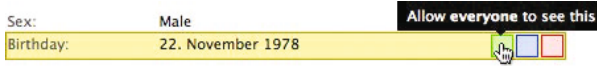


Fig. 1. Color coding for one attribute - birthday

Easy Access to Privacy Settings: $C4PS$ integrates in the mockup a new button under the profile picture to enter the $C4PS$ privacy settings page. This button is visible on each FB page and thus the $C4PS$ privacy settings page is easy to access ($P1$). After switching to the privacy editing mode and editing the privacy settings, the user can exit this mode by clicking a button labeled “Stop editing privacy settings” at the same place. In the improved version we enabled the visibility of color coding instantly without entering any privacy settings mode.

Information on Top of the Page: According to common practice ($P2$), general information about the color visualization and the meaning of each color are provided on top of the page in the editing mode.

Checking How Others See Their Own Profile: The privacy settings mode provides a button at the top of the page ‘How others see your profile’, which offers a simple visualization to check how selected other people - including friends - see the profile ($P1$).

Application to Photo Albums: The privacy settings for photo albums can be checked and modified with the same color mechanism. When visiting the Facebook “photos” tab, an overview of all photo albums of the user is displayed, as in the original Facebook interface. However, there is an additional button labeled “Edit Privacy Settings” (cmp. Fig. 2).

This button again activates the $C4PS$ privacy editing mode. Here, the photo album elements are highlighted with a color indicating the privacy setting (cmp. Fig. 3). Additionally, three colored buttons are shown on every item and allow to change the privacy setting as described before. Clicking on the colored buttons changes the privacy setting for the entire album, while individual restrictions, set to single photos, remain unchanged. To change the privacy settings of a single photo the user can open the photo album, in which the colored privacy buttons are placed under each photo.

With $C4PS$, checking and modifying privacy settings in Facebook takes a minimum of two steps:

1. Accessing the $C4PS$ privacy settings main page by clicking on “Edit Privacy Settings” (no longer required in the improved version).

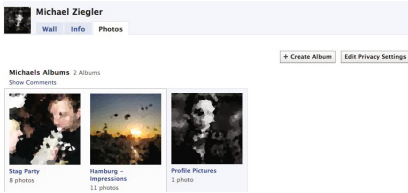


Fig. 2. Photo albums without privacy settings



Fig. 3. C4PS interface - photo albums

- 2a To inspect the current settings for the profile entry, the user only needs to properly interpret the color. In case of custom settings a third step is required.
- 2b To change the setting of any attribute, the user can simply click on the button colored accordingly.

4 User Study

To evaluate *C4PS*, we conducted an extensive, controlled lab study. We aimed at validating the following four hypotheses:

- H1.** *C4PS* makes it easier and faster to find out to whom a particular attribute is visible.
- H2.** Using *C4PS*, testing how the complete profile is presented to another user is easier and faster.
- H3.** Setting the visibility of attributes is easier and faster using *C4PS*.
- H4.** The group management can be handled easier and faster using *C4PS*.

These four are intended to cover all aspects that may concern users aiming to adjust their privacy settings. In addition, we were interested in the feedback about the concrete ideas implemented in *C4PS* to further improve it.

We decided to run a lab study because this enabled us to measure time and clicks while the participants solved some tasks with both interfaces - the improved one and the original one. Correspondingly, the participants were asked to use a lab PC and a Facebook profile we created, to set a controlled environment and without warring the user to expose his own profile.

4.1 Course of Action

The study contained the following phases:

All tasks had to be solved in this particular order while it was not required to start from the main page after login. This course of action is more realistic, as users usually want to check or edit the privacy setting for more than a single attribute.

| Nr. | Action |
|-----|--|
| 1. | OSN questionnaire [7] (on paper) containing eleven questions regarding the use of OSN in order to estimate the prior knowledge of the test person. |
| 2. | First practical part, during which several tasks have to be solved with one of the interfaces. Note, to prevent a possible learning effect due to the first use of one of the two interfaces, the order of presentation of the two interfaces was alternated for each test person. The answers were written down (on paper). |
| 3. | “System Usability Scale” (SUS) questionnaire (on paper) as introduced by Brooke [8]. It allows measurements concerning effectiveness, efficiency and user satisfaction, and due to its generality is applicable to various types of systems. |
| 4. | Second practical part |
| 5. | SUS questionnaire was applied to the second interface. |
| 6. | Usability questionnaire (on paper) containing 15 questions regarding the usability of the new interface and a field for general comments. |
| 7. | Demographic questions (on paper) concerning age, gender, and profession. |

| Nr. | In the practical part of the study, we asked the test persons to: |
|-----|---|
| 1. | Find out to which users or groups the birthday (Task 1) / hometown (Task 2) / relationship status (Task 3) / a particular photo album was visible (Task 4) |
| 2. | Find out which attributes were visible for a specific friend (Task 5) |
| 3. | Create a group “best friends” (Task 6) |
| 4. | Add two particular friends and the group “class mates” to the group “best friends” (Task 7) |
| 5. | Adjust the privacy settings of five attributes - mobile phone number to only two specific friends (Task 8.1) / interests to all (Task 8.2) / hometown to only one specific group (Task 8.3) / relationship to no one (Task 8.4) / religious and political views to all friends (Task 8.5) |
| 6. | Adjust the privacy settings of one selected photo album, granting access to a specific group, except a single particular friend, being part of the group (Task 9). |

4.2 Evaluation Criteria

The following information was deduced from the screencast:

- *Time*: Time a test person needs to perform a task
- *Hits*: Number of clicks a user needs to complete a task
- *Precision*: The task-solving precision of a study participant. It is only distinguished between the values 1 (task solved completely and correctly) and 0 (failure to precisely solve the task).

The measurement of time and clicks for a task was performed manually. The first goal-directed mouse movement was taken as starting point for the measurement of a task. The end of the measurement was chosen to be the successful or failed completion of a task, or the user canceling the task. We used the time frame without mouse movement before a new task was started as an indicator for canceling. We did not count clicks incidentally placed beyond any button or link as well as multiple clicks on a button or link to start a function (while waiting

for the website to respond). This should preserve the comparability of values. All other clicks to perform a task were counted. This includes clicks on scroll bars, selecting text or clicking into input forms. The time and clicks between tasks was stripped.

To evaluate our hypotheses, we measure both the time and clicks it takes to solve a task to evaluate if a system is *easier and faster*, and we consider the precision of a solution as its success. The usability questions from the SUS questionnaire, Attrakdiff(tm) questionnaire, and our final own usability questionnaire additionally are taken into account to gauge intelligibility and acceptance of *C4PS*.

4.3 Sample Description

Recruiting was done in lectures and via email lists. The information provided to the participants was that a new interface for the privacy settings in Facebook would be tested. Participants were rewarded with sweets.

The study was performed with 40 students, aged between 20 to 32 years. All were members of at least one OSN, except for three participants. 57,5% access their OSN profile(s) at least once a day and 25% even several times a day. Nearly two thirds of the test persons are Facebook users. Almost all study participants (90%) have already been in touch with the privacy settings of their OSN provider. However, many of them consider these settings to be confusing (57,5%). 15% of the participants were very concerned about their privacy settings and stated that they modify or check them every month. The rest did it less often. 30% did not change the privacy settings, after they have been set up once. The possibility to create lists or groups of friends, was only used by 25% of the participants and the possibility to set certain rights for groups or for individual friends was used by 37,5%. 62,5% of the participants stated that they are aware of the visibility of their profile's attributes to other network members.

5 Results

We first provide the results of the study regarding success rate and efficiency (Subsection 5.1). Afterwards, we discuss the feedback regarding the three usability questionnaires (Subsection 5.2 and 5.3). We show that the four hypotheses can all be confirmed in each category according to the evaluation criteria defined in Subsection 4.2. Based on these results we provide some ideas for further improvements.

5.1 Success Rates and Efficiency Analysis

In this subsection we show that the four hypotheses hold regarding the success rates, the time needed, and the number of clicks needed to complete the corresponding tasks.

Success Rates. The overall success rate for all tasks and all participants in the new interface is 91% while it is only 68% for the original Facebook interface. As shown in Figure 4 the success rate for the new interface is higher than the one for the original interface in almost all tasks. Only for task 3, the original Facebook interface performed better. Here, subjects were asked to list the friends or groups who have access to the attribute “Relationship Status”. Unfortunately the participants wrote down the privacy setting “selected friends” while we expected them to read out the actual list of friends who have access. In most cases the participant did not click on the blue button in order to get this information but only wrote down the tooltip text (selected friends) that was revealed when hovering over the button. Some other participants did not write down all groups having access to this attribute or the wrong ones. According to our definition both cases were interpreted as wrong answers.

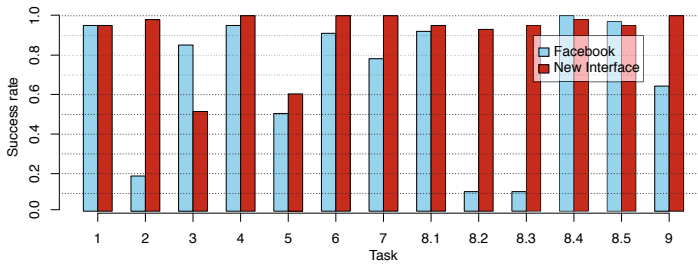


Fig. 4. Success rate per task

The biggest difference was measured at task 2 (visibility of the field “current city and hometown”). Only 17,5% of the participants solved this task correctly with the original interface, while all but one participant succeeded using the new interface. One reason for this is that this attribute is placed on Facebook in the slightly hidden “Connecting on Facebook”-section and not on the main privacy settings page. In addition, many participants wrote down the value of the incorrect attribute “Contact information”, which was displayed on the main privacy settings page on Facebook. The difference between both interfaces again is very large for Task 8.2 and 8.3, for a similar reason, and the participants hence changed the wrong attribute. For task 8.3, participants changed the field “Contact information” instead of “hometown” while for task 8.3 the incorrect attribute “Interested in” was changed, instead of “Interests”. The latter in this case represents the gender the user is interested in rather than the intended interest in his activities like sport, films, music or other.

Efficiency Analysis. The efficiency analysis with respect to time and clicks below compares only tasks 5 to 9, since in these tasks the participants actually had to change settings, rather than interpreting the current configuration.

Therefore, for these first four tasks it is not clear from the videos when a participant completed a task and started the next.

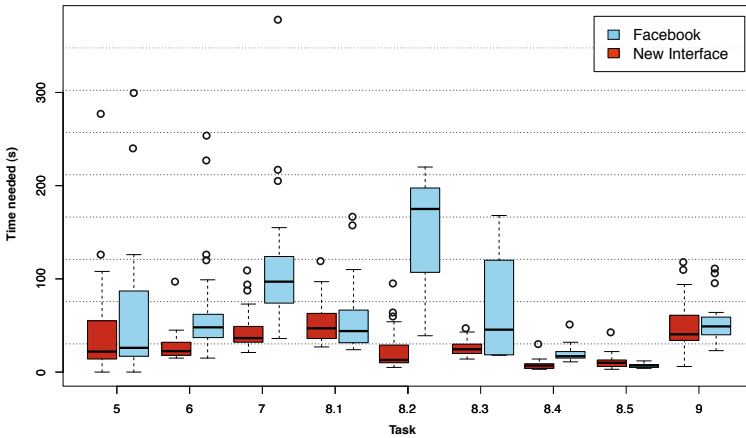


Fig. 5. Required time (per task)

The minimum number of clicks to properly execute all four Tasks (1-4) using *C4PS* is one click on “Edit Privacy Settings” from the main page, then interpreting the privacy settings for the first two requested attributes. In the case of task 3, a further click was required, as the displayed privacy level “selected friends” was not the proper answer, but it was necessary to interpret which selected friends were granted access by clicking on the blue button. Thus, one click was necessary to open the dialog, and another one to close it. Similarly, it was required to click on the photo album settings to discover this information. The minimum number of clicks in *C4PS* thus amounted to 4. The minimum number of clicks to execute these tasks properly in Facebook amounted to 8.

Time needed: Most tasks were completed faster when using *C4PS*, as shown in Fig. 5. Especially when adjusting privacy settings that are in the “Connecting on Facebook”-category and while creating groups. The test users on average need more than twice as much time to solve the tasks using the Facebook interface, as compared to *C4PS*. Fig. 5 also shows that the variance using *C4PS* is much lower for most tasks, indicating that all users achieved approximately the same efficiency.

Clicks needed: Considering the number of clicks (Fig. 6), the results are very similar to those from the time measurement. Most tasks can be solved with much fewer clicks using *C4PS*, and the variance is very low. The participants generally needed nearly three times more clicks to complete the task using the original interface. Note, that it can be assumed that a much greater deviation would have been achieved, if all privacy setting tasks had to be performed separately starting from the main menu. Using the Facebook interface, the user would have needed to perform at least three additional clicks to get to the settings menu, compared to a single click that is necessary using *C4PS*.

Comparing users with and without Facebook Accounts. The test persons who already use Facebook had an advantage when solving the tasks, because they already

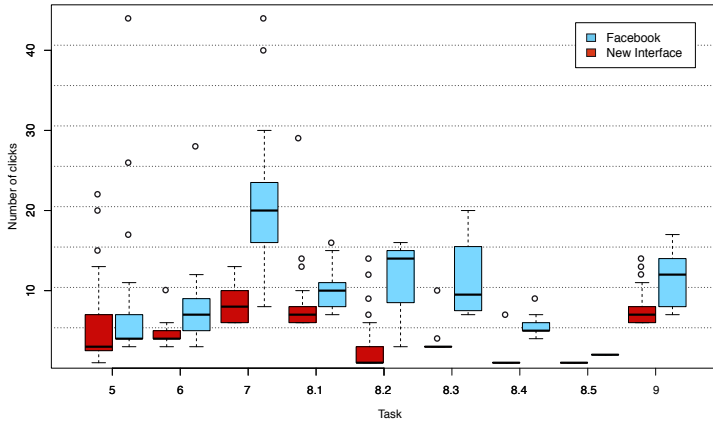


Fig. 6. Required number of clicks (per task)

knew the look and feel of the Facebook interface, or even the concerning privacy settings. However, even these participants achieved better success rates with *C4PS*, even if they could be considered Facebook experts for using it every day. In numbers, the success rate of Facebook experts for the tasks on Facebook was 73% compared to a success rate of 94% when using *C4PS*. Subjects who were not considered Facebook experts only reached a success rate of 60% for the task when using the original interface, rising to a success rate of 86% when using *C4PS*.

Almost all tasks have been solved better by participants that are Facebook users (in both interfaces). Solving the task on Facebook, the experts needed 1.65 less clicks on average. When using *C4PS*, the disparity between experts and normal users was smaller. The experts in this case completed the task with 0.89 less clicks. Measuring the time for completing tasks, the experts performed 1.76 times faster using Facebook. Using *C4PS*, however, the experts were only 0.75 times faster. This disparity shows an additional improvement of the usability of the systems, and the subjects who had not used Facebook before had a much harder time to cope with the original interface at all.

The results for all three criteria show that even users who consider themselves proficient with Facebook are unable to correctly perform precise and efficient privacy settings.

5.2 SUS - System Usability Scale

In this subsection we show that *C4PS* performs better regarding SUS.

The average System Usability Scale (SUS) [8] value for our interface (all users) has been evaluated to 82.6. The maximum possible SUS value of 100 was achieved at maximum, and the worst rating of the interface was valued at 37.5. Comparing this with Facebook, the users rated the interface with an average SUS value of 35. The maximum value was 75 and the minimum was 5. Referring to A. Bangor et al. who analyzed the results of 2324 studies with SUS in the last ten years [8],

acceptable products have a SUS-score of over 70. Better products start at the high 70s and end in the upper 80s range. Only truly excellent products have a score above 90. Products with scores less than 50 should be cause for significant concern and are judged to be unacceptable. Due to this scale, the usage of our interface is very good while Facebook itself reaches numbers below those for acceptable products.

5.3 Concept Evaluation

At the end of the study, we asked the study participants what they like and do not like as well as what they would improve. The results of this questionnaire are discussed in this Subsection. They show that *C4PS* also performs better regarding these interface specific usability questions, and that people like the general concepts.

57,5% of the participants rated the original Facebook privacy setting mechanisms as confusing (the worst level on a scale of 4 possibilities) and only one stated that it is very clearly arranged (the best level). 87,5% of the participants stated that *C4PS* improved the situation a lot (maximum improvement of a scale of 4 options). On a scale with 4 options 50% rated the visualization with colors as very good, 47,5% with good and the rest with level 3 while no one selected level four. The question whether the color coding is well-defined was agreed by 31 (77,5%) of the participants.

Only 20% of the participants answered that they cope ‘very well’ or ‘well’ with the original interfaces for group management while 97,5% of the participants made this statement for the new interface. The question regarding the usability of the privacy setting mechanisms was answered with ‘very good’ by 5% of the participants for the original Facebook interfaces and by 47,5% for the new interfaces while 22,5% (FB) and 50% (new interface) stated that these mechanisms in the corresponding interfaces provide a ‘good’ usability.

There were also two fields to provide comments. In the first one we asked the participants what they liked most about the interface. Almost everyone mentioned the colors while only a few also mentioned the group management. People stated for instance that the privacy settings are ‘easy’, ‘clearly arranged’, ‘directly accessible’, ‘easy to find’, ‘easy to use’, ‘everything is on one page’, ‘less clicks’, ‘quick’, ‘applicable for more attributes’ and ‘clear what to do’. In the second field we asked them to propose further improvements. Comments mainly addressed the group management and the profile preview in general and for the case that particular friends have the right to access this attribute. Some remarks were made regarding the colors - including only three colors, changing colors, self-defined colors; and also the fact that the order of the colored buttons in a row should stay the same.

6 Conclusion and Future Work

This paper deals with access authorization in Online Social Networks, and the specific case of Facebook. Even though users publish highly personal data on

such sites, several studies have shown that they are incapable of configuring their privacy settings correctly. The direct consequence is unwanted over-sharing of highly personal information by the users, which allows for various attacks, including information harvesting and various types of social engineering.

To increase the intelligibility of the authorization controls, we have proposed, evaluated, and implemented *C4PS* – *Colors for Privacy Settings*. *C4PS* introduces a new mental model for the privacy settings, and has been designed as simple and intuitive as possible, to minimize the cognitive overhead of the authorization task. It is based on the foundations of color coding, simple, one-click configuration, and group-based access control, including a simplified group management interface. We initially implemented *C4PS* as a mockup for controlled lab studies.

Evaluating *C4PS* in an extensive, controlled user study demonstrated two main insights:

1. *C4PS* greatly aids the authorization steps – it not only enables the user to grant exactly the desired authorization, but additionally helps the user comprehend their authorization activities and current settings.
2. Even users that are convinced of their expertise using Facebook are unable to employ the existing privacy controls correctly and efficiently, and are unable to precisely configure their profile according to the desired authorization.

Both interface and privacy configuration of Facebook have changed during the course of this study. The presentation of the profiles has changed entirely, and following Google+, the privacy settings have been made seemingly simpler to use. The service increasingly encourages to organize the friends in groups, to facilitate the authorization step. The interface additionally introduced an icon to hide items from the timeline. This control is not implemented for posts to other users' walls, though, and no enhancements have been made to help comprehending current settings, and the consequences of applied authorization changes. We hence adapted *C4PS* to Facebook Timeline and implemented it in a Firefox plugin, which is available for download from our web site.

C4PS being a concept of general applicability, we are aiming at applying it to other social networking services, like for instance Google+, Twitter and Foursquare, in future work. We are additionally aiming at analyzing the applicability of the main principles of *C4PS* to present other complex settings, configurations, and further properties of online services, thus making them easier to grasp and to handle.

References

1. Acquisti, A., Gross, R.: Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In: Danezis, G., Golle, P. (eds.) PET 2006. LNCS, vol. 4258, pp. 36–58. Springer, Heidelberg (2006)
2. Antón, A.I., Earp, J.B., Young, J.D.: How Internet Users' Privacy Concerns Have Evolved since 2002. IEEE Security & Privacy Magazine 8(1), 21–27 (2010)

3. Balfanz, D., et al.: In Search of Usable Security. *IEEE Security & Privacy* (2004)
4. Becker, J., Chen, H.: Measuring privacy risk in online social networks
5. Bilge, L., Strufe, T., Balzarotti, D., Kirda, E.: All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In: *WWW* (2009)
6. Boshmar, Y., et al.: The Socialbot Network: When Bots Socialize for Fame and Money. In: *ACSAC* (2011)
7. Boyd, D., Hargittai, E.: Facebook privacy settings: Who cares? *First Monday* (Online)
8. Brooke, J.: SUS - A quick and dirty usability scale. *Usability evaluation in industry* (1996)
9. Buchegger, S., et al.: PeerSoN: P2P Social Networking - Early Experiences and Insights. In: *SNS* (2009)
10. Castelluccia, C., Kafaar, D.: Owner-Centric Networking: Toward a Data Pollution-Free Internet. In: *SAINT* (2010)
11. Cutillo, L.-A., Molva, R., Strufe, T.: Safebook: a privacy preserving online social network leveraging on real-life trust. *IEEE Communications Magazine* (2009)
12. Egelman, S., Oates, A., Krishnamurthi, S.: Oops, i did it again: mitigating repeated access control errors on facebook. In: *CHI 2011* (2011)
13. Fang, L., Kim, H., LeFevre, K., Tami, A.: A Privacy Recommendation Wizard for Users of Social Networking Sites. In: *CCS* (2010)
14. Guha, S., Tang, K., Francis, P.: NOYB: Privacy in Online Social Networks. In: *WOSP* (2008)
15. Jagatic, T.N., Johnson, N.A., Jakobsson, M., Menczer, F.: Social Phishing. *Commun. ACM* (2007)
16. Jahid, S., Nilizadeh, S., Mittal, P., Borisov, N., Kapadia, A.: DECENT: A Decentralized Architecture for Enforcing Privacy in Online Social Networks
17. Kagal, L., Abelson, H.: Access Control is an Inadequate Framework for Privacy Protection. In: *W3C Privacy* (2010)
18. King, J., Lampinen, A., Smolen, A.: Privacy: Is There An App for That? In: *Symposium on Usable Privacy and Security, SOUPS* (2011)
19. Krishnamurthy, B., Naryshkin, K.: Privacy leakage vs. Protection measures: the growing disconnect. In: *W2SP* (May 2011)
20. Krug, S.: *Don't Make Me Think: A Common Sense Approach to the Web*, 2nd edn. New Riders Publishing (2005)
21. Lindamood, J., et al.: Inferring Private Information Using Social Network Data. In: *WWW* (2009)
22. Lipford, H.R., Besmer, A., Watson, J.: Understanding Privacy Settings in Facebook with an Audience View. In: *UPSEC* (2008)
23. Madejski, M., Johnson, M., Bellovin, S.: The Failure of Online Social Network Privacy Settings. Tech. rep., Columbia University (2011)
24. Mazzia, A., LeFevre, K., Adar, E.: The pviz comprehension tool for social network privacy settings. Tech. rep., University of Michigan (2011)
25. Strufe, T.: Profile Popularity in a Business-oriented Online Social Network. In: *EuroSys/SNS* (2010)

Dynamic “Participative Rules” in Serious Games, New Ways for Evaluation?

Jean-Pierre Cahier, Nour El Mawas, and Aurélien Béné

ICD/ Tech-Cico University of Technology of Troyes (UTT) Troyes, France
{cahier, nour.el_mawas, aurelien.benel}@utt.fr

Abstract. Rules are used by Computer Games to evaluate losses, gains, changing items and actions of the players. In addition, they reinforce realism and playability, especially in training situations where Knowledge is complex and expert (e.g. best practices acquisition in crisis management, decision making in complex socio-technical systems...). To evaluate items and actions, we propose a dynamic solution using “participative rules”. In this approach, based on Computer Supported Cooperative Work and Knowledge Engineering, the rules base is directly generated from a special discussion forum which contains successive versions of the textual rules continuously discussed and co-built by the designers’ community, in strong relation with the players’ community. This paper resumes a “Work in progress” recently presented with more details [1] to the Game Community, but it extends it by adding the point that, beyond the “Serious Games” field, the notion of “participative rule” that we are exploring, could interest more broadly Human and Social Scientists who seek new ways towards effective evaluation methods.

Keywords: evaluation, participative rules, serious game, participative design.

1 Introduction

Both quantitative and qualitative rules are classically used by “Serious Games” to evaluate items and actions of the players. They reinforce realism and playability, especially in complex and expert training situations (best practices acquisition in crisis management or in complex ecosystems...).

In one hand, these rules are a particular kind of knowledge needed by players to represent what they can do in the domain, what is the qualitative or quantitative value of their actions or of the used items, what are the stakes and priorities, and globally how activity makes sense and is valued in the game. These rules can be expressed by text or by more formal means (shared vocabulary, semantic categories, attribute-value peers, logical proposition, etc.). In classical games, rules and rules justifications are not well elicited or known by the player. They are not related to the experts’ discussions, at a detailed granularity. That represents problems we want to remediate.

In another hand, these rules, translated to semi-formal or formal computerized forms in a rules base, are processed by the game engine to compute, register, track, organize the progress of the player in the computerized game.

In almost computer games, including the Serious Games, the classical design process of a computer game disjoins the design step (of a given version of the game) and the gaming step. Players, considered as non-specialists, are not involved in the design. But in the case of serious games aiming expert knowledge in very complex domains, rules and heuristics are very numerous, changing and often controversial. Players (learning expert practices) should be more active: they can have, or invent during the playing, interesting knowledge useful to the game and to the game design.

For all these reasons, we propose to gather in a single loop the rationale, the design and the use of a rule, and to explore for games user-centered design practices using “participative rules”. This approach is based on Computer Supported Cooperative Work and Knowledge Engineering, and refers to “Open Source” Communities’ established principles, by considering that designers’ and gamers’ communities as mature ones. As shown on Figures 1, 5 and 6, the repository containing the game rationale, the game items and the game rules is organized on the basis on a discussion forum, to be permanently discussed and co-built by the designers’ community. The design is also made in strong association with the players’ community which has a large view and possibility of comment on detailed rules and their design rationale.

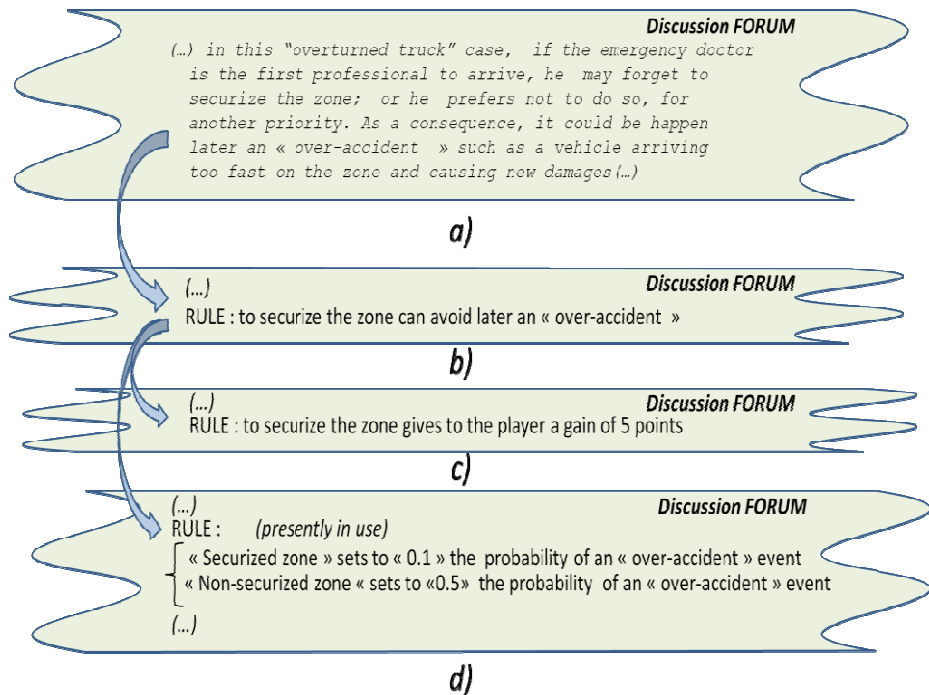


Fig. 1. From inquiry to participative rules, by using the designers’ forum : a) Verbatim of the case told by the expert. b) Induced rule (general knowledge). c) Pedagogic static participative rule. d) Pedagogic dynamic participative rule. (Example coming from the “Aidcrisisis” project).

The paper is organized as follows. The §2 proposes a quick overview on the notion of serious game, especially for expertise training in complex, changing and controversial domains. In §3 we justify some reasons for using participative rules, by examining present collective practices in the domain of the multiplayer games, and presenting the case of the serious game (of the multiplayer “adventure game” type) we are developing in answer to real world needs for Crisis Management [2]. In §4, we demonstrate in more details how participative rules work. Then in §5 we precise what are, at the social, cognitive and technical levels, some underlying proposed principles, models and architecture our solution requires. Finally in §6 we sketch possible applications of participative rules for searchers in other fields.

2 Serious Games for Expertise Training

The idea that the serious games facilitate the learning is confirmed in numerous domains like languages and health, economy and management [3]. The educational games can offer many learning benefits such as motivation, engagement and fun [4]. These “serious games” would rather be named “useful games”, as suggests it [5], because if they contribute to serious learning in complex domains knowledge, they haven't to be, however, serious or boring.

But on the contrary, no need to add “funny” dimensions. It is enough so that there is “game”, useful or not, that deploys all the tension of the associated personal investment. The player is not the spectator of a show. He/she is interactive, not a “user” or a consumer of an application conveyed by an Computer-Human Interface, “he does not behave towards the game as towards an object” [6]. What defines the player, and what defines the playful character, is that he/she is inside the game and takes seriously the purposes of his/her game, whatever is this game. The game is not thus characterized by a factor of joke, whim or surprise, which it would be necessary to strengthen by subtleties, but, thus according to the philosopher H.G. Gadamer, the game is characterized by the implication of the player subjectivity *“which forgets himself (...) in and out of the game movement. (...) The player is in a world which is determined by the seriousness of his purposes”* (ibid). Subjectively, this definition of the playful character seems completely to characterize the serious games intended for the professionals, as for the emergency doctors training with possible operations in case of disaster (see below): the professional mission (in this particular case, to limit the number of victims) is amply the enough element for making exist the playful dimension and, the priority objective of designers will be to analyze in closer the real situations lived by the professionals, to identify and model the playful elements.

The situations being arrested by actors through their activities and their knowledge, asks the question of knowledge. By considering the serious games as characterized by a learning aspect (memorization, personalization) and a playful aspect (motivation, interaction), the objective is to integrate the contents of learning into the playful aspect. [7] differentiates two ways to design the serious games by speaking about “extrinsic metaphor” - where the playful aspect is an overlayer without relationship with the didactic contents - and about “intrinsic metaphor” - where the learning is in the

heart of the playability. Because the game is a subjective implication, we have to consider the *participation* as a complementary and sometimes essential vector of the learning: “*I learn because I participate, because I make a commitment in an activity which offers me information elements, knowledge transformation or practices, as well by being conscious or no of educational effects of the process*” [8].

3 Why Participative Rules?

The transformation of knowledge, considered as "knowledge for the action" [9] includes the contribution of knowledge through the player. The knowledge through the player brings to the collective his "added value" to know, to know-how, to comment or to initiative and to facilitate contribution when the participative and knowledge-intensive game, as we propose it, allows to collect and to capitalize these contributions. The games allow developing at player's the "skills of the XXIth century" such as innovation, critical and systematic thought, and teamwork, to have knowledge producers and not only consumers [10]. It is about the informal strategic learning, between the formal learning and the informal learning of daily life [11].

In our works, we want in particular to evaluate the improvement of players learning when the game contains a discussion forum for players, we can lean for example on [12] and [13], which already deduced that the conversations in the forum help making the players masters of their learning and favor the passage from the transmissive model of knowledge towards the collaborative model of players communities.

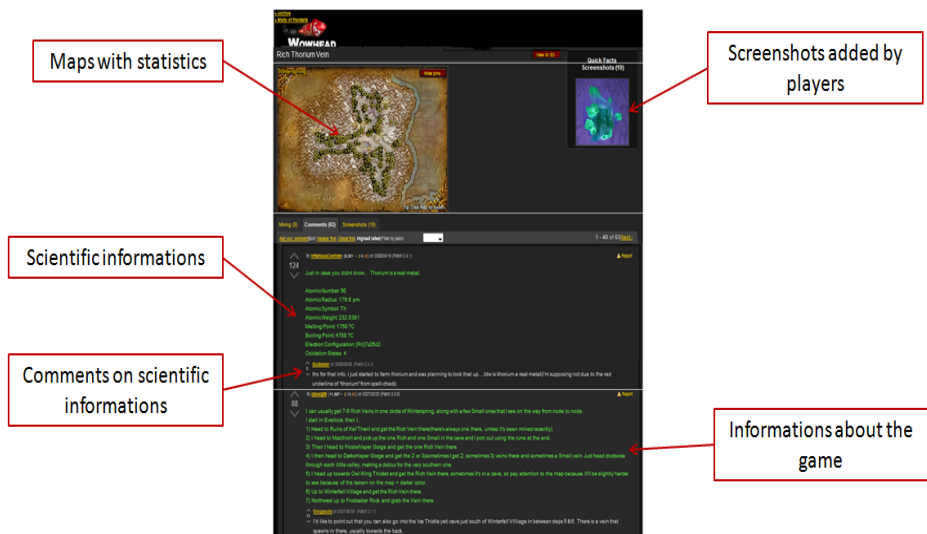


Fig. 2. The WOWHEAD Forum page dedicated to the “thorium” topic (for more details, see: <http://www.wowhead.com/object=175404>)

Players' discussion forums are today used on a very interesting manner, by the MMORPGs Players' Communities. MMORPGs (Massively Multiplayer Online Role-Playing Game Play) are a type of video game that allows many players to simultaneously interact in a virtual persistent world. Figure 2 shows a part of a page of the "WOWHEAD forum", created by World of Warcraft (WOW) players. This page is dedicated to the "Thorium" topic. The complete page contains hundreds of knowledge items, comments and discussions, and could be viewed as typical participative knowledge item, very near of what we want to do with "participative rules".

WOW is not a "serious game", nor a "useful game". Knowledge on "Thorium" in WOWHEAD does not represent valid scientific Knowledge. But, players' behavior is as serious as in a Serious Game! (As we said in §2, "*what defines the player is that he is inside the game and takes seriously the purposes of his game, whatever is this game*", so ironically we could say that WOW is a "serious game"). In this example, like in Wikipedia, crucial knowledge to operate the game, to make it more playful and to win, is co-constructed by the player's Community. Players complete for example the geographical map with statistics about the regions concerned by a given crucial resource (the "Thorium"). Note that WOW is a commercial product (Blizzard), but that the WOWHEAD forum is auto-organized by the Players community to exchange Knowledge independently of the WOW society, and is sometimes in conflict with official WOW knowledge (e.g. WOW do not diffuse statistics on "Thorium" localization).

Discussion of rules is necessary to develop Serious Games in high expertise fields. In the paper we take as an example a Serious Game development in the "Aidcrisis" project [2]. Started in 2011, the Aidcrisis project is led by the University of Technology of Troyes in association with the Emergency Service (SAMU) of Troyes Hospital (Fig.3). It concerns the management of NRBCE¹ crisis situations on the territory of the Aube department (France). One of the modules of this project concerns e-training of the emergency staffs (doctors, ambulance drivers, nurses etc.) and joins in the approach described in this paper. In the Aidcrisis project, the experiences of accident treatment cases or disasters are the object of a collection and an analysis leading, through the architecture we propose, in the specification of scenes facilitating the training of emergency staffs in the decisions and in the operations in the crisis cases.

In Aidcrisis, numerous factors plead for the use of the approach of participative and knowledge-intensive serious game: the cases of possible disasters although unlikely (but requiring for the emergency staffs to get ready for it) are very numerous, especially if we consider the conjunction of unforeseen factors or crossed causes of NRBCE / natural disasters.

The real exercises are very expensive and difficult to organize. And even if we want to proceed to virtual exercises with a classic approach of serious game, because of the current cost of the development of the game sequences, we could treat only some tens or hundreds of scenarios, what would not allow facing real stakes in the

¹ NRBCE: Nuclear Radioactive Biological Chemical Explosive.

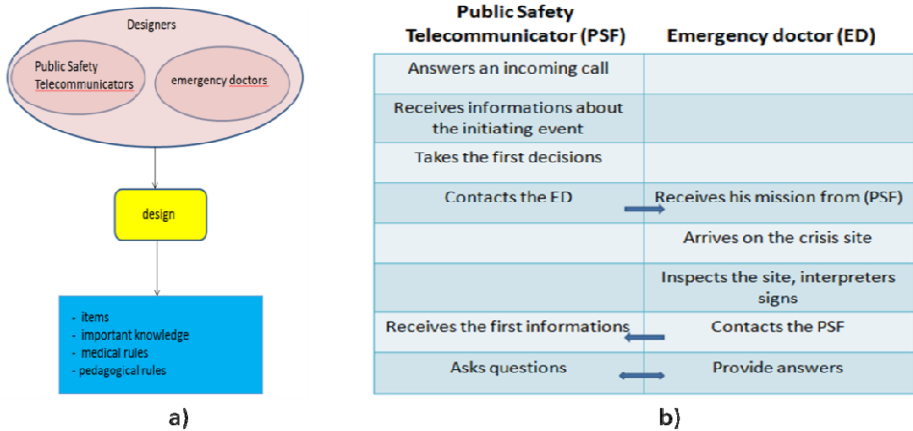


Fig. 3. A serious game for the Aidcrisis project a) Actors of the design, b) Example of scenario

preparation for real risk. The answer is to co-design the detailed specifications with and by the persons in charge of training at the emergency medical service, to benefit from the proposed interaction systems (sharing of specifications, designers’ and gamers’ forum), to give the system, master to persons in charge of the emergency trainings so they can develop easily any new scenario. Our objective is to prepare these professionals to be autonomous in the imagination of new cases and in the production of the new associated scenes (or variants of the existing scenes) based on their fine knowledge of the domain and their educational priorities.

4 How Do Participative Rules Work?

The notion of co-design in a participative approach goes back to the 80s projects, related to reflection on the democracy, in domains like the repair of locomotives or the publishing world in Scandinavian world [14] or the design of information and cooperation systems [15]. These authors underlined the necessity of including very early all actors concerned in the design. In our context, it means considering player as in the center of design: it is less a question of designing for player, that to design with him.

In agreement with [9] and “Social Semantic Web” Approaches [16][17] the co-design must be also accompanied by the construction of semantic structures by actors such as "maps" of their knowledge in connection with their practices. It allows actors themselves to map the shared items and to organize their space of cooperation even "to appear" this organization in continuous process. Globally the chosen approach allows supporting at the same time the design, the evaluation, but also the game itself like the players actions by notifying them about the changes which arise locally and globally in the game.

The possibility of ambitious “participation architectures” for the virtual universes, was consolidated by the success of applications like Wikipedia or more generally

applications of Web 2.0 [18]. But the realization of similar architectures for serious games still raises numerous problems, because it is necessary at the level of infrastructure to take into account the large number of players, to introduce a certain flexibility to take into account contributions of the multiple actors (players and designers). Furthermore, the actors have to cross their skills in situations for which knowledge and data are very numerous and strongly evolutionary, that is the case in games, where scenes and their items are numerous.

Figure 4 illustrates the infrastructure and the participative aspects at the level of both communities: co-designers and players. At the infrastructure level, Knowledge Organization System records the game items, the exchanged messages by actors, etc. The authorized actors can look, add, catalog, discuss, comment items. Intermediate plan is the one of the game. From the game, players can look and comment items.

For designers a permanent discussion forum must be able to organize on any object of the game in particular on every rule connected to an object, in particular when the game is used, because at this moment the differences of interpretation, opinions and approaches are the easiest to express and to discuss. The game has to base, on one hand, on a database cataloging and returning easily accessible and editable game items (including rules, designed as editable contents elements) and on the other hand the forum.

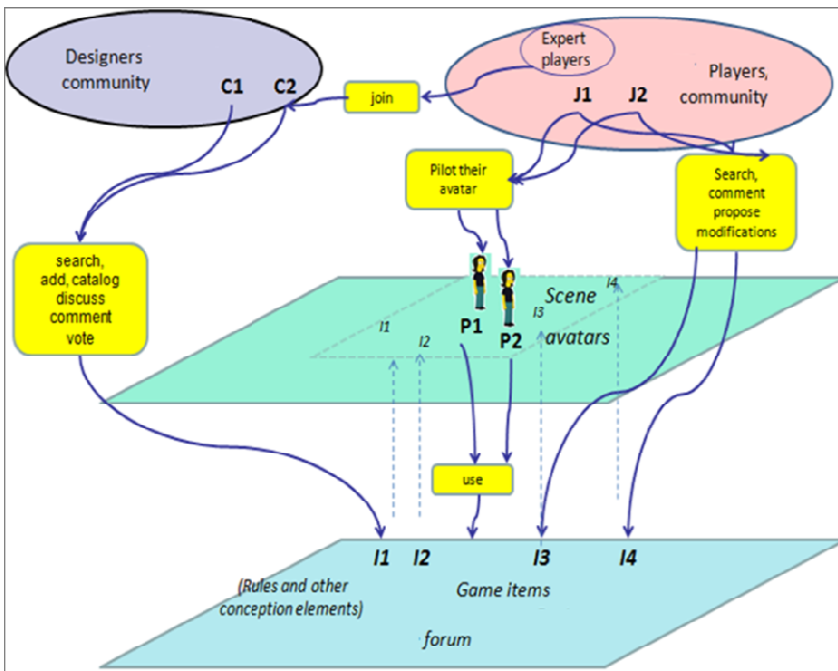


Fig. 4. Actors participation *i*) in game (*middle part*) and *ii*) in the Knowledge Organization System associated to the game, involving the discussion forum (*lower part*)

If we consider a game scene, this one is going to be constituted at first by knowledge elements and rules specified by trainers and designers, for example rules governing the penalty of an action on an object in terms of "points". Then, the scene is going to be played and the rules instantiated. The designers have to specify on one hand rules and items of the game, and on the other hand rules of the educational evaluation included in this game (in particular values such as the number of won or lost points, which are visible to the player, facilitating in particular his motivation in game, his auto training, etc.).

The architecture of the software platform proposed to the designers (see Fig.7a) has to allow editing these diverse specifications, finding easily knowledge, discussing them item by item, reaching the moderate values of attributes, etc. From their part, the players also can look for items and use them to treat assimilate and comment them - for example to confront their experiences, exchange hints and tips. It will be interesting that the players can reach (if they want to do it) the design rationale and the designers' discussions on the forum.

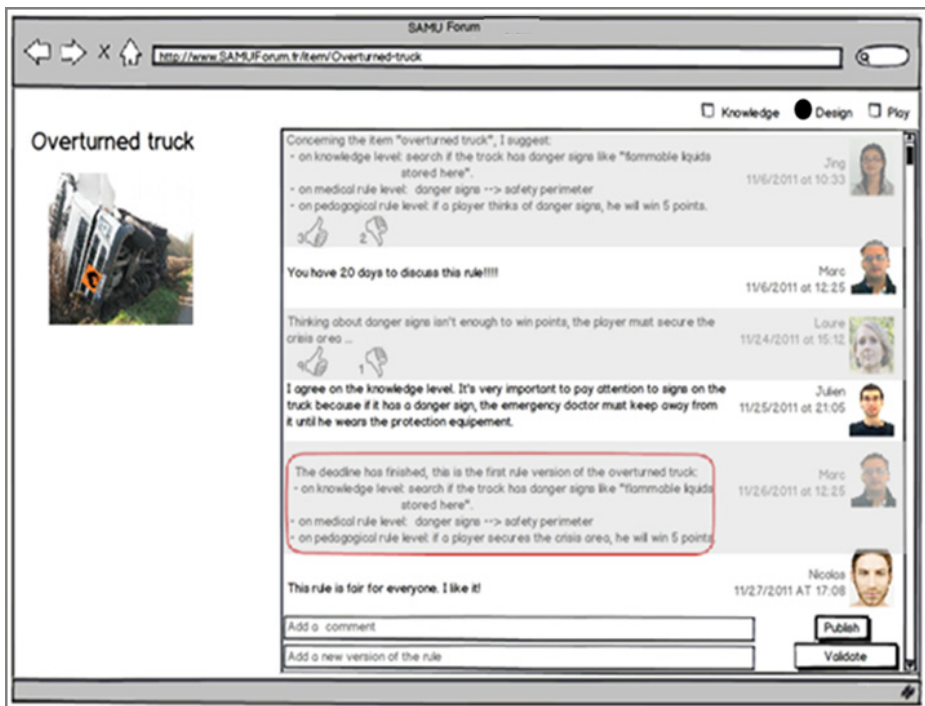


Fig. 5. Usage of the forum by designers for a participative rule (circled) concerning an “Overturned truck”, on the same case than in Figure 1 (*mock-up*)

Via the forum, the players can discuss, between them and with the designers, in an asynchronous way, on the rules of an action, the won and the loss of points, know and criticize the reason of this rule, etc. In the forum a mark allows to spot if messages are posted by a player or by a designer; the designers can, if they wish it, mask some of

their exchanges to the players. The proposed architecture is going to offer to the designers a working system which articulates:

- A specification system directed to a teamwork susceptible to associate skills resulting from several disciplines (jobs of expertise field, trainers of the field, pedagogy specialists, graphic designers and scriptwriters),
- A navigation system in the game objects (this point is particularly crucial in the knowledge-intensive serious games, which contain numerous objects and rules),
- A discussion forum type (see mock-up below, Fig.5 and 6).

In the complex domains (sustainability, crisis management) actors think and act locally according to rules which can depend on places, on seasons or on other factors. That is why for a designer who builds objects and rules of a scene, it is important to have design forum for the discussion between peers. For example if the community adopts the principle that a rule must not have "veto" of any other designer, all designers will be invited to join the " design forum " to discuss new rules and find the necessary compromise for their implementation in the game.

The figure 5, for example, shows the rule discussion corresponding to an action rule concerning an "Overturned truck". Does the player think to install a security perimeter? The initial designer suggests for this action a static rule (winning five points). When the rule "is released ", the discussion which it caused is available in the game scene: we suggest to the player the he can investigate rules attached to scene objects. The player is encouraged to mobilize the rule and his discussion thread as resource "to play better".

Additionally to the rules, the players can discover illustrated knowledge as well as related experts debates. Around the rule, they are invited to exchange "hints and tips". Complementary to the use of statistics, these "narratives" give designers more qualitative and richer returns. Finally, players can suggest improvement ideas for the game, introducing them, little by little, into the universe of designers and experts.



Fig. 6. Usage of the forum by players for the rule concerning an "Overturned truck" (*mock-up*)

A designer appreciating a player proposal for changing a rule cannot alone modify this rule, because changing a rule depends on the discussion between the designers groups participating to the design forum of this rule.

To encourage players to contribute and to improve the game, they have to feel that their proposals are examined and lead to improvements in the game. That is why, regularly, the group of designers should discuss new players’ proposals to decide if they reject, adopt or postpone the proposed modification of the rule.

5 Underlying Architecture and Model

The architecture of the software platform proposed to the designers (see Fig.7) has to allow editing these diverse specifications, finding easily knowledge, discussing them item by item, reaching the moderate values of attributes, etc. From their part, the players also can look for items and use them to treat assimilate and comment them - for example to confront their experiences, exchange hints and tips. It will be interesting that the players can reach certain parts of the designers' forum.

Our technical architecture depends on the particular status held by game rules in our approach. To be edited by non IT-specialists, these rules must be managed as data and not as programs anymore.

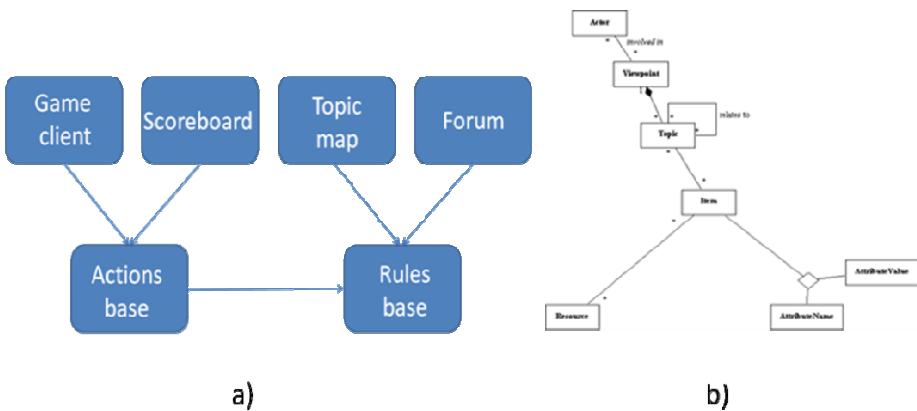


Fig.7. The proposed Architecture. a) Dependences between components b) Hypertopic model [19] including Viewpoints.

In addition, they are a reference point for the topic map, the forum, and the score calculation. That is why they need to be a share service. In a more classic way, another service allows the management of players' actions. This service doesn't allow only the mutual awareness between players but also allows to designers accessing to a scoreboard for the use traces analysis.

Item and rules are organized in the Knowledge Organization System following the Hypertopic model [19] which is conflict-tolerant [20] and allows building topic maps helping to make sense for users [16] using a few basic concepts: Item, Attribute, Point of view, Topic, Resource - and a few types of associations between them.

The Hypertopic model includes viewpoints and make easier participative editing of the rules and of other items in the virtual space. Points of view allow actors to qualify items by heuristic attributes ("topics") considered as linguistic terms (tags, topics, keywords...) and not as concepts. For example for a given item, Hypertopic attributes can be used to organize the knowledge, if there are different opinions about the value of an attribute (price of an item, number of winning points associated to a rule, the probability of happening of an event. Actors can be practitioners from various disciplines diverging on what are the important items, or on how to characterize them by topics, or on how to rely the topic, etc. According to Hypertopic an item is not a concept, just a node in the graph. It has a name (considered as an element of language) and, may be, topics, other attributes and links (to resources) put by actors.

6 Possible Extrapolations beyond Serious Games

We would now suggest that the "participative rules" concept, that we study and want to validate in the Serious Game field, prefigures a new kind of artifact of "intellectual technology" [21] useful in a "collective intelligence" perspective [22] for actors in the real world.

"Participative rules" could interest a large set of Scientists in Economics, Sociology, Management Sciences, Pedagogy, Ecology, etc., for theoretical reflection, experiment and prototyping. From a Peircean semiotic point of view, events in a situation are signs which can receive several types of normative interpretations: in the example Fig.1, the static rule prescribes only a *final interpretant* (*value as a result*), while the value in the dynamic rule depends more pragmatically of the potential consequences in the network of possible future events (*value as a set of potentialities*). In addition, the forum proposes another kind of *dynamic interpretant*: a collective "living" interpretation dimension, for both static and dynamic rules. May be, after a forum discussion, the value could be still confirmed to "5 points", but the significance of "5 points", could have nevertheless changed for the group.

Evaluation by the means of rules is an object for many researchers, especially in Human and Social Sciences. How do standards happen? What kind of pragmatic rule does structure the human activity? "A good rule", "a good principle of evaluation" could be considered as boundary objects [23] [24] on frontiers between various knowledge domains. These questions interest especially cross-boarded approaches seeking innovative ways for evaluating complex knowledge [25] [26], economic behaviors or human activity in complex socio-technical systems or ecosystems [27], etc.

For example, in future possible projects, various scientists in the fields of Sustainability could have ideas to use a "participative rules" architecture. They could explore dynamic evaluation solutions applied to real-world eco-systemic services. Rules could be "monetary" (e.g. by using "points") or "non-monetary", e.g. by using words, proximity in topic maps [28]...). But especially, because they are "living" and permanently discussed in an existing Community, *participative rules* should allow escaping the classic monetary/non-monetary dilemma and exploring new hypotheses (e.g. to exploit sociological approaches of the value [29] or illustrating approaches in the "economics of conventions" school [30] [31]).

It is to note here that an important prerequisite is the existence (or the progressive emergence encouraged by the artifact) of a Community of interacting participants, “designing when using” (or “when playing”) the service. In the Community the participants would be interested to continuously discuss, exchange viewpoints, design and apply the dynamic set of common rules. In such a service-oriented participative design, serious game could be a complementary means to explore ideas, to build and test theoretical models, to mock-up solutions and to train actors. Thus, beyond their benchmark in the Serious Game field, participative rules could be a means:

- To make emerge and evolve models coming from a permanent exercise of “collective intelligence” by a Community, by keeping track of controversial arguments and accepting conflict within the design rationale. To give a ruling about a theory or a model is never definitive. If a participant is a scientist, an expert, or simply an end-user more knowledgeable on a particular topic (cf. players becoming specialists on the “thorium”, see Fig.2), he/she can always add his/her stone to the model.
- To put emphasis on rhetoric, textual and semi-formal expression of rules, facilitating their expression and their understanding by actors. The discussion forum is composed by textual material with a high granularity and semi-formal structuring (icons, marks, tags, smileys...). Participative rules give an opportunity to readjust the balance between qualitative and quantitative approaches. An explicit quantification of rules is possible for certain models and can be useful for computer game deductions or automatic demonstrations. But in a participative rules context, a quantified rule in the engine is always indexed on an equivalent text in the forum, so that any quantified rule can be changed easily.
- To let the actors prototype and test, in a socially-controlled manner, the changing set of rules by gaming or by “open simulation” including “avatars” representing members of the Community. The validation criteria are decided by the group and can be continuously and endlessly changing. By that means, the Community can assimilate dynamic evaluation modes, test and control models proposed by scientists in a “collective intelligence” context.
- To train members to new models and evaluation modes accepted the Community important in their business or their every days life, in this context of permanent change of the Reality and of the Knowledge. Participative rules illustrate a emerging generation of Serious Games, useful for new “Web2.0” training systems where knowledge, viewpoints and priorities will be more easily updated.

7 Perspectives, Conclusion

We proposed in the paper a social method based on “participative rules”, that we intend to complete and to validate in the field on Serious Game. We described functional and technical solution elements, by indicating (on some examples involving changing and controversial expert Knowledge) why this solution would be the most suitable to these particular Serious Games.

In the near future we pursue a work plan for the computer implementation, to make a prototype which will allows us to validate gradually certain underlying hypotheses

of our proposal. The current stages include in particular the setting-up of the mock-ups of forum solutions presented in the present paper. We also develop the graphical game editor and the topic map allowing easily indexing and seeking items, rules and arguments in the debate and in the design rationale. We wish as soon as possible to realize an experiment implying a group of non-IT specialists' co-designers - within the framework of the mentioned Aidcrisis project - , so they can define scenes, create and modify them continuously, according to the proposed rapid prototyping and co-building method. We wish to prove that the Hypertopic model and the participation architecture underlying our platform, will favor structuring the objects and knowledge added in the game, so these are easier to find, to manipulate, to discuss in a wide community and to fluidly transfer in the game engine.

Thus “participative rules” concept could be at a short term implemented in serious game and tested in laboratory with designers’ and players’ communities, allowing to validate a part of the mentioned hypotheses. Concerning the extrapolation ways that we proposed in §6, a lot of additional field experiments with real communities would be to imagine and to realize, initiated with their additional hypotheses by scientific partners involved in Social Sciences. Such trans-disciplinary social experiments will be necessary to evaluate at which degree “participative rules” systems in real life stay utopian, or could be socially usable beyond the Serious Game field.

References

1. El Mawas, N., Cahier, J.-P., Bénel, A.: Serious games for expertise training. In: 17th Int. Computer Games Conf. CGame 2012, Saint-Louis, USA, July 30 (2012)
2. Matta, M., Lorientte, S., Cahier, J.-P., et al.: Representing experience on Road accident Management. In: IEEE 21st International WETICE Conference, 2nd CT2CM track (Collaborative Technology for Coordinating Crisis Management), Toulouse, France, June 25-27 (2012)
3. Blunt, R. (2009). Do serious games work? Results from three studies. eLearn. Magazine (December 1, 2009)
4. Ibrahim, R., Jaafar, A.: Using educational games in learning introductory programming: A pilot study on students’ perceptions. In: Conf. IADIS Game and Entertainment Technologies 2010 Freiburg, Germany, July 27 (2010)
5. Natkin, S., Dupire, J. (eds.): Entertainment Computing - ICEC 2009. LNCS, vol. 5709. Springer, Heidelberg (2009)
6. Gadamer, H.G.: Truth and method (1989); J. Weinsheimer & D. G. Marshal, Trans., 2nd rev. edn., New York, Continuum (original work published 1975)
7. Fabricatore, C.: Learning and Videogames: an Unexploited Synergy. In: 2000 AECT National Convention - A Recap. 2000 AECT National Convention, Long Beach, CA. Springer Science + Business Media, Secaucus (2000)
8. Brougère, G.: Using the Concept of Participation to Understand Intercultural Experience and Learning. In: International Seminar Research on Peace Education in Multilingual and Intercultural Contexts: the CISV Case, Modene University, Italy, March 27 (2009)
9. Zacklad, M.: Communities of Action: a Cognitive and Social Approach to the Design of CSCW Systems. In: Proc. of GROUP 2003, Sanibel Island, Florida, USA, pp. 190–197 (2003)
10. Gee, J.-P., Schaffer, D.W.: Looking where the light is bad: Video games and the future of assessment. Epistemic Group WP, no 2010-02, Univ. of Wisconsin-Madison (April 2010)

11. Protopsaltis, A., Pannese, L., Pappa, D., Hetzner, S., et al.: *Serious Games and Formal and Informal Learning*. eLearning Papers (25) (Juillet 2011)
12. Frasca, G.: *Videogames of the oppressed: Videogames as a means for critical thinking and debate*. PhD report, Georgia Institute of Technology (April 2001)
13. Riel, M.: *Cross-classroom collaboration in global learning circles*. In: Star, S. (ed.) *The Cultures of Computing*. Blackwell, Oxford (1995)
14. Ehn, P.: *Participatory Design and the Collective Designer*. In: Badham, R. (ed.) *Proceedings of Participatory Design 2002*, Malmö, June 23-25 (2002)
15. Winograd, T., Flores, F.: *Understanding Computers and Cognition*. Addison-Wesley, USA (1986)
16. Cahier, J.-P., Zaher, L.H., Leboeuf, J.-P., Pétard, X., Guittard, C.: *Experimentation of a socially constructed Topic Map*. In: 6th Ann. Conf. EurAM, Paper Session, 4 - Concepts and Practices of Organisational Learning, May 16-20, Oslo, Norway (2006)
17. Béné, A., Zhou, C., Cahier, J.-P.: *Beyond Web 2.0... And beyond the Semantic Web*. In: Randall, D., Salembier, P., et al. (eds.) *From CSCW to Web 2.0: European Developments in Collaborative Design*, pp. 155–171. Springer, London (2010)
18. O’Reilly, T.: *What is web 2.0: Design patterns and business models for the next generation of software*. Social Science Research Network (2005)
19. Zhou, C., Béné, A., Lejeune, C.: *Towards a standard protocol for community-driven organizations of knowledge*. In: *Proceedings of the Thirteenth International Conference on Concurrent Engineering. Frontiers in Artificial Intelligence and Applications*, vol. 143, pp. 438–449. IOS Press (2006)
20. Simmel, G.: *The Sociology of Conflict*. American Journal of Sociology (1903)
21. Goody, J.: *The Domestication of the Savage Mind*. Cambridge University Press (1977)
22. Levy, P.: *Collective Intelligence: Mankind’s Emerging World in Cyberspace*, 1st edn. La Découverte, Paris (1994); Translator, Bononno, R. (1999)
23. Star, S.L., Griesemer, J.: *Institutional ecology, ‘Translations’ and Boundary objects: amateurs and professionals on Berkeley’s museum of vertebrate zoologie*. *Social Studies of Science* 19(3), 387–420 (1989)
24. Vinck, D., Jeantet, A., Laureillard, P.: *Objects and Other Intermediaries in the Sociotechnical Process of Product Design: an exploratory approach*. In: Perrin, J., Vinck, D. (eds.) *The Role of Design in the Shaping of Technology*, COST A4 Social Sciences, Bruxelles, vol. 5, pp. 297–320. EC Directorate General Science R&D (1996)
25. Patton, M.Q.: *Qualitative evaluation and research methods*. Sage, London (1990)
26. Greene, J., Caracelli, V. (eds.): *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. New directions/br evaluation*, vol. 74. Jossey Bass, San Francisco (1997)
27. Maris, V., Béchet, A.: *From Adaptive Management to Adjustive Management: A Pragmatic Account of Biodiversity Values*. *Conservation Biology* 24, 966–973 (2010)
28. Cahier, J.-P., Zacklad, M.: *Towards a Knowledge-Based Marketplace model (KBM) for cooperation between agents*. In: *Proc. COOP 2002 Conference*, June 4-7, IOS Press, St. Raphael (2002)
29. Simmel, G.: *The Philosophy of Money*. In: Frisby, D. (ed.) (1907)
30. Boltanski, L., Thévenot, L.: *On Justification: Economies of Worth*. Princeton Studies in Cultural Sociology (2006); Transl., Porter, C. (1st publication: Gallimard 1991)
31. Orléan, A.: *L’empire de la valeur. Refonder l’économie*, Paris, Le Seuil, coll. La couleur des idées (2011)

Mobile Phones, Family and Personal Relationships: The Case of Indonesian Micro-entrepreneurs

Misita Anwar^{1,2} and Graeme Johanson¹

¹ Faculty of Information Technology, Monash University, Australia
{Misita.Anwar, Graeme.Johanson}@monash.edu

² State University of Makassar, Indonesia

Abstract. In the Indonesian context, the use of mobile phones has had many effects on the economic as well as the social fabric of communities. However, these impacts have not been thoroughly examined, particularly in relation to micro-enterprises, productivity or wellbeing. This paper evaluates the impact of mobiles from the perspective of human development where 'development' is seen as the expansion of people's choices and information and communications technologies (ICTs) are seen as supporting these choices. To test theories of development, it presents an empirical study undertaken in Indonesia about the impact of mobile phones on micro-entrepreneurs' wellbeing. Results show that micro-entrepreneurs regarded family is the most important aspect of their lives and that their own wellbeing was treated the same at that of their families. Accordingly, mobile phones are considered as a very significant force to maintain and improve their relationships with family, relatives and friends. Mobile phones contribute significantly to wellbeing.

Keywords: mobile phones, micro-entrepreneurs, human development, capability approach.

1 Introduction

Development has come to be conceived of and measured not only in economic terms, but also in terms of social wellbeing and political structures [14]. Such a perspective derives from a class of theories labelled "human development" or "people-centred development". Human development represents national development as "the enlargement of people's choices" and resonates with Amartya Sen's work on building capacities and entitlements. Sen [15] introduced a conceptual framework for evaluating social conditions by focussing on human wellbeing – it is called the Capability Approach (CA).

The human development perspective has given rise to a new strand of research studies where the use of ICTs is analyzed by using the lens of the many dimensions of human wellbeing, by using the CA as a framework (e.g., [1], [4], [18], [17], [11], [16], [8]). These studies suggest that ICTs have the potential to contribute to the many dimensions of human development.

This paper will analyse empirical field data to show how mobile phones have impacted on the lives and wellbeing of micro-entrepreneurs in Indonesia. The next section will outline the key concept of the CA, the existing research in this area, and how ICTs have been integrated into this framework. Then follows an example of how mobile phones potentially enable capability expansion. After presenting the research context, we present the preliminary findings of the study, where we analyse the relevance of mobile phones to micro-entrepreneurs' wellbeing, and how the micro-entrepreneurs describe the most important aspects of their lives.

2 The Capability Approach and ICTs

The CA is a framework that can be used for the evaluation and assessment of individual wellbeing, community development, social arrangements, or the design of practical policies and proposals about social change in society. Developed by Amartya Sen [15], and expanded later by other theorists and practitioners ([6], [12], [2], [13]), the CA critiques a welfare-based approach to evaluation. Sen [15] argued that in social evaluation and policy-making, it is essential to consider the "quality of life", which means including consideration of the freedom of people to live the life which they want, and which they find valuable. He believed that wellbeing and development should be analysed from the point of view of people's capabilities to function, and the opportunities and freedoms which they have "to be" and "to do" what they want to be and do.

The core ideas of the CA are *functionings* and *capabilities*. Functionings are described as the "beings and doings" of a person, whereas capability refers to a person's or group's "freedom to achieve" valuable functionings. Capability represents all potential actions that are available to a person from various combinations of functionings. Capability is a set of vectors of functionings (actions), demonstrating the person's freedom to lead one type of life or another [15]. In other words, the term functionings can refer to realised (actual) achievements and fulfilled expectations, whereas capabilities can refer to the effective possibilities of realising achievements and fulfilling expectations.

Questions that often arise as to which capabilities are the most important, and why? Sen has always refused to endorse any specific list of capabilities. He argues that to try to make one comprehensive list of capabilities should not be attempted, because these lists are used for different purposes, and each purpose might need its own list. Nussbaum [12] strongly advocates the outlining of central human capabilities and proposes a concrete list of basic capabilities, which is composed of the following ten categories: 1. Life; 2. Bodily health; 3. Bodily integrity; 4. Senses, imagination and thought; 5. Emotions; 6. Practical reason; 7. Affiliation; 8. Other species, that is, being able to live with concern for and in relation to animals, plants, and the world of nature; 9. Play; 10. Control over one's environment. Nussbaum however, does not claim that her list is definitive and unchanging and that states that it should be viewed not as a definition of good life but rather as the necessary conditions for a variety of lifestyles. A promising methodology for applying CA at the micro level has

developed by Alkire [2]. She addresses issues regarding the identification and pursuit of valuable dimensions of human development based in practical reason. Then, drawing on participatory tools and techniques she attempts to identify these valuable capabilities.

Having summarised the main concepts of the Capability Approach, we turn now to a discussion as to how ICTs can be integrated within this broad framework.

Although capability theory appears not to have been explicitly applied to technological domains before, a number of studies have referred to the relationship between ICTs and the CA (e.g., [1], [4], [18], [11]). Alampay [1] investigated how variables such as ICT ownership, age, gender, income and education, affect people's capability to use ICTs, and he also identified the barriers to access to ICTs. Giggler's [4] study concluded that ICTs can play an important role in enhancing the human capabilities of the poor (literacy-ICT skills) and that the extent to which the uses of ICTs expand peoples 'informational capabilities' was critical for determining the positive impact of ICTs on economic and social development. Johnstone argued for the need to enrich capability theory with knowledge capability; he asserted that ICTs at best are simply instruments of knowledge. Zheng [18] stressed that ICTs are relevant to the CA because of their special ability to shape the promotion of human rights, and the personal and social variables that affect people's capacity to achieve wellbeing. Finally, Kleine [11] operationalized the CA with what she called a "choice framework". Her framework asserts that a combination of individual agency and structural resources can be converted into capabilities. ICTs are an important part of structural resources which aid or constrain individual agency. They include rules, norms, culture, policies and dimensions of access (availability, affordability and skills needed for using ICTs). Additionally, human agencies can be categorised into 10 groups, including material resources, financial resources, geographical resources, health resources, human resources, educational resources, psychological resources, information, cultural resources, and social capital or social resources.

On the subject of capability expansion through the use of ICTs, several studies are notable, viz., [8], [16], [17]. Hamel [8] focused on the health, education, and income dimensions of human development with some notes on participation and empowerment. He asserted that ICTs can enhance capabilities for human development when applied with foresight, clear objectives, a firm understanding of the obstacles that exist in each context, and in situations where there are proper policies that establish an institutional framework that promote the use and benefits of ICTs for the poor. Toboso [16] analysed CA and its relation to ICTs in the context of people with disability. He argued that the importance of human diversity in the capabilities and functionings approach called for incorporating disability into any analysis of wellbeing and quality of life. He introduced the idea of 'diversity on functionings', describing the reality of persons who have the potential to access the same functionings as other people, but in a very different way — often through the use of technical tools and technological resources. Smith et al. [17] applied CA by categorising functionings created by the connectedness and information sharing characteristics of mobile phones into three networking dimensions, namely, social networks, economic networks and governance networks. They too argued that

mobiles are making substantial contributions to capabilities and freedoms in the economic, social, and governance spheres.

Our study extends the above research by providing empirical evidence of expansion of the concepts of capability through the use of mobile phones. It covers a wider range of dimensions, based on the perceived value of what is considered as important for wellbeing by the users of mobiles themselves. We interviewed face-to-face in depth. We present this analysis by working backwards, by first identifying what micro-entrepreneurs perceive as their most important achievements in their lives, i.e., their valued capabilities, then by examining how mobile phones achieve and expand valued capabilities. Grounded theory technique was employed. This paper focuses on one the most important wellbeing elements identified by the participants -- family, the need to be close to the family, and to nurture personal relationships. This element is very much in line in Nussbaum's seventh basic capability, which is affiliation. It consists of being able to live for and in relation to others, to recognize and show concern for other human beings, to engage in various forms of social interaction; being able to imagine the situation of another and to have empathy for that situation; having the capability for both justice and friendship; being able to be treated as a dignified being whose worth is equal to that of others.

3 Capability Expansion Potentially Enabled by Mobile Phones

The CA assesses human development by the expansion of capabilities. The more freedom people have in doing what they value most in their lives, the more developed they can be. To give people more freedom, any development initiatives need to provide them with more free choices and opportunities. In the CA, the relationship between commodities (goods and services), functionings, and capabilities is of particular importance. Sen argues that goods and services are important only in the sense that their characteristics enable people to do and to be, i.e., focusing on the question as to what a person can generate from goods and services. The goods and services are a potential resource for more capabilities.

A conceptualisation of ICTs within the CA framework is proposed by [18] and [9]. Following Sen, both authors described ICTs as a commodity with value only in relation to how they help individuals to do or to be. According to these authors, ICTs should claim a legitimate and central place in the overall capability account. When people are able to make use of ICT to maintain meaningful associations with one another or to earn a living, where they could not before, we can legitimately claim an instrumental role for technology in expanding capability and achieving valued forms of functioning [9]. Positioned in this way, mobile phones can contribute to the expansion of capabilities by creating new functionings and enhancing the existing ones. By way of example, creating the functioning of mobile communication and enhancing the functioning of long-distance communication, can also extend thinking capabilities and provide information. Therefore, mobile phones affect the personal, social and environmental conditions that enable or constrain the generation of

capabilities. They also affect personal preferences or needs which will in turn influence people's choices of realised functionings. The following example illustrates these propositions.

Let us think of the imagined life of Amir, a "penjual bakso" (meatball soup hawker), a very common type of micro-entrepreneur in Indonesia. A meatball hawker goes around the streets and alleys in his neighbourhood selling meatballs by knocking on a soup bowl with a spoon. It creates a very distinctive and well-known sound. Now, let us imagine Amir being given a mobile phone. First, provided that he knows how to use the mobile phone, he can call his family while he is working; thus Amir's mobile phone generates a new functioning, namely mobile communication. Then, let us say, Amir receives orders from his customers through his mobile phone. In this case, Amir's mobile phone is acting as a conversion factor that is helping to convert another commodity (the meatball business) into another functioning, namely a delivery service. It also makes his work more efficient, even profitable.

Then, Amir receives an SMS from his friends saying that a certain alley has more potential customers. Amir then could plan to stay longer in that alley to sell more meatballs. In this case, Amir's mobile phone is acting as a conversion factor enabler. His knowledge has been enhanced by the information he has just received. Finally, mobile phones can act as a choice developer. Amir usually sells meatballs on the street because he has to travel to his customers. Having a mobile phone enables him to receive orders from his customers at home. A new option has developed, i.e., selling more meatballs from his home, and Amir will probably choose this option since it requires less walking, is safer and he can remain closer to his family. It is clear that in Amir's case his capabilities and functionings are both improved a lot.

4 Research Context

This section of the paper describes the background context of this research. The study is based on research conducted in Indonesia in 2009-2010. The first part will outline the general context of micro-enterprise and the second part will describe the mobile phones industry, its penetration and adoption in Indonesia.

4.1 Indonesian Micro-enterprises

Small and medium enterprises (SMEs) in Indonesia have historically been the main player in domestic economic activities, especially as a large provider of employment opportunities, and hence a generator of primary or secondary source of income for many households. Micro-enterprises are the smallest, but most numerous businesses within the larger group of SMEs. Recent statistics [21] show that in 2010, micro-enterprises account for approximately 98% of all enterprises in Indonesia, absorbing 93 million labourers, of the 102 million which is the total labour force in Indonesia.



Fig. 1. Indonesian archipelago with marked sample location

In the reported research project, empirical data was taken from two cities, viz., Makassar and Bandung, which represent the eastern part and the western part of the country (see Figure 1). Using an in-depth semi-structured interview instrument, a total of 39 micro-entrepreneurs were interviewed. The interview instrument consisted of questions about their perceptions of wellbeing, that is, what are the most important values or elements of their wellbeing, and how mobile phones may contribute to these values.

Participants were chosen by snowballing technique from a broad range of businesses within the sectors for trade, manufacturing and services industries. In the study there are 51% female participants and 49% male participants. Their income varied from IDR 500,000 (AUD 60) to IDR 6 million (AUD 600) per month, with majority of interviewees earning an average of IDR 2.5 million (AUD 300) per month. Eleven of these micro-entrepreneurs were self-employed with no support staff. Most enterprises used simple technology, had limited access to credit, had limited managerial skills, and operated in the informal sector. For the research, micro-enterprises were grouped according to their location (Bandung or Makassar), products or services categories (e.g., shoe-maker, furniture-maker, blind masseur), or those belonging to the same service organisation. This grouping was based on the assumption (later shown to be true) that communication activities and interaction amongst similar businesses in such groupings are more likely to occur.

4.2 Mobile Phones in Indonesia

In Indonesia, mobile phones have become an essential part of people's daily lives. Almost everywhere one sees more and more Indonesians engaging with their mobile phones -- making calls, sending text messages, or simply listening to music. Mobile phones are no longer considered as a luxury item now that they are available in their

cheapest form for only USD 10. With an average growth of 60 percent per year since 2005, mobile phone penetration in Indonesia stood at 88 percent in 2010 with more than 200 million subscribers altogether [19]. At the same time, the number of landlines which stood at 25% has dropped 11% in 2010 [20].

The Indonesian telecommunications market is considered unique. This can be seen from the way consumers in Indonesia have mostly headed straight to mobile phones as their communication tool whereas consumers in most countries usually progress from having no connections to adopting landlines and subsequently mobile phone [20]. Most people have never really has experience with fixed lines. Research by Nielsen Company revealed that the Indonesian mobile consumers are getting younger. Much of the mobile growth is being driven by teens, with more than 70 percent having a mobile phone connection. The number of teens aged 10 to14 having mobile phones increased more than five times during the five year period of 2005 to 2010 [20]. Today's Indonesian teenagers are using mostly instant messaging or chatting of the phones which is preferable than voice calls or texting.

The research also found that Indonesian mobile subscribers are spending less now than they were five years ago, with 58 percent of consumers spending less than Rp. 50,000 (@USD 5) per month in 2010 compared to only 18 percent in 2005. The reason for this decline is partly because the prices of many services are down but more importantly, more new consumer segments with limited spending capacity are entering the market. Low rates remain the top factor for consumers when selecting a service provider, but most consider the reputation of networks and recommendations of friends and family, indicating that while dropping tariffs are starting to drive operator choice, consumers continue to be concerned about service quality when making their choice [20].

5 Family and Personal Relationship: Micro-entrepreneurs Valued Capabilities

Over time, the global community has in effect been moving towards conceiving development as the organised pursuit of human wellbeing. However, the notion of wellbeing is still a novel category such that no settled consensus on its meaning has yet emerged. According to Gough & McGregor (2007)

the conception of wellbeing is the one that takes account of the objective circumstances of the person and their subjective evaluation of these. But both of the objective circumstances and perceptions of them are located in society and also in the frames of meaning which we live. Thus wellbeing is also and necessarily both a relational and dynamic concept.

This implies that in order to understand the role of a technological artefact such as a mobile phone for an individual's wellbeing, his or her personal understanding of wellbeing must also be taken into account.

This section of the paper investigates the value of family and therefore personal understandings of relationships by participants in general, based on interviews with

micro-entrepreneurs in 2010. The next section will discuss how mobile phones are involved in shaping and facilitating these valued interactions. A preliminary finding in this study shows that participants put families as the top priority and the need to be close to family is very important. For the participants, the term ‘family’ would include not only parents and children but also grandparents, uncle and aunts, cousins, nephews, and other extended family. A sense of togetherness is so very strong that many families organise many gatherings or events to bring the family together regularly.

Family is very important. I won't feel good if we're not together (Sri).

The above quote captures the participants' perspective about family, that family is very important. Family was always uppermost when participants were asked about their wellbeing. The first and immediate response was almost always about family wellbeing. It seems that interviewees do not regard their wellbeing as their own, but rather it is an integral and inseparable part of their family wellbeing. This is depicted in the responses which almost always refer to “our” instead of “me”, the individual self:

Alhamdulillah [Thanks to God], my earnings can fulfil our daily needs (Hasna).

The wellbeing of the family mentioned by participants included family livelihood, children's education, maintaining good relationships with spouses, being dutiful to parents, and helping other relatives. Spontaneous concern for family livelihood is the most important aspiration. Many interviewees believed that they are materially well-off if they just earn subsistence for the family:

I have so many things that I want to do, but I don't want to think about it. I live my life as it is, as long as our daily needs are fulfilled (Timan).

Since family is very important, it is to be expected that participants expressed the need to be constantly close to the family. Some micro-entrepreneurs will bring their family to wherever they decide to set up a business. Others must leave their family in the village for business for various reasons. When asked, participants left their family behind because of their negative views of the urban lifestyle, and its potential damage to their children's upbringing. For those that have to leave their family to go somewhere else to find business opportunities, the goal is to someday go back to the village where they can reconvene with their family. Often when separated, there are issues that might compromise relationships with the family. These issues must be addressed and the key is communication. If not handled appropriately, separation might result and a weakened affinity towards each other. One masseur in Bandung expressed a very human concern:

My wife is also blind. She used to work with me here in Bandung but when we had children, I suggested that she stay home [in the village]. I know that our children might have some feeling of inferiority to have blind parents and if they do not have strong affinity to us, I am worried that they might feel embarrassed and that they will not acknowledge us as their parents. That is why I asked my wife to stay close with them (Eden).

The beneficiaries of the family earnings are largely the children. For many participants, it is very important for their children to be successful and they agree that education is the way forward. Therefore many believe that sending children to school is the utmost priority. They take all necessary means to ensure that their children receive an education. In some extreme cases, participants are prepared to set aside their personal dreams and aspirations so as not to interfere with this priority goal. They put a lot of effort into ascertaining that their children focus on their education. Many parents regularly monitor their children's study and academic achievements. Evidence of the role of mobiles in relation to education will be presented.

Maintaining good relations with spouse and parents were also mentioned frequently by the participants. They perceived harmonious relationships as the key to a happy marriage, and a happy marriage is one of the elements of prosperous life:

A prosperous life in my opinion is, first in economic terms, that our needs are fulfilled, and secondly, at home where there is harmony and peace with the family. I'll be happy if my desires were realized and if my wife agrees with me (Aris).

Parents are much respected in Indonesian families. It is part of long tradition in local culture that parents are a central place in the family. Their religious teachings endorse the highest regards for parents, that parents, in particular mothers, should be placed above other people. So, it is understandable that participants have also expressed the need to help parents and be dutiful to them.

I no longer live with my mom, so I try to do anything I can to help her. It is my way of being dutiful to my parents (Ani).

6 How Mobile Phones Enhance Personal Relationships

Good family and personal relationships are a much valued capability. Participants agreed that mobile phones had become a necessity for facilitating communication within family, with friends and wider community. As a communication enabler, the phone has provided valuable functionings for strengthening family relationships, providing ways of dealing with personal issues, and enhancing social relations. Mobile phones assisted with consultation and getting advice about personal and/or religious matters. Mobiles also helped to cope with loneliness.

6.1 Vignette: Ani and Her Sony Ericsson

Before discussing the valued functionings, we present a vignette from the fieldwork (by the first-named author) in Makassar, Indonesia, that reflects the range of mobile phone uses for promoting personal relationships.

Ani is a 32 year-old lady who owns a boutique. Her boutique is quite unique in that she only sells Moslem apparel (clothing suitable to Islamic sharia¹). Lately, she

¹ Sharia is the moral code and religious law of Islam. Sharia deals with many topics addressed by secular law, including crime, politics and economics, as well as personal matters such as sexual intercourse, hygiene, diet, prayer, and fasting.

complemented her range to add herb-based cosmetics and remedial products. She is married with 3 children and her husband works as an uztad (Islamic Scholar) who provides voluntary religious lectures and consultations in many Moslem communities in Makassar. He also helps Ani with the business. She owns a Sony Ericcson phone with call, text, and Internet capability, as well as big memory capacity. Even though she no longer lives with her mother (a widow who now lives with her younger sister not far from her house), Ani is very close to her and tries to maintain the same interaction as when she was still living with her.

She uses her phone very often to call the family, sending SMS and Facebook messages. She also likes to chat with a friend who has moved overseas. She feels that mobile phones have provided an efficient way of being close to people and to improve her relationships with family.

I feel closer with my family and friends and it is very convenient for keeping in contact with them. If we can't meet, then we'll call each other.

She further asserts the use of mobile phone helps to manage domestic activities, especially with helping her mother. She underlines her obligation to parents.

With a mobile phone I can finalise a lot of urgent matters. Sometimes my mother needs me to buy something for her, to send money to the bank, or to buy medicine for my sister, and it will be difficult for her if I don't have mobile phone. So it helps me to help my mother as it is a way to show my devotion to my parents.

Aside from running her business, she is also an Uztadzah (female Islamic Scholar) so she frequently uses her phone for this role. She distributes information to members of her congregation and provides or asks for consultation.

I use my phone to send information when required to the people. They can also call me for personal consultation on religious and/or personal problems, whereas I can call my mentor for consultation about my lectures.

Thus Ani agrees that mobile phones are very helpful in many ways. Nevertheless she is still cautious when getting a call from strangers, messages at night, or in the case where a woman kept calling and texting her husband:

I hate it when a stranger calls just to get acquainted, or I hate those who keep sending messages at night. There was also this woman who was trying to seduce my husband. She kept texting him with love poetry. I told her not to contact my husband directly as it is the norm that when someone is married, a lady should speak to his wife about her intentions, and she did, but all she said was how much she liked my husband ... aargh! (Ariani)

6.2 Mobile Phones Bring Family and Friends Closer

Ariani's story illustrates that communication is crucial for maintaining relationships. It helps to nurture interaction with family members, friends and other supportive

relationship. These communications have been greatly facilitated with mobile phones, as expressed by this blind masseur in Bandung who left his family in the village:

For me, the mobile phone is the most important tool because that's how I can communicate with my family in the village (Aris).

For those participants who are separated from their families, communication is even more imperative, and mobile phones seem to be the most affordable and efficient alternative.

My family is in Garut [200 km south east Bandung]. My wife stays there to look after the children, so other than for business, I use my phone mainly to call my family. I even give mobile phones to my children so that I can monitor them, particularly about their study, from here (Eden).

For some participants mobile phones are often regarded as the only option. Physical encounters are always preferable, but they require cost and extra time, which might not always be affordable for many, so communication over the phone is seen as a substitute for face-to-face encounters. Dewi, a masseur who runs her business from her rental studio, expressed the dilemma well:

My daughter is in Sukabumi [200 km west of Bandung] with my parents. I call her quite often via mobile phone ... She wants a Blackberry phone, but I don't have the money yet. I miss her a lot, but I can't go there every time, only when I have spared enough money for the bus. But at least I can still call her (Pur).

Mobile phones were used to maintain contact with friends, talk to and monitor children when away, get news from relatives, or for something as simple as asking about the whereabouts of a family member. The breadth of contacts is remarkable.

I have a friend who is from the same village. She has moved somewhere else. We were very close, so we both felt very lonely. We often send each other messages, but it is limited. So we agree to join the same provider, so we could call each other for a very cheap rate. That way we can talk as often as we want (Eti).

My husband travels a lot and I need to know where he is and how he's doing, so the mobile phone has been very helpful in that sense, such as now. He's already late coming home; I could call him to ask why he is late (Hasna).

It will be very difficult if we don't have a mobile phone. I wouldn't be able to know if, for example, my nephew got sick, or if my children need me to pick them up from school (Alo).

When separated or away, communication with children is so imperative, that many participants have equipped their children with their own mobile phones. Conversations with children during these “monitoring calls” is mostly about the children’s education, e.g., “Have you done your homework?”, “How’s your result this

semester?”, or the parents instruct and ask about their personal/social relations, such as “Take care of your mother”, or “Who are you going with?”

I use my phone to monitor the children as well. We do not spend much time together, as I go to my workshop in the morning and will not be home until late (Dewi).

The capacity of mobile phones to provide on-the-spot communication is very useful for many other situations where face-to-face encounters are not possible. An example was given by a natural cosmetic manufacturer in Makassar who used the mobile phone to guide his dying father:

It'd be very difficult for me without a mobile phone. Once, my father was very ill, and I was out of town. My sister said to come home straight away as his condition was critical. I am worried that I might not be able to get to him on time. So I called him even although I knew that he could no longer talk. I just read prayers to his ears over the phone for hours. I didn't want him to die not thinking of Allah. That's when I realised how important this mobile phone is. (Eda).

Another instance came from a catering owner whose husband was detained by a foreign government and held incognito in an isolated place overseas:

I can only communicate with my husband over the mobile phone, but because it is very expensive to call overseas, we use SMS more often. I don't know how he gets a mobile phone but I know that I can call or send him SMS at a certain time every day (Yani).

6.3 Mobile Phones Provide a Medium for Self-expression and Self-evaluation

Mobile phones help to enhance personal relations by providing a channel of self-expression and self-evaluation. Some participants have expressed that their relationship with their spouses are improving as the husband can now articulate romantic feelings over the phone, which he could not do in real life:

My husband is not a romantic person but with mobile phone he became one ... He is a working type, doesn't talk much, let alone say something romantic. But, I started sending him romantic SMSs, then he start texting me with “sweetheart”. Now he can say even say those words directly (Wiwik).

As in many Asian cultures and with religious backgrounds, partners are not supposed to show signs of affection on public, even when they are married couple. Mobile phones have provided a means around this limitation. Couples can express their affection in private ways like never before. That mobile phone has allowed them to be intimate and to talk about things that might have not been possible otherwise:

My husband is not very romantic, and we try to make sure not to show our affection in public. It was not until he had to go to Saudi Arabia that I learned

that he could be romantic -- and even more. He could discuss things that were meant to be discreet or were regarded as taboo, such as sex. He would have never talked about sex directly to me like he did over SMS or phone. So, I think the mobile does give you more freedom to express yourself (Tati).

Others have said that mobile phones have provided the means whereby they could control emotions when having arguments -- by sending messages using SMS. With SMS, words can be arranged in such a way that convey a special moderated meaning, at the same time avoiding coarse language. Also, stored messages can be revisited and have been used as self-reflection.

I discuss most problems with my husband over the phone, it is easier. Moreover, with face-to-face chats, expression is shown clearly in ways that sometimes are not easy, when you are expected to control your emotions. With SMS, you can think carefully about what to say. We even had fights using SMS, yet had no need to yell or say bad words. Texting polite sentences with piercing words are sufficient (Tita).

My children even prefer if we communicate using mobile phone because they say that I sound nicer over the phone. With face-to-face you can immediately see if someone is angry. Also, using SMS with caution is better when you're angry, as you can arrange the words and punctuation (Wiwik).

Sometimes when we're angry, we might say something different from what we actually want to say. With SMS, all the sentences can be structured. Also, with SMS, we can read it again, so it is a kind of self-reflection, like "Why did I say that?" or "Did I really say that?" [Tati].

6.4 Mobile Phone Helps to Cope with Loneliness and Depression

Mobile phones improve sharing with others. They create ways for dealing with psychological problems, such as loneliness. It was expressed sadly by one participant, a blind masseur who is a widow and lives by herself:

I live by myself, so I am very happy to get a call from anyone. I know someone, he is a Javanese. He calls me every night. I like it when he calls, he is a nice guy and very attentive to me. But I feel embarrassed and at the same time feel sorry for him, because he doesn't know that I'm blind. Sometimes I dream that I could marry someone who's not blind but then I think, there is no way that he would want me (Pur).

Another instance came from a participant whose father just passed away recently. Her sadness was lessened when she felt that many people cared about her, as indicated by the many messages and calls from friends and family:

I realised this when my father passed away. I received so many SMSs from friends, relatives and acquaintances. They made me happy as I feel that so many people care about me... (Eda).

Overcoming depression is also mentioned by one participant. This was conveyed by a clothing retailer who was divorced because her husband had a second wife. Although it is not the phone per se, but the content of the messages, that eased her depression. Her ability to store phone messages, and for her to read them over and over again, tremendously assisted:

Sometimes when I'm depressed, that feeling of being abused comes back. Then, I'd re-open SMSs from my friends, I read those messages of gratitude again for my help that has brought them together for me. I feel so happy that no money in the world can pay for it (Tati).

However, the use of mobile phone for personal relationships also has its drawbacks; as in the case of Ariani above, there are concerns over the potential of the technology to be used by those who are trying to disrupt family equanimity. Mobile phones are perceived as a way of facilitating extra-marital affairs.

I used to hate mobile phones because I thought they facilitated my husband's affair. It started with normal consultation, then there was intimacy and before you knew it, they were already married (Tati).

6.5 Desirable Mobile Services as Perceived by Indonesian Micro-entrepreneurs

The above extracts from interviews have demonstrated that facilities provided by mobile phones have a profound ability to enhance family and personal relationships. Mobile features that are cheap and that improve the quality of communication are generally desirable. Services such as same-provider cheaper or free calls, 3G capability that allows multi-media communications, as well as the capacity to access the Internet, have been used or are desired by participants.

I use mobile phone to call my son who is studying in a boarding school in Bandung. I use Simpati's Time-On² for that because it's a lot cheaper. He's not allowed to have a mobile phone in boarding school, so if he needs to contact me, he can use his teacher's phone to SMS or beep me (Eti).

I teach in boarding school 5 days a week, so I can only see my children on weekends. They stay with my parents. The mobile phone has been very crucial to control my emotions because I feel much better if I know what my children are doing, and where they are going, and all. I even read their 'status' from Facebook in my phone (Tati).

My husband works far from here. We only see each other once a month as he is not allowed to bring his family [to work]. It feels really awful. Luckily there is a mobile phone, so I can still call him even if I can't meet him. It feels so good just listen to his voice. I wish I had that 3G thing so I could see him (Lilis).

² Time-on: A service provided by a telco company, that, when registered, customers can call to any number within the network for IDR 3000 or AUD 0.35 a day.

7 Conclusions

This paper has shown that the primary aspects of wellbeing, as perceived by Indonesian micro-entrepreneurs, are those of family and personal relationships. It is evident that mobile phones not only have the potential to improve relationships, but that they do deepen, expand and enrich relationships in many ways.

The characteristics of mobile phones which provide connectedness and information-sharing have created a range of functionings that were impossible to achieve in other ways. The addition of mobile functionings has certainly affected micro-entrepreneurs' capability sets. They can call family using mobile phones, manage family arrangements, and enhance their personal relations in many ways, as demonstrated by the interview evidence presented. Being able to contact family members anywhere, anytime, or monitor children when away, flows on to contribute to wellbeing by enhancing a sense of belonging and of security. The ability to convey feelings more freely to one another over mobile phones has provided more of a sense of freedom, which is at the core of the capability approach. Although there are some drawbacks -- where mobile phones are used for immoral affairs, which work against norms of wellbeing -- looking at the potential and uses of the phones, we conclude that mobile phones have many more constructive social advantages than disadvantages.

This article is in line with and extends the cited studies on CA and ICTs by suggesting that mobile phones enhance human capabilities. We add to existing knowledge by concluding that CA relates to the technological domain in the sense that technology should not be viewed only as an artefact, but as a commodity and human resource that helps individuals and groups to achieve their valued capabilities. The paper has also shown that mobile phones have expanded many functionings that promote family and personal relationships in the dimensions of education, psychological well-being, and for people with a disability. It can be safely recommended that mobile services fulfil many of a micro-entrepreneur's human needs. Finally, the functionings presented here may not represent a complete set of functionings for all individuals, but they are sufficient to show that the use of mobile phones has expanded the capability sets of Indonesian micro-entrepreneurs in urban contexts, smoothed their relationships, and helped to enable them to live their chosen, valued lives.

References

1. Alampay, E.A.: Beyond access to ICTs: measuring capabilities in the information society. *International Journal of Education and Development Using Information and Communications Technology* 2(3), 4–22 (2006)
2. Alkire, S.: Dimensions of Human Development. *World Development* 30(2), 181–205 (2002)
3. Donner, J.: Research Approaches to Mobile Use in Developing Countries: Review of Literature. *Information Society* 24(3), 140–159 (2008)

4. Giger, B.-S.: Informational Capabilities- The Missing Link for the Impact of ICT on development. World Bank ICT Sector Week. E-Transformation Working Paper Series: World Bank Working Paper Series (2011)
5. Deneulin, S., McGregor, J.A.: 'Capabilities Approaches and the Politics of a Social Conception of Wellbeing'. *European Journal of Social Theory* 13, 4 (2010)
6. Gasper, D.: Sen's capability approach and Nussbaum's capabilities ethics. *Journal of International Development* 9(2) (1997)
7. Gough, I., McGregor, J.A., et al.: Theorising well-being in International development. In: Gough, I., McGregor, J.A. (eds.) *Well-being in Developing Countries*, Cambridge University Press, Cambridge (2007)
8. Hamel, J.: ICT4D and the Human Development and Capability Approach: The Potentials of Information and Communication Technology. *Human Development Research Paper* 37 (2010)
9. Heeks, R., Molla, A.: Impact Assessment of ICT-for-Development Projects: A Compendium of Approaches. *Development Informatics Working Paper Series* 36 (2009)
10. Johnstone, J.: Knowledge, development and technology: Internet use among voluntary secto: AIDS organizations in KwaZulu-Natal. Department of Information Systems, London School of Economics and Political Science, London (2005), <http://csrc.lse.ac.uk/research/theses/johnstone.pdf>
11. Kleine, D.: The ideology behind the technology – Chilean microentrepreneurs and public ICT policies. *Geoforum* 40, 171–183 (2009)
12. Nussbaum, M.: Capabilities as fundamental entitlement. *Capabilities equality: basic issues and problems*, pp. 44–70. Kaufman, Routledge, New York (2003)
13. Robeyns, I.: The Capability Approach: A Theoretical Survey. *The Journal of Human Development* (2005)
14. Sein, M.K., Harindranath, G.: Conceptualizing the ICT Artifact: Toward Understanding the Role of ICT in National Development. *The Information Society* 20(1), 15–24 (2004)
15. Sen, A.: *Development as Freedom*. Oxford University Press, New York (1999)
16. Smith, M.L., Rasyid, A.T., Spence, R.: Mobile Phones and Expanding Human Capabilities. *Mobile Telephony Special Issue* 3, 77–88 (2011)
17. Toboso: Rethinking disability in Amartya Sen's approach: ICT and equality of opportunity. *Ethics Information Technology* 13, 107–118 (2011)
18. Zheng: Exploring the Value of the Capability Approach for E-Development. In: *Proceedings of the 9th International Conference on Social Implications of Computers in Developing Countries*, São Paulo, Brazil
19. ITU, ICT Data and Statistics (2012), <http://www.itu.int/ITU-D/ict/statistics/material/excel/Mobile-cellular2000-2011.xls> (retrieved July 26, 2012)
20. Nielsen Company, Mobile phone penetration in Indonesia triples in five years (2011), <http://blog.nielsen.com/nielsenwire/global/mobile-phone-penetration-in-indonesia-triples-in-five-years/> (retrieved July 18, 2012)
21. Ministry of Cooperatives and SME, SME Data (2011), http://www.depkop.go.id/index.php?option=com_phocadownload&view=sections&Itemid=93 (retrieved March 28, 2012)

An Analysis of Topical Proximity in the Twitter Social Graph

Markus Schaal¹, John O'Donovan², and Barry Smyth¹

¹ University College Dublin
Belfield, Dublin 4, Ireland
forname.surname@ucd.ie

² University of California
Santa Barbara, USA
jod@cs.ucsb.edu

Abstract. Standard approaches of information retrieval are increasingly complemented by social search even when it comes to rational information needs. Twitter, as a popular source of real-time information, plays an important role in this respect, as both the follower-followee graph and the many relationships among users provide a rich set of information pieces about the social network. However, many hidden factors must be considered if social data are to successfully support the search for high-quality information. Here we focus on one of these factors, namely the relationship between content similarity and social distance in the social network. We compared two methods to compute content similarity among twitter users in a one-per-user document collection, one based on standard term frequency vectors, the other based on topic associations obtained by Latent Dirichlet Allocation (LDA). By comparing these metrics at different hop distances in the social graph we investigated the utility of prominent features such as Retweets and Hashtags as predictors of similarity, and demonstrated the advantages of topical proximity vs. textual similarity for friend recommendations.

Keywords: micro blogs, topical proximity, social network distance, friend recommendation.

1 Introduction

The quality and relevance of information is crucial for a wide range of information-related activities, ranging from rational information retrieval to engagement in social interactions. Recently, recommendation systems have tried to improve the quality of search results by incorporating social signals, see for example [1] for an approach to developing a search engine on top of Twitter or the work of [2, 3] to incorporate social signals into mainstream search engines like Google. The objective with these approaches to social search is to increase the likelihood of addressing the searcher's genuine information needs by leveraging the content sharing and search experiences of their social networks in addition to more conventional index-based techniques. Social approaches have become especially relevant in the

context of providing users with more *novel*, *diverse*, and *timely* results; see [4, 5] for recent trends in recommendation diversification. Also, feedback signals in social communities have been used to assess two types of signals per user independently, with respect to both of their roles as target and source of feedback respectively. This way, while processing feedback events in their temporal order, not only the quality of the user as content creator is evaluated, but also the quality of the user as a content evaluator, see [6].

Conventional recommendation systems aim to proactively suggest relevant results to users based on the information consumption histories of similar, albeit anonymous, users. More recently, researchers have acknowledged that similarity alone does not guarantee relevant or optimal recommendations. In particular the importance of real-world user relationships has been highlighted as an important factor to leverage improved recommendation quality. For example, there has been considerable interest in modeling the reputation of users to bias future recommendations from users who are both relevant *and* reputable; see for example [7-9]. By leveraging the social web, research such as [10] explores the use of services like Twitter as a new source of item data and user opinions, showing how even this noisy signal can be used to make reliable recommendations.

In this paper, we quantify the extent to which social neighbors share interests in similar topics. We do this by a novel metric for topic similarity which is analyzed along star chain sequences emanating from a set of seed users. This is a first step towards the development of a concise framework for the evaluation of social recommendation against both rational and social criteria of quality. In this work we focus on Twitter. Twitter is an exceedingly popular microblogging platform and members typically choose their followee connections (friends) according to their topical interests rather than according to their social connections. Prior research in this area has investigated the prediction accuracy for Twitter followees based on latent topic similarity, see [11, 12], for Twitter followees based on text similarity, see [13], and for social link prediction in general, see [14]. We aim to test whether and to which extent distance in a social network correlates with content similarity. Clearly, social proximity facilitates content migration thereby increasing social similarity. But does this similarity extend to content itself, i.e. do people in close relationships talk about similar things? We will give some answers to this question throughout the result section of this paper. However, our main concern and target is the proposition of a *topical proximity measure* and its validation by comparing its expressiveness to a baseline measure using term frequencies. As a consequence, we looked at a variety of methods to compute content similarity based on Latent Dirichlet Allocation, Term Frequency Vectors, and Pearson's Correlation and compared them with respect to network distance.

We base our investigation on a Twitter data set collected during November 2011 at the University of California. We took a reasonably sized sample of tweets per user as the source for a summary document, which was created in various ways; see Section 2.

The remainder of this paper is structured as follows: Section 2 describes our crawling algorithm and the various ways to build user summary documents from the resulting data. Next, we introduce our use of Latent Dirichlet Allocation (LDA) and resulting similarity measures in Section 3. We discuss the potential of our novel topic similarity measure for friend recommendation in Section 4, followed by a detailed analysis of our results in Section 5. We conclude with a discussion and suggest a number of opportunities for future work.

2 Data Crawl, Star Chaining, and User Models

Twitter is an interesting social platform because it combines the follower-followee social network structure with many dynamic sub-networks based on information flow. For example, mentions and retweets link users in ad-hoc networks as they discuss particular topics and events. We are interested in capturing topic associations from free-text content and from hashtags, and performing various similarity assessments based on properties in the underlying social structure.

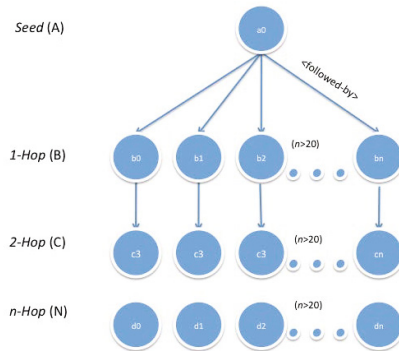


Fig. 1. Example of a “star-chain” pattern from the data crawl. Topic similarity is computed between seed and each hop B-N. This view is for a single user.

Figure 1 shows an example of the structures obtained from the Twitter API, in this case for a single seed user. We started with a seed crawl using the Twitter API and searched for particular keywords, here *Libya*, to allow for a good network density among the Tweepers. After this initial crawl, we had 33740 seed users and considered them as a network sample. We then crawled the other tweets for those users and all of their followers and followees. Starting from the set of seed users, we created chains of length 4, each consisting of a seed user, a friend (or followee) of the seed user, a fof (friend of the friend) of the seed user, and a fofof (friend of the friend of the friend) of the seed user. We only considered seed users for which at least 50 such chains could be found, with members across chains all distinct. In total, we crawled 5628 seed users with more than 20,000 followee 4-chains involving in total 22,549 users. The details of the algorithm used to obtain

the chains are given in Figure 2. Note, we can guarantee that all of the 1-hop and 2-hop members of the chains we created are indeed shortest distance to the seed user, but we cannot give such a guarantee for the 3-hop chain members. We also collected data about mentions, re-tweets and other statistical information about users such as profile age and number of friends and followers.

For each user, we subsequently generated a *user profile document* based on 50 randomly selected tweets from the crawl. Depending on the exact method, we either included retweets, excluded retweets, or selected only retweets. Each user is represented as the collection of words in their tweets, after filtering mechanisms are applied. Filtering is a crucial part of our research since we want to investigate the influence and use of certain elements of language and communication within tweets. Therefore, we give special consideration to *hashtags* as well, in addition to *retweets*. Hashtags are named entities that are used for content tagging and search both within twitter and by external services such as Listorious. Therefore we can either consider them as part of the text (by simply removing the hashtag symbol #), or leave them out, or create the summary document per tweet user solely from them. All in all, we considered five different methods as depicted in Table 1.

Table 1. Representation of Twitter Users

| No | User Representation | ID | Description |
|----|------------------------------|--------|--|
| 1 | All Text | All | The collection of all words in the selection of tweets, but with the hashtag symbol # removed |
| 2 | No Retweets | All-RT | The collection of all words in the selection of tweets without retweets, but with the hashtag symbol # removed |
| 3 | Only Retweets | RT | The collection of all words in the selection of tweets without retweets, but with the hashtag symbol # removed |
| 4 | No Hashtags | All-HT | The collection of all words in <i>All</i> , but without hashtags |
| 5 | No Hashtags, No Retweets | Filt | The collection of all words in <i>All-RT</i> , but without hashtags, fully filtered |
| 6 | Simple Hashtags | HT | Only the set of hashtags in <i>All</i> |
| 7 | Simple Hashtags, No Retweets | HT-RT | Only the set of hashtags in <i>All-RT</i> |

3 Similarity Measures

Based on the user profile documents described earlier, we computed pairwise similarity among all pairs of users and their followees. We used simple Pearson's correlation over the term frequency vectors as a *text-based* baseline and compared it to our *topic-similarity* measure based on Latent Dirichlet Allocation (LDA). For LDA, we consider users as being represented by documents, so the result of training an LDA model for our set of user profile documents is a vector containing topic associations across all topics for each user. In accordance with the *text-based*

```

1: procedure GETSTARCHAINS(Users, Links)
2:   chains =  $\emptyset$ 
3:   for all seed  $\in$  Users do
4:     seedChains =  $\emptyset$ 
5:     F1 = Links(seed)
6:     2A = Links(F1) - F1
7:     2U =  $\emptyset$ 
8:     3U =  $\emptyset$ 
9:     for all f1  $\in$  F1 do
10:      F2 = Links(f1) - {seed} - F1 - 2U
11:      for all f2  $\in$  F2 do
12:        found = false
13:        F3 = Links(f2) - {seed} - F1 - 2A - 3U
14:        for all f3  $\in$  F3 do
15:          found = true
16:          seedChains  $\leftarrow$  (seed, f1, f2, f3)
17:          2U  $\leftarrow$  f2
18:          3U  $\leftarrow$  f3
19:          break
20:        end for
21:        if found then
22:          break
23:        end if
24:      end for
25:    end for
26:    if Count(seedChains)  $\geq$  50 then
27:      chains = chains  $\cup$  seedChains
28:    end if
29:  end for
30:  return chains
31: end procedure

```

Fig. 2. The Star Chaining Algorithm: The outermost loop iterates through all the seed users. The three inner loops are meant to find a sequence of neighbors without repetition and without shortcuts. F_x is the set of candidate users for the x -th position of the chain for $x \in \{1, 2, 3\}$, $2A$ is the set of all friends of a friend for one particular seed user, and xU is the set of users previously used at the x -th position of the chain for $x \in \{2, 3\}$. Note also, that the algorithm does simple backtracking, i.e. if it does not find a third neighbor for a particular second neighbor, then it will try other second neighbors until a third neighbor is found, if possible. However, since the selection of chains is arbitrary, the algorithm does not guarantee the maximum number of chains per seed user to be found.

baseline, we apply Pearson’s correlation among the topic vectors to obtain user-to-user *topic-similarity*. We used an implementation of Blei’s LDA algorithm [15] from the Stanford Topic Detection Toolbox¹, which provides multiple different implementations of Latent Dirichlet Allocation, see [15, 16]. We removed terms consisting of less than 3 characters, occurring in less than 3 profiles and the 30 most common terms. Stop words were also removed. We used 20 topics. In general the term lists (topics) output by the LDA algorithm were representative of human-understandable themes. Table 2 shows example topics from our data set. The two methods are summarized in Table 3.

Table 2. Example LDA term-lists (topics) mined from the Twitter API

| <i>Comment</i> | <i>Topic</i> |
|----------------|---------------------------------------|
| Libya Crisis | gaddafi tripoli nato libyan feb17 |
| Social | today photo stories facebook google |
| US Elections | gop cain president romney perry |
| UK Related | bbc london guardian cameron telegraph |

Table 3. Methods to Measure Content Similarity

| Method | ID | Description |
|-------------------|-----|--|
| Topical Proximity | LDA | Compute User-Topic Associations for a fixed Set of Topics based on Latent Dirichlet Allocation (LDA) on the user profile documents, then compute similarity as Pearson’s Correlation |
| Text Similarity | TF | Compute term frequency vectors for the user profile documents, then compute similarity as Pearson’s Correlation |

4 Friend Recommendation - A Potential Application

Whether or not topical proximity can be used for friend recommendation, that depends on many factors. To get a first understanding about the potential of this application, we conducted the following experiment. For each seed user, we choose 20 neighbors from the hop-1 and the hop-2 layers (10 each). We then recommended the top-10 similar users among the whole list of 20 to the seed user as hop-1 neighbors and measured $B-C \rightarrow B$ *precision*, i.e. the ratio of recommended neighbors that are indeed in the hop-1 layer. Obviously, this experiment is somewhat flawed by the actual social influence among neighbors, but it may give some qualitative insight in the potential of our various methods to support social recommendation.

¹ <http://nlp.stanford.edu/software/tmt/tmt-0.4/>

Table 4. Results of Topic-based Similarity Analysis

| <i>ID</i> | <i>Representation</i> | <i>Method</i> | $sim(A, B)$ | $sim(A, C)$ | $sim(A, D)$ | $sim(A, N)$ | <i>B-C → B Precision</i> |
|-----------|-----------------------|---------------|-------------|-------------|-------------|-------------|--------------------------|
| A | All | LDA | 0.3894 | 0.1550 | 0.0618 | 0.0326 | 0.7225 |
| B | All-RT | LDA | 0.3268 | 0.1500 | 0.0769 | 0.0454 | 0.6912 |
| C | HT | LDA | 0.2490 | 0.0850 | 0.0203 | 0.0089 | 0.5965 |
| D | HT-RT | LDA | 0.1847 | 0.0644 | 0.0123 | 0.0067 | 0.6399 |
| E | All-HT | LDA | 0.3841 | 0.1545 | 0.0602 | 0.0320 | 0.7306 |
| F | Filt | LDA | 0.3129 | 0.1459 | 0.0789 | 0.0446 | 0.7077 |
| G | All | TF | 0.4587 | 0.4323 | 0.4218 | 0.3834 | 0.5147 |
| H | RT | TF | 0.4964 | 0.5373 | 0.5000 | 0.3990 | 0.5258 |
| I | HT | TF | 0.0712 | -0.0023 | -0.0745 | -0.0607 | 0.6091 |

5 Results

To analyze the influence of network distance on topical or text-based similarity between Twitter users, we begin by averaging the topic similarities according to hop distances. In Figure 1, these are $sim(A, B)$, $sim(A, C)$, $sim(A, D)$, and $sim(A, N)$ respectively, where N is an arbitrary set of Twitter profiles randomly taken from our crawled set of seed users. In addition to the average similarity values, we also computed $B-C \rightarrow B$ precision, as discussed in Section 4, to provide a more complete picture. Results are shown for all methods in Table 4 and discussed in the sequel.

5.1 Similarity Analysis of Hop Distance

Table 4 shows a table of 9 experimental conditions. The second column represents the method for user representation from Table 1 above, while column three represents the method for computing content similarity, i.e. either *text-based* (TF - Term Frequency) or *topic-based* (LDA - Latent Dirichlet Allocation). The fourth to seventh column represent the average similarity scores between the seed user and 1-Hop, 2-Hop, and 3-Hop, and n-Hop users along the star-chain, where n-Hop is actually computed based on randomly selected users. These scores are calculated by a simple application of Pearson’s correlation over the topic associations between the seed profile (a_0) in Figure 1, and each profile in the one-hop layer ($B = b_0 \dots b_n$ in Figure 1), the two hop layer ($C = c_0 \dots c_n$ in Figure 1) and finally the n-hop layer (the disconnected component of Figure 1). Averages are calculated across all chains irrespective of the seed user they belong to.

Results vary significantly depending on the use of Retweets, the method to create the user profile document, and the chosen similarity measure, albeit the removal of Hashtags does not seem to be important, see cases *E* and *F* which are not very different from cases *A* and *B*. For the same reason, we skipped

computing their corresponding *text-based* (TF) similarity measures. In the sequel, we will discuss the results in a deeper fashion and extend the analysis in a couple of ways. Even though our main interest is the decay for topical similarity across hop distances, we will first take a look at the differences between baseline text-based similarity and topic similarity.

5.2 Topic vs. Text Similarity

We look at the unfiltered cases *All, LDA* and *All, TF* to examine the differences between LDA-based topic similarity and TF-based text similarity. The topic similarity case exhibits a score of 0.3894 for the hop-1 correlation $sim(A, B)$ decaying to 0.1550, and 0.0618 along the chain ($sim(A, C)$ and $sim(A, D)$) and to 0.0318 with respect to a random node among the seed users. For the text similarity case, we have quite different numbers. However, the decay along the chain is obvious in both cases. For *All, TF*, the values are falling for each pair along the chain, confirmed by a one-sided t-test with significant level 0.05. A deep analysis of the differences is given in Figure 3. Here we looked at the distribution of similarity values, basically a normalized histogram of the similarity values with 7 bins of size 0.2 between -0.4 and 1. Similarity values lower than -0.4 occurred rarely and have been omitted here for presentational clarity.

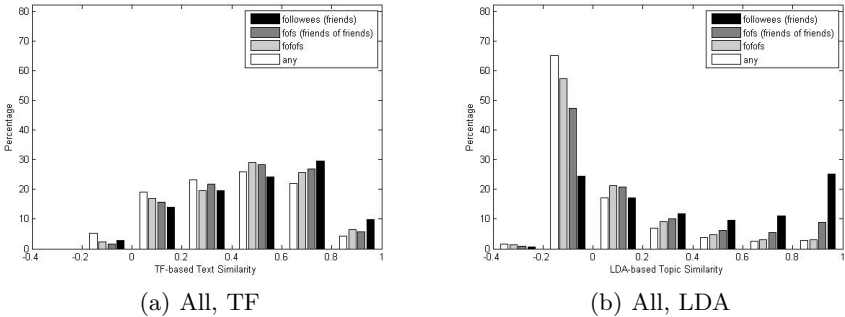


Fig. 3. Text Similarity (TF-based) and Topic Similarity (LDA-based) for Conditions *A* and *G*

The difference in signal clarity between *text similarity* and *topic similarity* is obvious and so the generally higher values of $B-C \rightarrow B$ precision as a proxy for friend recommendation quality or link prediction accuracy are not surprising.

5.3 Retweets and Hashtags

Looking at the same conditions but this time with no retweets (*All - RT, LDA*), we see a slight decrease of -16% in the hop-1 similarity score $sim(A, B)$ which

falls to 0.3268 but there is no large difference at hop-2 and hop-3, so obviously Retweets have no influence on topic similarity beyond direct neighbors, as expected. Again, there is no difference when we apply the hashtag filter, see case *F*. The use of hashtags alone produced lower correlations, see cases *C* and *D* in Table 4. This could be credited to the smaller amount of available text data as Blei reports in [15] that LDA requires large amounts of text as input to function well. However it is more likely that hastags are free of shared vocabulary or irrelevant chat topics and therefore provide a better means to see the true topical relationships between neighbors, especially if we remove retweets as in the *false*, *HT*, *LDA* case.

In order to understand the exact difference between the methods according to different hop distances, we produced distribution graphs for conditions *A-D*, see Figure 4.

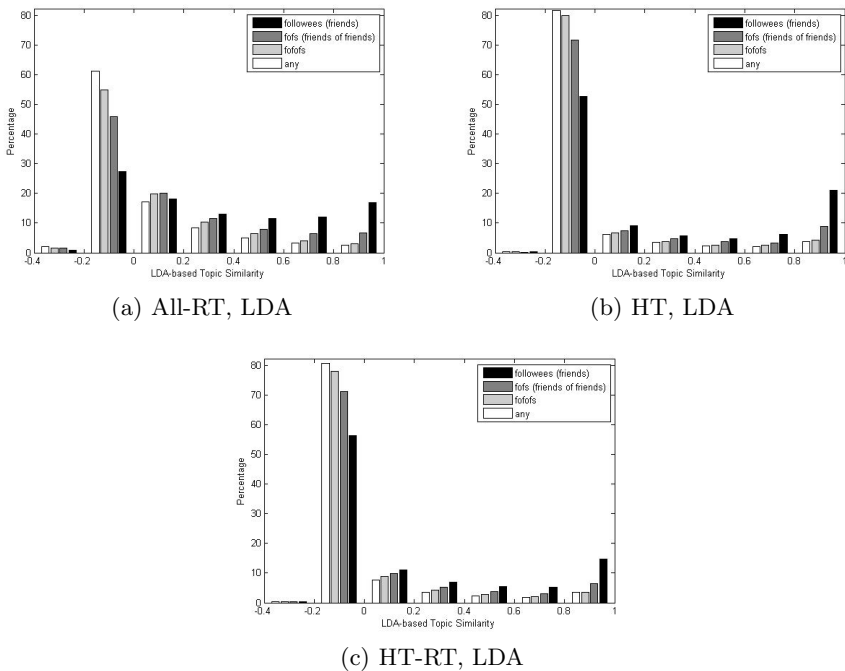


Fig. 4. Network-based Comparison of Topic Similarity (LDA-based) for Conditions *B*, *C*, and *D*

As expected, 1-hop neighbors are leaning towards higher similarities. There is a fraction of 1-hop neighbors with a very high correlation and this may indicate selectiveness in choosing your friends according to particular topics, but it could also have other reasons. Neighbors 3-hops away from the seed user did not exhibit a big similarity difference from the randomly chosen (n -hop) set.

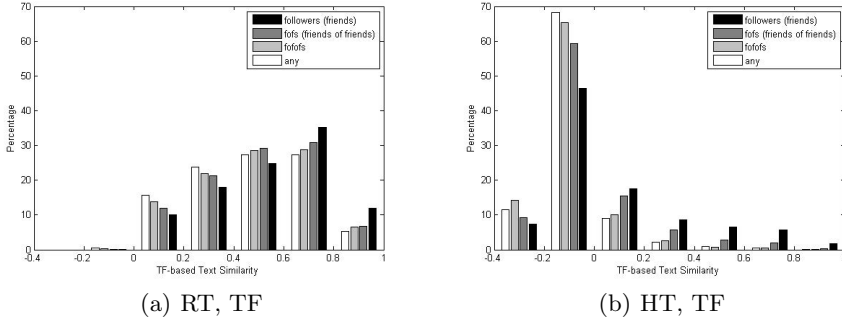


Fig. 5. Network-based Comparison of Text Similarity (TF-based) for Conditions H and I

5.4 Analysis of Decay Patterns

To analyze whether or not the decay pattern is different for different groups of users, we classified groups according to their average $sim(A, B)$ value. Without loss of generality, we chose the $All - RT, LDA$ condition for this analysis. We classified the users into three groups depending on the mean value of their Hop-1 topical similarity value. The results are shown in Table 5. The left two columns define the min and max of the range of mean $sim(A, B)$ values for the group, the third column is the count of users in our result data set, and the third and fourth column are the averages $sim(A, B)$ and $sim(A, C)$ across that group of users. Note, the aggregated results in the bottom row are not exactly the same with Table 4, due to a different aggregation technique and due to different processing, i.e. we needed to remove an entire chain here whenever one similarity value was missing. The last column is the *decay* factor, i.e. the percentage of the value in column 5 with respect to the value in column 4.

Table 5. Decay Pattern for different User Groups

| Group | Min | Max | Count | $sim(A, B)$ | $sim(A, C)$ | <i>decay</i> |
|--------|-----|-----|-------|-------------|-------------|--------------|
| low | -1 | 0.1 | 2192 | 0.0559 | 0.0758 | 1.3556 |
| medium | 0.1 | 0.4 | 2271 | 0.2976 | 0.1426 | 0.4790 |
| high | 0.4 | 1 | 2963 | 0.6117 | 0.2200 | 0.3597 |
| all | -1 | 1 | 7426 | 0.3369 | 0.1507 | 0.4473 |

Not surprisingly, the *low* group seems to have a random distribution of topical similarity across its neighbors, i.e. their similarity is not very different from the similarity of Hop-2 neighbors. Users with high Hop-1 topical similarity also have a higher Hop-2 topical similarity with respect to the the *medium* group.

5.5 Single User Analysis

Our star-chaining approach extracts at least 50 chains per user and thereby facilitates the analysis of similarity distributions for individual seed users. We chose 5 such users from the *All – RT, LDA* case and show their results in Table 6. The new second column is the range of the chain count that was used for computing the *Hop-1*, *Hop-2*, *Hop-3*, and *Any* averages in columns 3 – 6. We illustrate the distributions across similarity values as before, for *Users 2-5*, see Figure 6.

Table 6. Topic-based Similarity Analysis for Individual Users

| User ID | # | $sim(A,B)$ | $sim(A,C)$ | $sim(A,D)$ | $sim(A,N)$ |
|---------|-------|------------|------------|------------|------------|
| 1 | 26-38 | 0.2522 | 0.0445 | 0.0040 | -0.0665 |
| 2 | 54-72 | 0.1570 | 0.2499 | 0.2455 | 0.1491 |
| 3 | 64-67 | 0.4028 | 0.1600 | 0.0980 | 0.0506 |
| 4 | 78-86 | 0.5657 | 0.2086 | -0.0076 | 0.0290 |
| 5 | 72-77 | 0.1429 | 0.0224 | -0.0155 | 0.0162 |

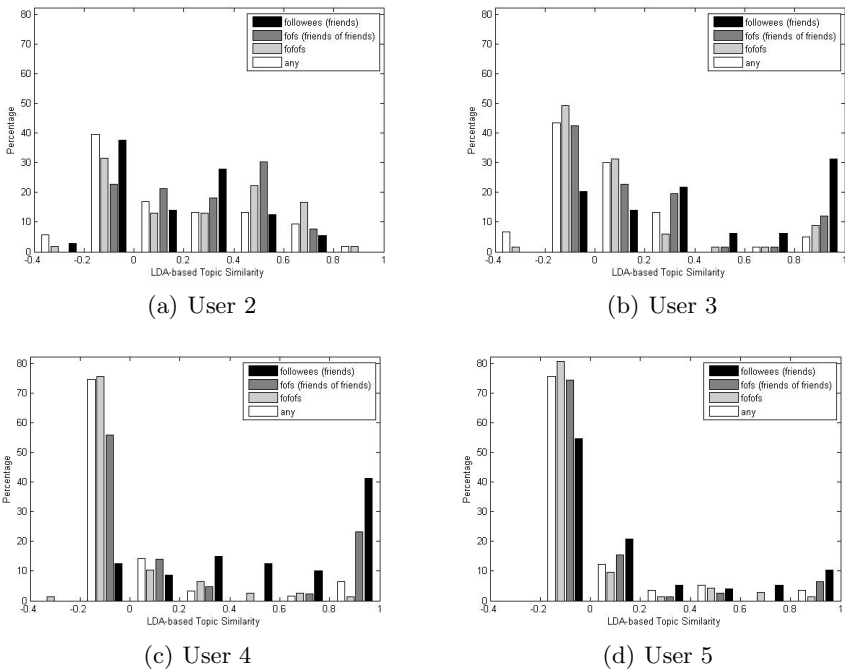


Fig. 6. Comparison of Individual Users

We want to share a couple of observations, even though their value is limited by the fact that such a small sample of users with a small number of chains is not representative by any statistical means.

- *User 2* does not seem to be topically selective with respect to his followees, since there is no big difference between Hop-1 and Hop-N.
- *User 3* is very similar to some of his followees, but this might be a result of direct Retweets, as there is no clear difference between *Hop-2*, *Hop-3*, and *Any*.
- *User 4* is very similar to some of his followees, and topically very different from many other users on all Hop-distances. Probably this is a user with a special interest shared by her direct neighbors.
- *User 5* may also have a special interest explaining his low overall score in topical proximity, but he is different from *User 4* since he does not even show similarity to some of his followees.

5.6 Potential as a Friend Recommender

$B-C \rightarrow B$ precision as shown in Table 5 is just an indicator for the decay of topical proximity between hop-1 and hop-2 neighbors. Nonetheless, it is remarkable that the values are much higher for LDA-based topical proximity than for term-frequency-based text similarity. For text similarity methods, hashtags are better than the full text, but for LDA-based proximity the opposite is true, probably because hashtags are easily adopted among neighbors and therefore provide proximity clues that are automatically detected by LDA.

5.7 Discussion of Results

The results illustrate the power and applicability of LDA-based Pearson’s correlation as a measure for topical proximity in Twitter. Not only are topic similarity signals stronger than text similarity signals and support better link prediction, but also they allow for a fine-grained and detailed analysis and juxtaposition of topical proximity vs. network proximity. We showed that topic similarity with full texts is a stronger predictor of 1-hop neighbors than hashtags, no matter whether we use LDA- or term-based Pearson’s correlation for the latter.

In Section 5.5, we demonstrate by a detailed analysis of 5 selected users that there is a huge variety of different cases which could easily be distinguished by the presented research, and which is one of the venues to be followed up in future research. We showed that the removal of hashtags from the tweets prior to processing only has a very minor influence on topical proximity.

6 Conclusion and Future Work

In this paper we have presented an analysis of topic and content similarity along the twitter social graph. By computing similarity between users based on their underlying topical proximity, and comparing it against a text-based term frequency approach, we quantify the notion of topical similarity among micro bloggers and demonstrate distinct similarity patterns through a novel use of LDA.

There are a number of implications of the results of this study for Twitter users and application designers, particularly with respect to applications like social recommender systems [17] wherein there is usually an accompanying set of social links to augment traditional nearest-neighbor selection strategies. We believe that our approach and findings can be applied across any corpus of text-based content that has an associated underlying network structure. The results of our topic-similarity analysis show that 1) there is generally a strong LDA topic similarity between direct neighbors in the social graph, and 2) this value decays to a standard value usually by the third hop, and in some cases at only two hops in the Twitter graph. We showed that and how those findings can easily be used to develop an LDA-based topical friend recommender.

Future research is directed towards the evolution of a concise framework for automatic measurement of various qualities relevant to the user in a social network. The exploitation of this framework for building applications and evaluation of recommender systems is an obvious opportunity for us and others alike.

Acknowledgment. This work is supported by Science Foundation Ireland under grant 07/CE/I1147.

References

1. Phelan, O., McCarthy, K., Bennett, M., Smyth, B.: Terms of a Feather: Content-Based News Recommendation and Discovery Using Twitter. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudooh, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 448–459. Springer, Heidelberg (2011)
2. Smyth, B., Briggs, P., Coyle, M., O'Mahony, M.: Google Shared. A Case-Study in Social Search. In: Houben, G.-J., McCalla, G., Pianesi, F., Zancanaro, M. (eds.) UMAP 2009. LNCS, vol. 5535, pp. 283–294. Springer, Heidelberg (2009)
3. McNally, K., O'Mahony, M.P., Smyth, B., Coyle, M., Briggs, P.: Social and collaborative web search: an evaluation study. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, IUI 2011, pp. 387–390. ACM, New York (2011)
4. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM 2009, pp. 5–14. ACM, New York (2009)
5. Vargas, S., Castells, P.: Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23–27, pp. 109–116 (2011)
6. Schaal, M., Fidan, G., Müller, R.M., Dagli, O.: Quality Assessment in the Blog Space. *The Learning Organization* 17(6), 529–536 (2010)
7. Bourke, S., McCarthy, K., Smyth, B.: Power to the people: exploring neighbourhood formations in social recommender system. In: Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys 2011, pp. 337–340. ACM, New York (2011)
8. Jamali, M., Ester, M.: A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, p. 135. ACM Press, New York (2010)

9. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI 2005, pp. 167–174. ACM, New York (2005)
10. Esparza, S.G., O'Mahony, M.P., Smyth, B.: Effective Product Recommendation using the Real-Time Web. In: Bramer, M., Petridis, M., Hopgood, A. (eds.) Research and Development in Intelligent Systems XXVII. Springer, London (2011); Proceedings of AI 2010
11. Puniyani, K., Eisenstein, J., Cohen, S., Xing, E.P.: Social links from latent topics in microblogs. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, WSA 2010, pp. 19–20. Association for Computational Linguistics, Stroudsburg (2010)
12. Kang, B., O'Donovan, J., Höllerer, T.: Modeling topic specific credibility on twitter. In: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, IUI 2012, pp. 179–188 (2012)
13. Hannon, J., Bennett, M., Smyth, B.: Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 2010, pp. 199–206. ACM, New York (2010)
14. Pennacchiotti, M., Gurumurthy, S.: Investigating topic models for social media user recommendation. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, pp. 101–102 (2011)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (2003)
16. Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A.U., Newman, D., Smyth, P.: Topicnets: Visual analysis of large text corpora with topic modeling. *ACM TIST* 3(2), 23 (2012)
17. Woerndl, W., Groh, G.: Utilizing physical and social context to improve recommender systems. In: Web Intelligence/IAT Workshops, pp. 123–128 (2007)

A Foresight Support System to Manage Knowledge on Information Society Evolution

Andrzej M.J. Skulimowski^{1,2}

¹ AGH University of Science and Technology,
Chair of Automatic Control and Biomedical Engineering,
Decision Science Laboratory, Al. Mickiewicza 30, 30-050 Kraków, Poland

² International Centre for Decision Sciences and Forecasting,
Progress & Business Foundation, 30-041 Kraków, Poland
ams@agh.edu.pl

Abstract. In this paper we present an intelligent knowledge fusion and decision support system tailored to manage information on future social and technological trends. It focuses on gathering and managing the rules that govern the evolution of selected information society technologies (IST) and their applications. The main idea of information gathering and processing here presented refers to so-called real-time expert Delphi, where an expert community works on the same research problems by responding to structured questionnaires, elaborating complex dynamical system models, providing recommendations, and verifying the models so arisen. The knowledge base is structured in layers that correspond to the selected kinds of information on the technology and social evolution, uses, markets, and management. An analytical engine uses labelled hypermultigraphs to process the mutual impacts of objects from each layer to elicit the technological evolution rules and calculate future trends and scenarios. The processing rules are represented within discrete-time and discrete-event control models. Multicriteria decision support procedures make it possible to aggregate individual expert recommendations. The resulting foresight support system can process uncertain information using a fuzzy-random-variable-based model, while a coupled reputation management system can verify collective expert judgments and assign trust vectors to experts and other sources of information.

Keywords: Foresight Support Systems, Complex Socioeconomic Models, Group Model Building, Knowledge Fusion, Intelligent Decision Support.

1 Introduction

The evolution of modern societies cannot be sufficiently explained without a penetrative study of its technological research, economic, political, and social context. A universe of objects, events and dynamical phenomena, and relations between them, has to be taken into account in order to carry out such a study. These form a complex system [1], [4], [6], [10], [18], [19], usually referred to as the Information Society (IS). Therefore, research on IS modelling methodology can provide clues to building foresight scenarios, eliciting social and technological trends, and planning of future development of information technologies (IT) and their application areas.

A collection of new IS/IT modelling approaches has been elaborated as part of the 'Foresight of the Information Society in the European Research Area' (FISTERA) project [11] financed by the EU within the 5th Framework Programme and applied successfully to model the cohesion processes in the EU New Member States. Some of these methods, in particular the 8-element IS model [11] and the interdependence of trends, events and scenarios constitute the background to defining the knowledge base and information processing requirements of the Foresight Support System presented in this paper. Compared to earlier work on IS/IT models [1],[4],[6],[18], the research results presented in [11] have shown that it is possible to model the essential trends and phenomena of an IS as a discrete-continuous event system. However, building a holistic information system that would require large data sets and massive computation was beyond the research goals of earlier EU-funded projects focusing on qualitative analysis of IS phenomena, such as FISTERA. Creating a comprehensive computational model has been therefore left as an open challenge, cf. e.g. [9].

Having studied the findings of the above-mentioned research on IS/IT modelling, it becomes apparent that in order to follow the continuous emergence of essential new IT and consumer usage scenarios, it is necessary to analyze very large data sets that link different aspects of the IS/IT evolution in one future-oriented model. A straightforward conclusion, relevant to the scope and results of this paper, is that any individual product or technology is embedded in a complex technological, economic and social system in such a way that its evolution cannot be explained without investigating this system in a holistic way. The latter task can be accomplished by establishing an appropriate knowledge base, capable of acquiring, storing and processing information from heterogeneous sources and characterized by different types of uncertainty. Furthermore, the results of the above projects have shown that the sole use of both classical econometric methods and narrative descriptions have proved to be insufficient for achieving adequate IT forecasts or scenarios. Therefore, it has been necessary to elaborate new computational methods based on discrete-event and time-series driven models, and novel decision support methodology to derive foresight-specific rankings and recommendations.

In this paper we provide a report on the design, implementation and use of such an innovative information system equipped with an ontological knowledge base and endowed with analytical data processing mechanisms. It serves as the main component of an IT foresight-oriented decision support system, which will be termed also *foresight support system* (FSS) [20],[14], [2] to emphasize its specificity. The design and implementation of this kind of information system is based on a prior extraction, formulation and analysis of the general rules and principles that govern the evolution of key technologies [14]. In the implementation of the FSS here presented, special attention is paid to models used in the selected areas of information technology under review [9], in the full context of the information society and digital economy. However, the ideas applied to design the FSS described in this paper explore the general principles of organizing foresight research (cf. e.g. [2],[9],[11]) so that they are applicable to support prospective technological studies in other areas. The studies of different kinds of interactions within an IS led to the joint application of discrete-event systems, multicriteria analysis and discrete-time control.

To provide decision support to industrial enterprises, research institutions and governing bodies concerning IT-related R&D management and investment strategies, as well as the definition of legal regulations, a research project has recently been carried out in Poland [9]. Its results are to be constructively applied to developing technological policies and strategies at different levels, from corporate to international. A number of partial goals have been defined, which might help to achieve the ultimate objective described above, as well as being independent research aims in their own right. These include:

- Implementation of an ontological knowledge base which stores heterogeneous data together with suitable technological models, trends and scenarios in the form of so-called proceedings (records of operations) containing data together with records of their step-by-step analyses, results and assessments.
- Elaborating or adjusting methods of multicriteria rankings suitable for IT management and capable of generating constructive recommendations for decision makers as regards the prioritization of IT investments.
- An in-depth analysis of several real-life industrial applications of the decision support system so arisen. The selected technological areas are submitted by industrial partners cooperating on the implementation of the project results.
- A detailed analysis of technological trends and scenarios in areas such as 3D-based e-commerce, expert systems, decision-support systems, recommenders, m-health, neurocognitive technologies, quantum and molecular computing.

Any of the above partial objectives should provide useful solutions to the technology management problems presented by the industrial stakeholders involved in the exercise. This would allow them to apply the knowledge gained to set strategic technological priorities and formulate IT and R&D investment strategies. This is discussed further in Sec.4.

Although the general applicability field of the models studied in [9] is generating trends and foresight scenarios, they can also be used to better understand the role of global Information Society Technology (IST) development trends and to elaborate IS and IT policies in an optimal control framework.

To sum up, the data processing methods presented here as a background to elicit trends and elaborate scenarios of decision-support and decision-making systems can be applied as a universal framework in any future-oriented socio-economic or socio-technological study. As an example application that has been elaborated within the foresight project [9], we will present recommendations concerning the development of some types of decision support systems.

2 The Principles of Information Society Modeling

A user of a technological knowledge base could pose the following question: how does the development of selected information technology depend on global IT development, diffusion processes and on the integration of the IS around the world, driven by global socio-economic trends? We will investigate this question in more detail in the next sections. As regards the global environment, various factors should be considered such as falling telecommunication prices, the growth of information exchange through the internet, rapid diffusion of information on innovations and

technologies, the development of e-commerce, and free access to web information sources. The civil society evolution, driven by the growing availability of e-government services and related web content, has been taken into consideration as well. Finally, the psychological and social evolution of IT users, including all positive and negative i-inclusion phenomena, has to be taken into account as a set of feedback factors influencing the legal and political environment of the IS.

Due to the complex nature of decision technologies that rely heavily on cognitive and social phenomena, it is difficult to create a technology evolution model that is clear, unambiguous and concise. One of the aforementioned earlier findings [11] was that the composite indicators merging data from users with statistical information can rarely provide an adequate description of the technology parameter dynamics. Therefore, when performing the research described in this paper, it was decided that the use of aggregates as the basis of forecasts and recommendations should be restricted to the final visualization of the information retrieved. Instead, during the quantitative analysis phase we have used the basic social, economic and technological data embedded in a new class of input-output models that fit well into the specificity of this kind of technology. For instance, we can separately analyze different groups of potential users characterized by different preferences. Even though their full statistical characteristics are missing, we can explain the development of the market of complementary products that are distinguished by features corresponding to the consumers' preferences. The actual parameters of the groups such as their size, spatial distribution etc. can be estimated *ex-post* based on market data.

In [11] we have defined eight major subsystems of an IS, such as its population, demographics, legislation, IS policies, IT infrastructure, R&D etc. (cf. Fig.1). It has been applied to model the IS evolution in the EU New Member States.

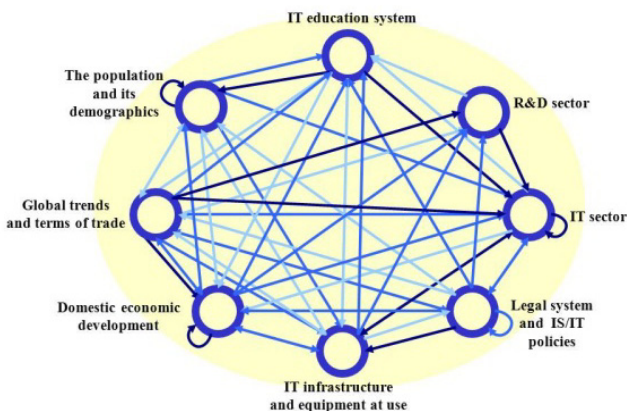


Fig. 1. An example of a causal graph linking the major groups of data used in the IS/IT model

The causal graph presented in Fig.1 contains direct impacts only, i.e. those which are shown within one modelling step. Indirect impacts may be obtained by multiplying by itself the coincidence matrix associated with the direct impact graph. Dark blue edges denote strong direct dependence, medium blue indicates average relevance of causal dependence, and light blue denotes weak dependence between subsystems. The feedback directions are not marked as they may vary for different subsystem variables.

The above assumptions allow us to define the scope of data to be gathered and processed, and they hint at as regards procedures and models to apply. The characteristics of the information stored in the knowledge base are given in Tab.1.

Table 1. The IS model characteristics stored in the knowledge base

| No. | Data description | Data type | Data sources | Typical size of a data set |
|-----|--|--|---|---|
| 1. | Time series describing quantitative variables in the IS model (macroeconomics, demographics etc.) | monthly to yearly quotes | Eurostat, national statistics | 80 to 100 x (20 to 200) |
| 2. | Auxiliary financial time series (stock prices of IT companies, specialized equity indices, exchange rates, IT-related commodity prices etc.) | from tick-by-tick to daily quotes | financial information providers | from $10^4 \times 20$ to $10^5 \times 500$ |
| 3. | Metadata as an ontology: system description, definitions of subsystems, variables, descriptions of event classes, assignment of variables and events to subsystems, relations between them | OWL code, text, graphics, graph incidence matrices | Experts and analysts involved in modelling | Can vary strongly, e.g. about 10 MB in the example in Sec.4 |
| 4. | Qualitative and quantitative characteristics of past events with the corresponding states of the system, links to data sources | records with heterogeneous information | event streams, news agencies, experts | 100 to 10^5 events, 10kB per record |
| 5. | Qualitative assessments and quantitative characteristics of relations between IS subsystems and between system variables | structured expressions | expert Delphi, statistical calculations | $\sim(10^4 + 64) \times$ (no. of experts) assessments |
| 6. | Annotated source files (bibliographic, patent, personal, research projects, research institutions, IT companies etc. databases) | texts, spreadsheets, files with heterogeneous data | automatic updates, manual data input and annotation | from 10GB to 1 TB |

During analysis of an IS, each subsystem shown in Fig.1. appears as a bundle of discrete events, continuous trends and continuous or discretized state variables. For instance, in the initial model of the Polish IS used in [9] there are 92 variables in total, while the number of variables describing subsystems ranged from 7 for the ICT sector to 17 for the R&D sector. The final set of quantitative characteristics has been selected from a total of 337 variables considered, based on an iterated two-stage procedure: an expert Delphi and calculating statistical relevance of causal relations with standard tests. This approach is justified by the insufficient length of time series (cf. Tab. 1) to rely solely on statistical methods and by the need to verify expert judgments with statistical tests even when their relevance was not perfect.

The dynamics of the system can be derived from past observations forming vector time series. It can be described [15] by the following discrete-time dynamical system

$$x_{t+1} = f(x_t, \dots, x_{t-k}, u_p, \dots, u_m, \eta_p, \dots, \eta_n), \quad (1)$$

where

x_1, \dots, x_{t-k} , are state variables, $x_j := (x_{j1}, \dots, x_{jN}) \in \mathbb{R}^N$,
 u_1, \dots, u_m are controls, $u_i \in [u_i, u_{i+}]$, for $i=1, \dots, m$, and
 η_1, \dots, η_n are external non-controllable or random variables,
 f is linear non-stationary with respect to x , and stationary with respect to u and η .

The coefficients of f can be identified using least-squares or maximum likelihood methods on each subinterval of the modelling period where they were stationary. To cope with the non-stationarity in (1) that usually manifests in abrupt changes of parameters caused by internal (legal system, R&D) or external events, the evolution model was supplemented by a discrete-event system P [7], [10] that represents the dynamics of discontinuous variables, namely

$$P=(Q, V, \delta, Q(0), Q_f) \tag{2}$$

The notation used in (2) is explained in the following Tab.2.

Table 2. The data characterizing the discrete-event component of the IS model

| No. | Symbol in eq.(2) | Data description | Data sources | Typical size of a data set |
|-----|------------------|---|--|--|
| | Q | The set of all feasible states of event-driven model components, stored as labeled narrative descriptions combined with Boolean or fuzzy logic vectors that model the occurrence of predefined state properties | expert analysis of appropriate IS components | 10 to 100 per component |
| | $Q(0)$ | The set of initial states of event-driven model components, used together with causal links as a base to derive transformation rules for P | legislation, R&D state-of-the-art | 10-20 (=no. of discrete-valued components) |
| | Q_f | The set of reference (or final) states of event-driven model components corresponding to alerts or to reporting the modelling results | experts involved in modelling | ~10 MB (including descriptions) |
| | V | The set of admissible operations over the states of discrete system components, derived from rules governing legislation, principles of generating R&D results and innovations | legislation, expert analysis of R&D | 10 to 100 operations per each component |
| | δ | $\delta : V \times Q \rightarrow Q$ – the transition function governing the results of operations over states, stored in form of rules | expert Delphi, rules inferred from cases | 100-1000 rules |

Events in P are defined as pairs of states $e := (q_1, q_2)$, such that $q_2 = \delta_V(q_1)$. Following the above assumptions concerning the controlled discrete-event variables, the operations from V may be either controls, i.e. the decision-maker’s actions over Q , or may occur spontaneously as the results of random processes. Furthermore, we assume that there exists a set $X(Q)$ of quantitative or ordinal characteristics of states, which can be deterministic, interval, stochastic, fuzzy etc. Although (2) is in principle asynchronous, one of the coordinates of $X(Q)$ can be identified with time to couple (2) with (1).

The evolution of the IS can then be modelled as a discrete-time/discrete-event system, where the mutual impacts of each of its elements are represented either in

symbolic form, as causal diagrams, or within state-space models. Some external controls, such as legal regulations and policies, are modelled as discrete-event controls, while the others, such as tax parameters or the central bank's interest rates are included as discrete-time control variables in (1). The exogenous non-controlled variables include exchange rates, energy prices, demographic structure, attitudes towards IT-related learning and so on. Both serve as inputs to the system (1)-(2), while basic social, technological, and economic characteristics are state variables in (1) linked by feedback loops. The parameters of (1) are functions of the states of (2), changing their values when the output $X(Q)$ from (2) is modified by an event. After performing a simulation of external and random variables, and assuming a sequence of controls, output trends can be calculated, allowing us to model the influence of consumer and industrial demand on the IT development, research, production and supply of selected IT or IT-dependent products, as well as on GDP growth rates. Scenarios appear as the results of grouping trends and sequences of events, for different variants of decision variables, random events, and external drivers.

3 The Architecture of the Foresight Support System

The above-presented expert system can serve as a framework to organize the overall information processing during future-oriented research, such as socio-technological foresight. Its main component is the ontological knowledge base fed by collective expert judgments, autonomous webcrawlers and updates by users. The information is verified, then processed by analytical engines. The knowledge base includes ontology management functionality, specifically ontology merging and splitting, evolution registering, operations on metadata and metadata updating protocols as well as the usual data warehousing functionalities such as automatic verification and updating.

The main functionality of the above knowledge-based system, together with its analytical capabilities and automatic or supervised knowledge acquisition, update and verification, is to respond to queries submitted by users and to support their decisions. Further functionalities, such as content marketing, can be included as separate modules. The structure of the system is illustrated in Fig.2 on the next page. It shows the generic structure of the knowledge base, while the architecture of a particular instance of the system will depend on the scope of applications. The focus areas of the research reported in this paper, which are also reflected in the scope of the knowledge gathered and processed and in the system architecture, are listed below:

- Key IS application areas (e-government, e-health, e-learning, e-commerce)
- Expert systems, including decision support systems and recommenders
- Machine vision and neurocognitive systems, including man-machine interfaces

The use of the system consists in applying the research and modeling results in the first two focus areas listed above to elicit development trends and scenarios for more specific IT areas that depend on basic technological trends and socio-economic processes. Thematic databases store the area-specific information, while a common data block contains interdisciplinary information, such as macroeconomic data, social characteristics (employment, education, demographics), geographic information and other potentially useful data. The common block is used for providing decision-making support during specific thematic analyses.

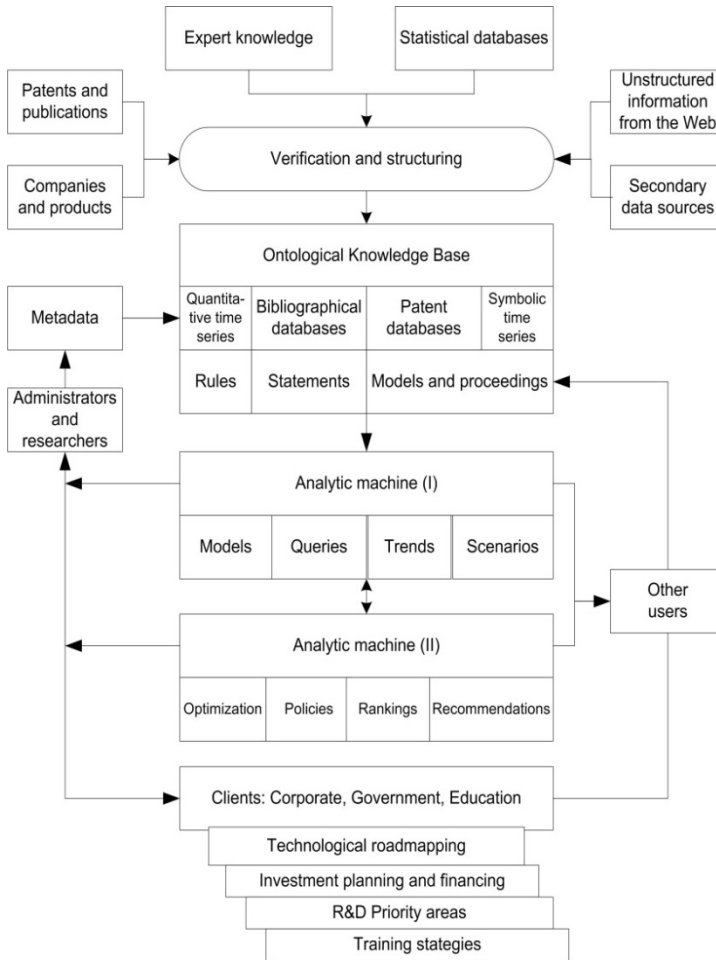


Fig. 2. The scheme of the knowledge base within the foresight support system

The Analytic Machine I contains specialized data fusion algorithms such as:

- Delphi questionnaire analysis, where each Delphi question is associated with a trend or a future event in (1)-(2),
- trend-impact and cross-impact analysis based on experts' judgments,
- consumer preference models, which are to be integrated according to [3], [9],
- specific sector and market models concerning education, health care services, media, internet advertising, quantitative information markets,
- a package of simulation procedures, adaptive trend algorithms and autoregressive time series forecasts.

The R&D trends are derived primarily from biblio- and patentometric data that are partitioned according to the time of appearance and the syntax of the query as proposed in [17]. Other relevant sources of information are technical product characteristics.

As a result of heterogeneous data fusion and processing, an instance of the system (1)-(2) is obtained. Specifically, the data stored in the knowledge base is used to derive the parameters of coupled discrete-event and discrete-time control systems thus providing a tool to elicit trends of state variables as trajectories to (1)-(2). The foresight scenarios are constructed as sequences of output trends and events. Mutually contradictory scenarios are filtered out using causal-anticipatory models [16].

A user's query input to the system may define the time span and parameters for simulations, specify variables or events to be displayed or a composite output function to be calculated from the system variables. The user can also specify the data to be taken into account as a subset of resources available in the knowledge base and provide assumptions as regards his own or a stakeholder's future decisions in the form of rules. They are then taken into consideration during the simulation. The knowledge gathered in the system is continuously updated, represented and processed using causal networks, rule generation from cases, and anticipatory feedback. The system is capable of indicating which data is missing to obtain trustworthy results, thus allowing for interaction with the user, who can input new data, extend its scope under consideration, or start a new round of data acquisition from experts and external sources.

Responding to a query, the system performs the required calculations and returns the specified trends, scenarios and indicator forecasts. All parameters of the model used, including the scope of data, and the results of calculations can be stored in the dedicated object-oriented database as simulation proceedings. They are available for further analysis by Analytic Machine II, for modifications or comparative analysis by the user, who can investigate model sensitivity to parameter change.

The optimization and recommendation results are generated by Analytic Machine II based on outputs from Analytic Machine I. Its engine makes use of multicriteria optimization, outranking methods, ranking forecasts [12] and so on. Consequently, when using models (1)-(2) to generate optimal technological strategies or investment policies, the criteria and goals should be quantified by the user or external customer and associated with state variables, events and decision scenarios. A generalization of the multicriteria shortest-path algorithm [10] combined with discrete-time optimal control [14], [15] can then be applied to variable-structure networks that appear in the simultaneous optimal control of discrete-events [10] and discrete-time dynamical systems.

If the recommendations require multicriteria rankings, their dynamic character has to be taken into account, i.e. future changes in the customer's priorities after certain goals have been achieved or as a result of changing external circumstances. In the present implementation of the model, dynamic prioritization algorithms have been developed as tailored for IT ranking problems.

While Analytic Machine I can respond mostly to research-related questions, any query from an external client will usually refer to the capabilities of Analytic Machine II. However, it must be processed making use of all the system components and databases, even if the global trends and general development models used are not visible in a reply. At this stage, if context-dependent information related to the specific area of the query is missing, the system may require additional input data, and to combine them with those input to the Analytic Machine I.

In Sec.4 we will provide an example application of the above system to derive development trends for decision support and autonomous decision-making systems, which has been selected as the first focus area for the prospective study. The data

necessary to describe the development of decision-support technologies and related social, economic, and technological evolution characteristics has been extracted from the knowledge base created within the research project [9].

4 An Application Example

After selecting the social, technological and geographical focus areas, which can correspond to an instance of the knowledge base outlined in Sec.3, the first task of an analyst is to obtain rankings of the technologies, markets and application areas with the most potential to be subjects of a detailed study. These rankings result from an expert and practitioner pre-delphi, which are merged interactively with the analytical outcomes from the knowledge-based system above. This stage of the prospective study is the least formalized as it should be tailored to the specific needs and customs of stakeholders. The common feature is a simple online questionnaire research, called “Delphi round 0”, where the users mark the most relevant items. Then the experts assess the results, review the data resources available in the knowledge base and determine the effort needed to gather the additional necessary information. The final selection of specific topics and a time horizon for the study is a result of a trade-off between the users’ needs expressed by aggregated ranks and resource requirements.

In the case of decision support systems (DSS) and recommenders, which have been selected as the exploratory focus area in [9], cf. also [13], the pre-delphi phase resulted in the identification of the following key subareas, listed according to their pre-delphi relevance scores (best first):

- recommenders for e-commerce (excluding banking and finance)
 - graphical (content-based) recommenders for multimedia
 - graphical (content-based) recommenders for 3D-e-commerce
- recommenders for security and commodity trading
- intelligent intermediary agents for negotiations, partner matching, e-commerce.

Then for each of the subareas selected, the experts retrieved the keywords, geographic, temporal and other characteristics to be used when surveying the information in the knowledge base and acquiring additional data. A topical DSS-related ontology was created, which ended the preparatory phase of the study.

According to the scheme presented in Secs. 2 and 3, the subsequent knowledge acquisition, processing and analysis comprises the following five steps:

Step 1. A common discrete-time model of the IS in Poland (1) with 92 variables has been updated to include the recent economic data and input price trends. The model parameters have been re-calculated and verified using the Granger causality tests. After the model iteration for the blocks of variables with statistically relevant A -matrix (1) coefficients, or trend extrapolation for the remaining ones, we received the forecasts of GDP per capita, digital literacy indicator, unemployment, mobile technology penetration and other relevant trends. They all have been input to the new analysis instance (called DSS2025) in the knowledge base.

Step 2. The DSS-related ontology served to retrieve from the knowledge base the list of most relevant technologies, methods and models to be used in the DSS. This phase was based on an automatic webcrawler search in external information sources

(bibliographic and patent databases, smart web search for products and technologies [14]). It yielded a.o. the following results:

- GIS technologies able to evaluate or elicit geographic preferences within a specified area, capable of using advanced visualization techniques coupled with GPS,
- DSS endowed with cognitive features, making it possible to avoid the negative consequences of decisions made by an irrational decision maker etc.
- Mobile decision support technologies that can explore a Personal Preference Record available in the cloud.

Step 3. An online expert Delphi was performed to complement the data gathered so far: verify the causal relations in the models (1)-(2), identify the market trends and the external events that may influence the model parameters, such as expected legislation, political decisions and IPR impact. It included the elements of risk analysis as well: the respondents could identify barriers, opportunities, threats and challenges for the DSS production and use.

Step 4. The fusion of all information gathered so far has been the most crucial point in the overall study as the quality of output information influences directly the success chances of the users. A dedicated information processing method has been elaborated as a merger of discrete-event simulation (2) and the well-known trend-impact analysis [8]. All events have been originally regarded as 0-1-valued functions on the predefined time interval Ω , in our case $\Omega=[2012,2025]$. Then, depending on the event character, the event variables have been converted to continuous functions, using the Delphi responses and technological trends elicited from the bibliographic and patent analysis. They have been interpreted either as fuzzy (partial) events or cumulative distribution of the event occurrence probability, or as a combination of both. The influence of events generated from (2) on the outputs from (1) was performed by multiplying the latter by the event variables. The whole process was repeated iteratively until convergence was reached. Finally, the following salient trends concerning future development of decision support systems (DSS) until 2025 have been obtained (cf. Tab.3 below).

Table 3. Selected DSS-related trends until 2025

| Technological/consumer trend description | Present value | Value in 2020 | Value in 2025 |
|--|---------------|---------------|---------------|
| Penetration of the mobile DSS in OECD countries (in % of mobile phone users) | 3% | 60% | 80% |
| Seeking advice from an online medical DSS (in % of Internet users, EU) | 18% | 45% | 70% |
| Share of financial investment decisions made with DSS (in %, OECD) | 65% | 80% | 95% |
| DSS as a component of social media | 5% | 60% | 95% |
| Share of DSS using multicriteria analysis (except simple scoring) | 35% | 50% | 80% |

Source: Delphi and causal trend analysis in [9],[13]

Step 5. The quantitative results have been described in form of recommendations to the software analysts and researchers, specifically, we claim that:

- the role and degree of sophistication of OR-based methods applied in DSS will grow, especially multicriteria optimization, uncertainty models and management,
- the class of decision problems regarded as numerically non-tractable will shrink,
- DSS will converge with search engines and intelligent data mining agents; the latter will complete missing data that might help in solving decision problems supplied in the client's queries.

The above presented example shows how the results produced by the information system allow the users to characterize the evolution of selected technologies as well as rank and position the companies, countries or regions under review in terms of development of a particular technological area. For instance, the above presented future characteristics of the DSS market are helpful in assessing the competitiveness of DSS suppliers. The systematic specification of key technologies, focus areas, methods and models within the above presented approach allows us, in turn, to perform targeted research on trends and scenarios concerning the objects selected in an efficient way. Its results can then be used to re-examine technological evolution principles in the knowledge base, thus forming a consistent interactive and adaptive model.

A real-life example of a typical industrial user of the above research results is an investment fund focused on 3D and virtual reality technologies for modern e-commerce applications. When making investment decisions the management of the fund takes into account IT development trends and rankings of prospective products, technologies and markets elicited during a foresight exercise. At a higher decision-making level, dynamic ranking methods [12] are used to rank corporate development policies, which concern the sector, size, or regional preferences regarding targeted markets or portfolio structure [8]. At the lower decision-making level, rankings are implemented as investment rules, by assigning funds to specific undertakings. Each assignment is a function of time and of external logical variables, the latter representing the changes in higher-level ranking and the states of external socio-economic (including financial markets) situation and research environments. The trends and scenarios generated by the system (1)-(2) are used to establish future investment rankings in an adaptive way. In particular, based on the feasible scenarios found at moment t_0 , the management of the fund can calculate corresponding future rankings for $t = t_0 + 1, t_0 + 2, \dots, t_0 + k$. This makes it possible to input into fund allocation planning more knowledge coming from systematically updated foresight results in the form of future recommendations and real options. Apart from rationalizing the time order, financing IT and market expansion projects, investment policy ranking may also help to determine organizational structure, future human resource and budgetary needs, and actions to be taken when priorities change as a result of external events [12].

5 Conclusions

The main user group of the recommendations and future prospects produced by the system described in Secs. 2 and 3 are policy makers at different levels as well as R&D and educational institutions on the key directions of development, and on the demand for IT professionals. Moreover, the global trends concerning the economy- and consumer-behavior-driven diffusion of IT innovations and technological characteristics of the IS evolution can provide clues to innovative IT companies seeking technological recommendations and advice concerning R&D priorities. This information will also be useful for corporations from different sectors that invest in IT. Foresight outcomes can

situate the IT project portfolio management and fund allocation strategies within the macroeconomic, political, technological and research environment by providing recommendations, relative importance rankings, trends and scenarios [12]. More objective and quantifiable future technological and economic characteristics will enable us to define more appropriate policy goals and measures to implement. The quantitative characteristics of the technological evolution can provide direct clues to IT providers, specifically DSS, as regards future demand for their products.

Comparing quantitative and descriptive approaches to elicit technological trends and build scenarios, it is noticeable that the approach of extracting evolution rules prior to a scenario analysis proved especially useful in the case of converging information societies, as exemplified by the IS/IT trends in the EU States which acceded in 2004 and 2007 [11]. The progress of the cohesion process seven years after the IS foresight results in these countries were published [11] confirms the adequacy of the modelling methods developed by a good coherence of forecasts and their ex-post verification. Furthermore, the architecture of the knowledge base designed originally as a foresight support system, and the hints resulting from its applications can contribute to the mainstream of knowledge science development (cf. e.g. [5]), as an example of an information system based on participatory modelling by experts and stakeholders.

The foresight results provided in [11] can be used as arguments supporting our claim that trustworthy Information Society trends, scenarios and rankings for the following 12-15 years can be derived using the methods applied in [9], some of them described in this paper. Such results can have useful applications in planning corporate strategic IT development. In particular, the investigation of selected technology areas within the IT foresight project [9] can provide constructive recommendations to companies interested in the development of DSS for e-commerce applications.

Another type of result that can be derived from the information gathered in the knowledge base is the model of adaptation of new software versions to the changes in consumer behaviour and technological progress. The product line evolution model described by (1)-(2) together with the research on the evolution of the consumers' preferences can provide clues to IT providers about future demand. They can also give R&D and educational institutions some idea of the most likely directions of development and demand for IT professionals, exploring the interdependence of the corresponding components of the model (1)-(2). Moreover, the general IS evolution model presented in Secs.2 and 3 can be useful for the analysis of global socio-economic trends that influence the development of the digital economy in a country or region, thus useful to the policy makers at national or regional levels.

Acknowledgment. The results presented in this paper have been obtained during the research project "Scenarios and Development Trends of Selected Information Society Technologies until 2025" financed by the ERDF within the Innovative Economy Operational Program 2006-2013, Contract No. WND-POIG.01.01.01-00-021/09.

References

1. Antoniou, M.R., Stenning, V.: The Information Society as a Complex System. *Journal of Universal Computer Science* 6(3), 272-288 (2000)
2. Bañuls, V.A., Salmeron, J.L.: Scope and Design Issues in Foresight Support Systems. *International Journal of Foresight and Innovation Policy* 7(4), 338-351 (2011)

3. Górecki, H., Skulimowski, A.M.J.: A Joint Consideration of Multiple Reference Points in Multicriteria Decision Making. *Found. Control Engrg.* 11(2), 81–94 (1986)
4. Lane, D., Pumain, D., van der Leeuw, S.E., West, G. (eds.): *Complexity Perspectives in Innovation and Social Change*. Springer Science+Business Media B.V (2009)
5. Nakamori, Y. (ed.): *Knowledge Science. Modelling the Knowledge Creation Process*. CRC Press, Boca Raton (2012)
6. Olivera, N.L., Proto, A.N., Ausloos, M.: Information Society: Modeling A Complex System With Scarce Data. *Proc. of the V Meeting on Dynamics of Social and Economic Systems* 6, 443–460 (2011) (arXiv:1201.1547)
7. Ramadge, P.J., Wohnam, W.M.: Supervisory control of a class of discrete event processes. *SIAM J. Control* 25(1), 206–230 (1987)
8. Salo, A., Mild, P., Pentikäinen, T.: Exploring causal relationships in an innovation program with Robust Portfolio Modeling. *Tech. Forecasting Soc. Change* 73, 1028–1044 (2006)
9. *Scenarios and Development Trends of Selected Information Society Technologies until 2025*, Progress & Business Foundation, Kraków (2012), <http://www.ict.foresight.pl>
10. Skulimowski, A.M.J.: Optimal Control of a Class of Asynchronous Discrete-Event Systems. In: *Proceedings of the 11th IFAC World Congress, Automatic Control in the Service of Mankind*, Tallinn, Estonia, vol. 3, pp. 489–495. Pergamon Press, London (1991)
11. Skulimowski, A.M.J.: Framing New Member States and Candidate Countries Information Society Insights. In: Compañó, R., Pascu, C. (eds.) *Prospects For a Knowledge-Based Society in the New Members States and Candidate Countries*, Publishing House of the Romanian Academy, pp. 9–51 (2006)
12. Skulimowski, A.M.J.: Application of dynamic rankings to portfolio selection. In: Soares, J.O., Pina, J.P., Catalão-Lopes, M. (eds.) *New Developments in Financial Modelling*, pp. 196–212. CSP Cambridge Scholars Publishing, Newcastle (2008)
13. Skulimowski, A.M.J.: Future Trends of Intelligent Decision Support Systems and Models. In: Park, J.J., Yang, L.T., Lee, C. (eds.) *FutureTech 2011, Part I. CCIS*, vol. 184, pp. 11–20. Springer, Heidelberg (2011)
14. Skulimowski, A.M.J.: Fusion of Expert Information on Future Technological Trends and Scenarios. In: Kunifujii, S., Tang, X.J., Theeramunkong, T. (eds.): *Proc. of the 6th International Conference on Knowledge, Information and Creativity Support Systems*, Beijing, China, October 22–24 (KICSS 2011), pp. 10–20. JAIST Press, Beijing (2011)
15. Skulimowski, A.M.J.: Discovering Complex System Dynamics with Intelligent Data Retrieval Tools. In: Zhang, Y., Zhou, Z.-H., Zhang, C., Li, Y. (eds.) *IScIDE 2011. LNCS*, vol. 7202, pp. 614–626. Springer, Heidelberg (2012)
16. Skulimowski, A.M.J.: Anticipatory Network Models of Multicriteria Decision-Making Processes. *Int. J. Systems Sci.* 44 (2012), doi: 10.1080/00207721.2012.670308
17. Skulimowski, A.M.J., Schmid, B.F.: Redundancy-free description of partitioned complex systems. *Mathematical and Computer Modelling* 16(10), 71–92 (1992)
18. Sudár, E., Peto, D., Gábor, A.: Modeling the Penetration of the Information Society Paradigm. In: Wimmer, M.A. (ed.) *KMGov 2004. LNCS (LNAI)*, vol. 3035, pp. 201–209. Springer, Heidelberg (2004)
19. Tadeusiewicz, R.: A need of scientific reflection on the information society development. In: Bliźniuk, G., Nowak, J.S. (eds.) *Information Society 2005*, pp. 11–38. PTI (2005)
20. Walden, P., Carlsson, C., Liu, S.: Industry foresight with intelligent agents. *Human Systems Management* 19(3), 169–180 (2000)

How Many Answers Are Enough? Optimal Number of Answers for Q&A Sites

Pnina Fichman

Indiana University, Bloomington, USA
fichman@indiana.edu

Abstract. With the proliferation of the social web questions about information quality and optimization attract the attention of IS scholars. Question-answering (QA) sites, such as Yahoo!Answers, have the potential to produce good answers, but at the same time not all answers are good and not all QA sites are alike. When organizations design and plan for the integration of question answering services on their sites, identification of good answers and process optimization become critical. Arguing that ‘given enough answers all questions are answered successfully,’ this paper identifies the optimal number of posts that generate high quality answers. Based on content analysis of Yahoo! Answers’ informational questions (n=174) and their answers (n=1,023), the study found that seven answers per question are ‘enough’ to provide a good answer.

Keywords: Q&A sites, CQA, Optimization, Web 2.0, Information Quality.

1 Introduction

One of the goals of IS research is to find ways “to increase the timeliness, accuracy, and completeness of information at a minimum of costs---economic, cognitive, political, social, affective, and physical. At the heart of IS research, then, is a complex optimization problem” [1, p. 13]. As such, it is not surprising that information quality is a focus of much IS research (e.g., [2, 3]). The challenges associated with information quality, both conceptual and practical, are not new but, with the adoption of information technology, organizations are faced with additional challenges. This complexity further intensifies as organizations try to leverage the potential of the social Web, mass collaboration, and free and open source software (FOSS). Thus, scholars have examined the potential and challenges associated with organizations using FOSS [4], and the potential of cost reduction and innovation by means of crowdsourcing [5, 6, 7]. With these complexities in mind, optimization is still one of the core challenges in IS research and practice.

The proliferation of mass information production on the social Web (e.g., Wikipedia, Yahoo! Answers) raises many questions about the reliability of user-created content. Empirical support for the potential of crowdsourcing, for example, is provided by consistent reports that the quality of Wikipedia entries is as good as those in traditional encyclopedias (e.g., [8]) and that the Wikipedia Reference Desk is as good as reference services provided by libraries [9]. At the same time, concerns about the rise

of a culture of mediocrity fostering a cult of amateurs [10] where everything is miscellaneous [11].

Scholars try to explain why and how the participatory nature of Web 2.0 provides an infrastructure for achieving high quality knowledge production. A popular explanation suggests that it is the “wisdom of crowds” [12]. Another explanation comes in the form of Linus’ Law: “given enough eyeballs, all bugs are shallow” [13]. However, ‘enough’ may mean some but not too many, as the cliché argues that too many cooks can spoil the broth. In the context of FOSS, this rationale leads to Brooks’ Law [14], which claims that increasing the number of developers in a project can introduce inherent coordination complexity that may hinder group performance.

Like FOSS and Wikipedia, Question Answering (QA) sites draw on mass collaboration and user participation. They are based on the idea that “everyone knows something” [15, p. A01], and that through collaborative knowledge production, users can provide answers to questions that are being asked. The growing popularity of these sites in terms of the number of users, questions, and answers is fascinating. For example, Yahoo! Answers is among the most frequently consulted reference sites, second only to Wikipedia. By the end of 2009, Yahoo! Answers boasted 1 billion questions and answers, 179 million users, and over 200 million visitors worldwide [16]. If QA sites provide high quality information while reducing costs, then organizations can utilize similar mechanisms for mass user participation to improve their services; specifically, information intermediation services can leverage this potential through crowdsourcing their services. While the potential benefit of QA sites providing quality information has been empirically documented (e.g., [9, 17]), great caution must be advised because information quality varies between answers and across different QA sites (e.g., [17, 18]). Assuming that answer multiplication (many answers can be posted for a single question) is beneficial, a few questions should be addressed: Is there an optimal number of answers/answerers per question that leads to the best outcomes in terms of information quality? How many answers per question are ‘enough’ to produce a good answer? Is it possible that after an optimal number of answers have been posted, the added value of additional answers is minimal or may even hinder answer quality? Is it likewise possible that many answers are still not ‘enough’ and that, regardless of their number, answer quality is low? This optimization issue is critical when organizations design and plan for the integration of QA services on their sites. The goal of this study is to answer the question: How many answers does it take to provide a good answer on QA sites?

Content analysis of informational questions ($n=174$) and their multiple answers ($n=1,023$) from Yahoo! Answers was performed at two levels of analysis. Findings reveal that answer multiplication significantly improves answer quality and that, in order to provide a reliable answer, seven answers per question are ‘enough’.

2 Background

2.1 Question Answering Sites

There is a growing body of research on QA sites that focuses on information retrieval, information seeking behavior and use, information intermediation, and the social

dynamics of these online communities (e.g., [9, 17, 19, 20, 21, 22, 23, 24, 25]). In their respective areas researchers argue that QA sites change information creation, dissemination, intermediation, retrieval, seeking, and use. Most of these studies have focused on Yahoo! Answers; some have examined other QA sites, such as Answerbag, [20, 26, 27], Wikipedia Reference Desk [25], and Naver [28], while several have examined and compared multiple QA sites in their studies [9, 17, 24, 29]. One common motivation for research in these domains follows the assumption that there is added value in achieving a better understanding of the question answering process (information intermediation, information reuse, and information retrieval) and outcomes (information quality in terms of answer quality). Information retrieval researchers, for example, assume that the crowd produces information that should be archived and reused because of its quality. This assumption justifies their efforts to identify high quality answers, incorporating social aspects such as user reputation and user ranking of answers.

The popular assumption about the potential benefits of collaborative question answering should not be taken for granted; it has been challenged because empirical findings show that information quality varies not only among answers but also across different QA sites [9, 17, 29]. Despite the fact that all QA sites exploit similar collaborative mechanisms to enable mass user participation, answer quality varies amongst them [9, 17]. Therefore, it is still unclear whether the crowd improves answer quality at all. The present study tries to address this gap, aiming to determine whether answer multiplication improves information quality.

This study tries then to uncover the conditions that can produce good answers, mainly by identifying the optimal number of answers per question and by asking how many answers are needed to yield a reliable answer. This optimization effort is critical for the future design and implementations of next-generation QA systems. It is also useful to examine whether common FOSS laws are applicable to QA sites. Specifically, assuming that bugs resemble questions in that they need to be identified or asked, processed or answered, and solved by the crowd, the study aims specifically to test whether Linus' Law is relevant here. In the context of QA sites, Linus' Law can be stated as follows: 'given enough answers, all questions are answered successfully.'

Posing this statement in the context of QA leads to three main challenges; the meaning of 'all' questions, the meaning of being 'answered', and finally, the meaning of being 'answered successfully'. First, not 'all' questions that are posted on QA sites are answered (e.g., [9, 17]). Response rates range between 16%-96% per QA site (rates of no response ranges between 4%-84%) [9]. Second, different types of questions might call for different answers and might require different evaluation criteria (e.g. [21, 23]); thus considering 'all' questions becomes a complex task. Third, what constitutes an answer is yet another challenge. For example, simply responding to a question with a random statement does not seem to be an answer to the question. Moreover, an answer could be 1) an individual post; 2) all posts for one particular question; 3) an answer that collaboratively co-authored by more than one user; or 4) a chosen "best answer". Fourth, having an answer does not guarantee that the answer is of high quality (even when it is chosen as "best answer"). Thus, that a question has been successfully answered could mean different things to different scholars and the

challenge of determining what makes a good answer becomes apparent. There are multiple points of view as to what constitutes a good answer and how answer quality should be evaluated, which include user rankings of “best answers”, user reputation, user satisfaction, and content criteria of answers, such as answer accuracy and completeness [9]. Taking into account these challenges, this paper aims to identify what constitutes ‘enough’ in the context of question answering.

2.2 Information Quality and Answer Quality

Scholarly publications about information quality are mostly practical and less theoretical [2]. Likewise in reference research, where answer quality has been assessed, “a lack of attention [has been] given to theory” [30, p. 3]. Information quality has attracted much research attention across many scholarly communities; among them are scholars engaged in information systems (IS) research and library and information science (LIS). In IS research for example, information quality is one of the key factors that affect IS success [e.g., 31] and in LIS information quality was examined, for example, through the lenses of information seeking behavior research [e.g., 32] and reference research [33, 34].

Information quality is a multidimensional construct with many different definitions and attributes [35]; it has been the center of attention well before the introduction of the social web. With the increase interest in the quality of user-generated information, the concept continues to capture scholarly attention. Two different approaches to information quality seem to be prominent [35]. The first is subjective, focusing on users’ judgment of information credibility [22, 23, 32] or user perceptions of fitness of use [35], and the other focuses on objective measures of an information artifact (a website or an answer), such as accuracy and completeness [25, 33, 36]. The utilization of the second approach to information quality in the study of answer quality on Q&A sites can be useful, but poses certain challenges, as the artifact is dynamic and multifaceted. Under the objective approach, high quality answers were determined based on content analysis of the answers [17, 19, 25, 26, 37]. Scholars that analyzed the content of answers have found, for example, that better answers are longer [17, 37, 38], or include references to external sources [26]. Interestingly, question category, answer accuracy and completeness, and length of answer are significant predictors of answer quality, whereas asker’s and answerer’s authority and reputation are not [37].

Prior research on answer quality on QA sites has primarily assessed quality using the subjective approach and was based on user rankings of “best answers”. However, user rankings are problematic because they provide a subjective measure of answer quality. Poston and Speier [39] argue that, “rating validity, [which] describes the degree to which the rating reflects the intrinsic quality of the content ... may be low for a variety of reasons ... [it is] inherently subjective and voluntarily provided, resulting in mismatch between the true quality of the content and the rating given ... [and] those submitting ratings may manipulate ratings...” [39, p. 223]. For example, in 29.8% of cases where users chose “best answers” in Yahoo! Answers, their selections were based on socio-emotional criteria rather than on the content or utility of the answer [22]. Another method to identify answer quality is by tracking user reputation.

This method is based on the assumption that certain users are more likely to provide better answers than others [40]. Examples of this approach include the ranking of authoritative responders using link analysis [41, 42]. Other ranking methods measure users' reputations based on their activity levels (e.g., [37, 40, 43]), their focus on one subject area [44], their credibility (authority), or the number of "best answers" they have previously posted [37, 45]. However, this approach is also problematic because even users with good reputations do not always provide high quality answers.

Measuring answer reliability by focusing on answer accuracy and completeness is another common approach in quality assessment on QA sites [9, 17, 25, 37, 46]. Under this objective approach, high quality answers have been determined based on content analysis of the answers [17, 19, 25, 26, 37]. Scholars analyzing the content of answers have found, for example, that better answers are longer [17, 37, 39], and include references to external sources [26]. Researchers argue that different questions warrant a different type of answers and that not all measures of quality should apply to all answers [9, 23, 47]. They differentiate between conversational and informational questions [47], subjective and objective questions [48], or navigational, informational, transactional, and social questions [49].

QA sites are socio-technical systems where many different facilitating conditions can affect the quality of answers that can be found on them [18, 20, 24]. This led to the development of theoretical frameworks that integrate both the objective and subjective approach to determine QA sites effectiveness [e.g., 18, 37, 40]; answer quality is an important component in all of these frameworks. The present study examines the relationships between two components in the social reference model [18]: number of users (counting their posts) and answer quality (using reliability measures) under the objective approach to information quality. Given the lack of attention in these frameworks to the issue of optimization, the present study focuses on optimization. It also tests their underlying assumption that the crowd, by providing multiple answers to a given question, answers questions well enough.

3 Method

3.1 Data Collection

Data were harvested from Yahoo! Answers, using a Perl program that was set up to collect on July 10th 2008 the most recent question per category over a 24-hour period at a random minute of every hour (24 points of time), and, 24 hours later, to collect all of the relevant answers. Using this method, a random sample of 585 transactions was collected. Yahoo! Answers was chosen because it is the most popular QA site [50]. A transaction includes a question and a whole answer. A whole answer includes any number of answers; most of the time the whole answer includes multiple answers. Transactions that include a "best answer" are resolved transactions. A question becomes a Resolved Question when a Best Answer is chosen. After a question becomes Resolved it stays in Yahoo! Answers and is available for searching and browsing. The Best Answer remains open to receive comments and ratings from the community.

Very few questions received a high number of answers and many of the questions received either very few or no answers at all. The number of answers per question varied between zero and 60 ($M=6.12$, $SD=7.52$), and while 70% of the questions received more than one answer, 16% of questions received no answers.

The amount of time that passed between the initial posting of questions and the posting of first answers ranged between 0:01 and 34:33 hours ($M=1:01$, $SD=3:41$). The amount of time that passed before last answers (in the data set) were posted ranged between 0:03 and 57:31 hours ($M=6:24$, $SD=9:52$).

Because different credibility criteria for informational and conversational questions were reported by users of Yahoo! Answers [23], the aim in this study was to focus further analysis only on informational questions. The questions were categorized, either as conversational or informational, using the following definitions [47, p. 759]:

Informational questions are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. An example: *What's the difference between Burma and Myanmar?*

Conversational questions are asked with the intent of stimulating discussion. They may be aimed at getting opinions, or they may be acts of self-expression. An example: *Do you drink Coke or Pepsi?*

First, the transactions were sorted into one of the two categories by one coder and later, 30% of the data was sorted into categories by a second coder to assure inter-coder reliability and strengthen the validity of the study results [51]. Inter-coder reliability was determined using simple agreement, also called percent agreement, which is based on the percentage of all codes that a pair of coders agreed on [52]. Inter-coder reliability resulted in 90% agreement, which is high; as a rule of thumb, co-efficiency of .90 or greater would be acceptable to all [51, 52].

Seventy-three percent of the questions were informational ($n=422$), and the rest were conversational ($n=163$). Conversational transactions had significantly more answers per question ($M=9.74$, $SD=10.73$) compared with informational transactions ($M=4.92$, $SD=6.39$).

Two samples that complement each other were drawn from the informational questions data set for manual content analysis. The first sample was a purposeful sample of resolved transactions (questions with “best answers”) in line with prior research tendencies to include only resolved transactions; it included 74 transactions. However, because the resolved transaction sample included only 17% of the 422 informational transactions and only 12% of the entire data set of 585 transactions, a second with 100 random transactions was collected. The random sample included transactions with questions but no answers ($n=19$), transactions with answers but no “best answer” ($n=65$), and resolved transactions with “best answers” ($n=16$). The average number of answers per question was higher in the resolved transactions sample ($M=7.81$, $SD=8.87$) and lower in the random sample ($M=4.45$, $SD=5.54$) than it was in the entire informational questions data set.

3.2 Data Analysis

To determine answer reliability level, a content analysis of 174 transactions and 1,197 posts from Yahoo! Answers (174 questions and 1,023 answers in two samples), was conducted [53]. Content analysis of answers is a widely used method to evaluate answer quality on QA sites (e.g., [17, 19, 26, 37]) because it enables the evaluation of answer quality based on quantifying the presence or absence of quality measures (codes) in the answer.

Analysis was conducted at two levels: 1) transaction (n=174) – whole answer; 2) question-answer pair (n=1,023) – first answer and “best answer”. Quality rates for the whole answer, the first answer, and the “best answer” were coded. The first answer is the first answer posted in response to a question. The “best answer” is the answer chosen as “best answer” by the asker or by a community vote. The “best answer” encompasses feedback about the fit between question and answer and a selection of one answer as being of good quality. Coders do not define the “best answer” but determine the quality of the individual answer that was chosen, in some transactions, as the “best answer.” Frequencies of reliability codes were aggregated for: whole answer, first answer, and “best answer”, using three reliability measures: accuracy, completeness and verifiability. These three measures have been widely used in prior research on QA sites (e.g., [9, 37, 46]), and have been frequently used by Yahoo! Answers’ users in their information credibility judgments [23]. Accuracy, completeness, and verifiability are of particular importance in judging the credibility of answers to informational questions [23]:

1. **Accuracy** of an answer refers to a correct response.
2. **Completeness** of an answer refers to an answer that is thorough, provides enough information, and answers all parts of a multi-part question.
3. **Verifiability** of an answer refers to an answer that provides a link or a reference to another source where the information can be found.

Using these codes, two coders each coded the entire data set, assigning a value (yes/no) for each code (accuracy, completeness, and verifiability) to the transactions and the question-answer pairs. Coders were graduate students studying library and information science at a Midwestern university. They were instructed to determine the accuracy, completeness and verifiability of the answers “on the surface” [52] and based on their best judgment to verify information with external sources. Inter-coder reliability between the two iterations of coding of all the transactions was determined using simple agreement and Cohen’s Kappa. Inter-coder reliability was 92%, which is high [51, 52]; Cohen’s Kappa was .84, which means that there was almost perfect agreement between the two coders [54].

First, frequency tables were created for each of the two samples (the random sample and the resolved transactions sample), tallying the presence of codes (yes values) for the whole answer, first answer, and “best answer”. Then, the percentages of codes per answer were marked and statistical analysis using SPSS 17.0 was done. Later, the location of “best answer” was marked and cumulative quality rates were examined; data about the users, those who asked and answered questions in the random sample, were tallied as well.

4 Findings

The results of the analysis of both samples at two levels of analysis (transaction and question-answer pair) are presented in Table 1. The two samples were compared on all three reliability measures (accuracy, completeness, and verifiability), for all three types of answers (first answer, whole answer, and “best answer”). The differences between the samples were not statistically significant; the level of accuracy, completeness and verifiability for “best answer” and first answer did not differ between the samples, but the level of completeness for the whole answer was higher in the resolved transactions. In both samples, the whole answer and the “best answer” are significantly better than the first answer, and the “best answer” shows the highest levels of accuracy and completeness (Tables 1). Verifiability levels are very low for both samples (Table 1).

Table 1. Rates on single variables

| | | Accurate | Complete | Verifiable |
|-------------------------------------|-----------------------|----------|----------|------------|
| Resolved transactions (n=74) | | | | |
| Best Answers | % | 95% | 96% | 16% |
| | # | 70 | 71 | 12 |
| Whole Answers | % | 89% | 96% | 18% |
| | # | 66 | 71 | 13 |
| First Answers | % | 68% | 62% | 9% |
| | # | 50 | 46 | 7 |
| Random sample (n=100) | | | | |
| Best Answers | 16 resolved questions | 88% | 94% | 13% |
| | 81 answered questions | 17% | 18% | 2% |
| | 100 posted questions | 14% | 15% | 2% |
| | # | 14 | 15 | 2 |
| Whole Answers | 81 answered questions | 89% | 84% | 14% |
| | 100 posted questions | 72% | 68% | 11% |
| | # | 72 | 68 | 11 |
| First Answers | 81 answered questions | 78% | 57% | 11% |
| | 100 posted questions | 63% | 46% | 9% |
| | # | 63 | 46 | 9 |

The findings indicate that answer multiplication significantly increases answer quality in terms of accuracy and completeness. Information reliability for whole answers is higher than for first answers.

While in both samples there are small differences in verifiability levels between the first answer, “best answer”, and whole answer, these differences are not statistically significant. Completeness levels in both samples, and accuracy levels in the resolved sample, are significantly different. The level of answer reliability in the resolved sample, differs between the first answers, “best answers”, and whole answers in terms of accuracy ($\chi^2=29.91$, $df=2$) and completeness ($\chi^2=59.36$, $df=2$), but not in terms of verifiability ($\chi^2=5.82$, $df=2$). Follow-up pair-wise comparisons show that: 1) first answers are significantly less accurate than whole answers ($\chi^2=13.06$, $df=1$) or “best answers” ($\chi^2=24.18$, $df=1$); 2) “best answers” and whole answers are equally accurate; 3) first answers are significantly less complete than whole answers ($\chi^2=34.84$, $df=1$) and “best answers” ($\chi^2=34.84$, $df=1$); and 4) “best answers” and whole answers are equally complete.

In the random sample, the level of completeness is significantly different between the first answers, “best answers”, and whole answers ($\chi^2=43.17$, $df=2$). Follow-up pair-wise comparisons show that first answers are significantly less complete than whole answers ($\chi^2=17.53$, $df=1$) and “best answers” ($\chi^2=37.01$, $df=1$).

As the results above indicate, answer accuracy and completeness improve for whole answers in comparison with first answers. Still, it is unclear how many answers are required to reach a quality of answer that is good enough. Looking at the average number of answers per transaction can provide one solution to this question. Accuracy and completeness levels improve with an average of 7.81 answers per question (resolved sample), and level of accuracy improves with 4.45 answers per question (random sample). In other words, 5 answers are enough for an increase in levels of completeness from first answers to whole answers, while 8 answers are enough for an increase in levels of completeness and accuracy. However, because the quality of “best answers” is equal to that of whole answers on all measures in both samples, it is

Table 2. “Best answer” location in transaction

| “Best Answer” Location in Answer | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th |
|--|------------|------------|------------|------------|------------|------------|------------|
| Resolved Sample | | | | | | | |
| Number of Answers (n=74) | 21 | 12 | 9 | 4 | 4 | 4 | 6 |
| Cumulative Percent of Resolved Transactions (n=74) | 28% | 44% | 56% | 62% | 67% | 73% | 81% |
| Random Sample (2010 data) | | | | | | | |
| “Best Answer” Location in Answer (By Vote; n=48) | 26 | 11 | 2 | 3 | 2 | 1 | 3 |
| Cumulative Percent of Resolved Transactions (By Vote; n=48) | 54% | 77% | 81% | 88% | 92% | 94% | 100% |
| “Best Answer” Location in Answer (By Asker; n=25) | 10 | 3 | 3 | 1 | 1 | 1 | 1 |
| Cumulative Percent of Resolved Transactions (By Asker; n=25) | 40% | 52% | 64% | 68% | 72% | 76% | 80% |

possible that, if a “best answer” has been selected, fewer answers than the average are enough. It is likewise possible that the quality can improve beyond the levels of whole answers or “best answers” for resolved transactions with more answers than the average. Further analysis moving beyond the simple averaging of numbers was done next, looking into the optimization challenge.

First, using data from the resolved sample, the cumulative percentage and frequency of “best answer” location have been noted (Table 2). In most of the transactions the “best answer” was one of the first three answers (59.45%), and in many cases the “best answer” was the first answer (28%). Only by the seventh answer did 80% of the resolved transactions have a “best answer”.

Next, the random sample was revisited in November 2010, when a higher percentage of the transactions had been resolved ($n=73$ in 2010 compared with $n=17$ at the time of original data collection); these “best answers” were chosen either by the asker ($n=25$) or by a community vote ($n=48$). The cumulative percentage and frequency showing the location of “best answers” have been noted (Table 2) for these transactions. All “best answers” chosen by the asker, and 80% of “best answers” chosen by a community vote were selected from the first 7 answers posted.

The findings from both samples indicate that 7 answers are enough to achieve an accuracy rate of 95%. Further, while the findings indicate that it takes at least seven answers to achieve high quality information in the form of a “best answer”, high quality answers appear also before and after the “best answer”. The number of accurate and correct answers that are posted before the “best answer” strongly correlates with the “best answer” location ($r=.86$) and with the number of answers ($r=.90$). In fact, the total number of accurate and complete answers correlates with the total number of answers ($r=.87$) and there is a strong Pearson product-moment correlation coefficient between the number of answers and the location of the “best answer” in the transactions ($r=.93$). Moreover, it was evident that only two answers were needed to achieve accuracy in more than 80% of the resolved transactions.

To sum up, the findings indicate that: 1) answer multiplication, with or without the “best answer” feature, results in higher answer quality than the first answer; 2) for over 80% of the transactions, seven answers are enough to get good answers.

5 Discussion

Questioning the almost unquestioned belief that “given enough eyeballs, all bugs are shallow,” could be done through multiple lenses; philosophically, statistically, or empirically. This study treats empirically a variation of this belief and postulates that ‘given enough answers, all questions are answered successfully’. It defines success as measured by answer accuracy and completeness. While probabilistically the argument that ‘given enough answers, all questions are answered successfully’ is a sound argument, in reality it can take an endless number of users (or answers) and a long time. The study determines that the number of answers that are needed for 80% of the questions to be answered successfully is the optimum. Under these conditions, the findings show that seven answers are enough to yield good answers for over 80% the questions, that answer quality improved with additional answers, and that there was no evidence of a number of answers, after which additional answers reduce quality.

As such, the findings do not provide evidence to support Brooks' Law, which is a competing theory used in the context of FOSS. Brooks' Law argues that there might be an optimum number of people involved in one successful project, but that adding more participants after reaching that point may hinder performance [14]. According to Brooks' Law, Linus' Law may not hold up ad infinitum. In the context of QA sites, it could mean that answers produced by large groups may be of a lower quality than those of smaller groups. In fact, resembling the inverse relationship between incentives to contribute and group size [55], despite the fact the Yahoo! Answers is the most popular QA site (it has more users and questions than other QA sites), and that a question on Yahoo! Answer gets more answers on average than the Wikipedia Reference Desk, for example, answer quality on Yahoo! Answers was lower than that of the Wikipedia Reference Desk [9]. The findings of the present study, focusing only on Yahoo! Answers, show that additional answers only improved answer quality. Thus it can be concluded that the findings are in alignment with prior FOSS research that found evidence in support of Linus's Law, rather than Brooks' Law [4]. Schweik et al. [4] found that adding more developers improves the chances that the FOSS project will be successful, but they caution that the correlation between size and success does not necessarily mean that bigger groups produce better software, and that it is likewise possible that successful projects attract more contributors. Schweik et al. [4] claimed that size is only one factor that may contribute to the success of the FOSS project and, because they did not conduct multivariate analysis, it is possible that other factors could well serve as competing explanations. Similarly, Meneely and Williams [48] found empirical support for Linus' Law, but they also found some support for Brooks' Law and argue that their findings "do not necessarily negate Linus' Law ... [but that] they are a legitimate opposing force" [56, p. 460].

In addition to the support for Linus' Law the findings of the study also show that answer multiplication leads to quality improvement (better accuracy and completeness of answers). There was no evidence in prior research that supports this assumption, yet, under the assumption that the crowd can produce good answers, scholars have made efforts to describe and understand the process of social question-answering [18, 20, 24]. In order to provide empirical support for this assumption, data was analyzed and compared at two levels of analysis; this comparison is essential when looking into the benefits of answer multiplication. While quality improvement was evident for all three variables (verifiability, accuracy, and completeness), it was only statistically significant for two of them, accuracy and completeness. Answer verifiability was very low at both levels of analysis, for whole answers, "best answers", and first answers, echoing prior research. For example, and only one out of ten messages on the Wikipedia Reference Desk includes references [25]. This dimension of answer quality is perceived to be very important by Yahoo! Answers' users [23]. Low verifiability levels not only correspond with, but also give rise to, concerns about information quality and the lack of authority on the social Web.

It is important to note that significant improvement in answer quality, in terms of accuracy and completeness, was also associated with "best answers." While "best answers" are individual answers, their quality was equal to that of whole answers. This may be due to the selection process of "best answers", which involves an

additional step of information processing that includes feedback about the quality of the answer in light of the question. In fact, in two third of the transactions with “best answers” the choice of a “best answer” was a result of a community vote (48 “best answers” have been chosen by the community and 25 have been chosen by the asker).

6 Conclusion

This study shows that answer multiplication and user’s choice of best answers on QA sites significantly improves answer quality in terms of accuracy and completeness when compared with an individual (first) answer. However, the collaborative process did not produce a significant change in the (low) levels of answer verifiability. In support of the argument that given enough answers all questions are answered successfully, the findings reveal that it takes seven answers to achieve a 95% accuracy level for resolved transactions and, on average, seven answers to achieve an 89% accuracy level for transactions that have not yet been resolved.

References

- [1] Briggs, R.O., Nunamaker, J., Sprague, R.: Introduction to the special section: social aspects of sociotechnical systems. *Journal of Management Information Systems* 27(1), 13–16 (2010)
- [2] Ballou, D., Madnick, S., Wang, R.: Special section: assuring information quality. *Journal of Management Information & Systems* 20(3), 9–11 (2003)
- [3] Nelson, R.R., Todd, P.A., Wixom, B.: Antecedents of information and system quality: an empirical examination within the context of data warehousing. *Journal of Management Information Systems* 21(4), 199–235 (2005)
- [4] Schweik, C.M., English, R.C., Kisting, M., Haire, S.: Brooks’ versus Linus’ law: an empirical test of open source projects. In: *Proceedings of the 2008 International Conference on Digital Government Research*, pp. 423–424. ACM, Montreal (2008)
- [5] Howe, J.: The rise of crowdsourcing. *Wired* 14(6) (2006), <http://www.wired.com/wired/archive/14.06/crowds.html>
- [6] Howe, J.: *Crowdsourcing*. Crown Publishing Group, New York (2008)
- [7] Leimeister, J.M., Huber, M., Bretschneider, U., Krcmar, H.: Leveraging crowdsourcing: activation-supporting components for IT-based ideas competition. *Journal of Management Information Systems* 26(1), 197–224 (2009)
- [8] Giles, J.: Internet encyclopedias go head to head. *Nature* 438, 900–901 (2005), <http://www.nature.com/news/2005/051212/full/438900a.html>
- [9] Fichman, P.: A comparative assessment of answer quality on four question answering sites. *Journal of Information Science* 37(5), 476–486 (2011)
- [10] Keen, E.: *The Cult of the Amateur: How Today’s Internet is Killing Our Culture*. Doubleday/Currency, New York (2008)
- [11] Weinberger, D.: *Everything is Miscellaneous: The Power of the New Digital Disorder*. Henry Holt & Co., New York (2007)
- [12] Surowiecki, J.: *The Wisdom of Crowds*. Anchor Books, New York (2004)
- [13] Raymond, E.: The cathedral and the bazaar. *Knowledge, Technology & Policy* 12(3), 23–49 (1999)

- [14] Brooks Jr., F.P.: *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley Publishing Company, Reading (1975)
- [15] Noguchi, Y.: Web searches go low-tech: you ask, a person answers. *Washington Post*, p. A01 (2006), <http://www.washingtonpost.com/wp-dyn/content/article/2006/08/15/AR2006081501142.htm>
- [16] Yahoo Answers hits 200 million visitors worldwide! *Yahoo Answers Blog*. Yahoo (2009), <http://yanswersblog.com/index.php/archives/2009/12/14/yahoo-answers-hits-200-million-visitors-worldwide/>
- [17] Harper, F.M., Raban, D., Rafaeli, S., Konstan, J.: Predictors of answer quality in online Q&A sites. In: *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 865–874. ACM, New York (2008)
- [18] Shachaf, P.: Social reference: a unifying theory. *Library & Information Science Research* 32(1), 66–76 (2010)
- [19] Agichtein, E., Castillo, C., Donato, D., Gionides, A., Mishne, G.: Finding high-quality content in social media. In: *Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 183–194. ACM, Palo Alto (2008)
- [20] Gazan, R.: Microcollaborations in a social Q&A community. *Information Processing & Management* 46(6), 693–702 (2010)
- [21] Harper, F.M., Weinberg, J., Logie, J., Konstan, J.: Question types in social Q&A sites. *First Monday* 15(7) (2010), <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2913/2571>
- [22] Kim, S., Oh, S.: Users' relevance criteria for evaluating answers in social Q&A site. *Journal of the American Society for Information Science and Technology* 60(4), 716–727 (2009)
- [23] Kim, S.: Questioners' credibility judgments of answers in a social question and answer site. *Information Research* 15(2), paper 432 (2010), <http://InformationR.net/ir/15-2/paper432.html>
- [24] Rosenbaum, H., Shachaf, P.: A structuration approach to online communities of practice: the case of Q&A communities. *Journal of the American Society for Information Science and Technology* 61(9), 1933–1944 (2010)
- [25] Shachaf, P.: The paradox of expertise: is the Wikipedia Reference Desk as good as your library? *Journal of Documentation* 65(6), 977–996 (2009)
- [26] Gazan, R.: Specialists and synthesists in a question answering community. In: *Proceedings of the American Society for Information Science & Technology Annual Meeting, ASIST, Austin*, pp. 1–10 (2006)
- [27] Gazan, R.: Seekers, sloths and social reference: Homework questions submitted to a question-answering community. *New Review of Hypermedia & Multimedia* 13(2), 239–248 (2007)
- [28] Nam, K.K., Ackerman, M.S., Adamic, L.A.: Questions in, knowledge in?: a study of Naver's question answering community. In: *Proceedings of the 27th International Conference on Human Factors in Computing Systems*, pp. 779–788. ACM, Boston (2009)
- [29] O'Neill, N.: Chacha, Yahoo!, and Amazon. *Searcher* 15(4), 7–11 (2007)
- [30] Saxton, M.L., Richardson, J.: *Understanding Reference Transactions: Transforming an Art into a Science*. Academic Press, San Diego (2002)
- [31] DeLone, W.H., McLean, E.: The DeLone and McLean model of information systems success: a ten-year update. *Journal of Management Information Systems* 19(4), 9–30 (2003)

- [32] Rieh, S.: Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology* 53(2), 145–161 (2002)
- [33] Fallis, D.: On verifying the accuracy of information: Philosophical perspectives. *Library Trends* 52(3), 463–487 (2004)
- [34] Frické, M., Fallis, D.: Indicators of accuracy for answers to ready reference questions on the Internet. *Journal of the American Society for Information Science and Technology* 55(3), 238–245 (2004)
- [35] Arazy, O., Nov, O., Patterson, R., Yeo, L.: Information quality in Wikipedia: the effects of group composition and task conflict. *Journal of Management Information Systems* 27(4), 71–98 (2011)
- [36] Stvilia, B., Twidale, M.D., Smith, L.C., Gasser, L.: Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology* 59(6), 983–1001 (2008)
- [37] Blooma, J.M., Chua, A.Y.K., Goh, D.: A predictive framework for retrieving the best answer. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*. ACM, Fortaleza (2008)
- [38] Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.: Knowledge sharing and Yahoo! Answers: Everyone knows something. In: *Proceedings of the International World Wide Web Conference*, ACM, Beijing (2008)
- [39] Poston, R., Speier, C.: Effective use of knowledge management systems: A process model of content ratings and credibility indicators. *MIS Quarterly* 29(2), 221–244 (2005)
- [40] Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums: The case of Yahoo! Answers. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 866–874. ACM, Las Vegas (2009)
- [41] Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 919–922. ACM, New York (2007a)
- [42] Jurczyk, P., Agichtein, E.: Hits on question answer portals: exploration of link analysis for author ranking. In: *Annual ACM Conference on Research and Development in Information Retrieval*, pp. 845–846. ACM, Amsterdam (2007b)
- [43] Chen, W., Zeng, Q., Wenyin, L.: A user reputation model for a user-interactive question answering system. In: *Proceedings of the Second International Conference on Semantics, Knowledge, and Grid*, pp. 40–45. IEEE Computer Society, Washington D.C (2006)
- [44] Adamic, L.A., Wei, X., et al.: Individual focus and knowledge contribution. *First Monday* 5(3) (2010)
- [45] Dom, B., Paranjpe, D.: A Bayesian technique for estimating the credibility of question answerers. *Proceedings of the Society for Industrial and Applied Mathematics (SIAM)*, pp. 399–409. SIAM, Atlanta (2008), http://www.siam.org/proceedings/datamining/2008/dm08_36_Dom.pdf
- [46] Ong, C., Day, M., Hsu, M.: The measurement of user satisfaction with question answering systems. *Information & Management* 46(7), 397–403 (2009)
- [47] Harper, F.M., Moy, D., Konstan, J.: Facts or friends?: Distinguishing informational and conversational questions in social Q&A sites. In: *Conference on Human Factors in Computing Systems*, pp. 759–768. ACM, Boston (2009)

- [48] Li, B., Liu, Y., Ram, A., Garcia, E.V., Agichtein, E.: Exploring question subjectivity prediction in community QA. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 735–736. ACM, Singapore (2009)
- [49] Liu, Y., Li, S., Cao, Y., et al.: Understanding and summarizing answers in community-based question answering services. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 497–504. ACL, Manchester (2008)
- [50] Hitwise. U.S. visits to question and answer websites increased 118 percent year-over-year. Hitwise, New York (March 19, 2008), <http://www.hitwise.com/news/us200803.html>
- [51] Neuendorf, K.: *The Content Analysis Guidebook*. Sage, Thousand Oaks (2002)
- [52] Lombard, M., Snyder-Duch, J., Bracken, C.: Content analysis in mass communication: assessment and reporting of intercoder reliability. *Human Communication Research* 28(4), 587–604 (2002)
- [53] Krippendorff, K.: *Content Analysis: An Introduction to its Methodology*, 2nd edn. Sage, Thousand Oaks (2004)
- [54] Landis, J.R., Koch, G.: An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33(2), 363–374 (1977)
- [55] Zhang, X., Feng, Z.: Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101(4), 1601–1615 (2011)
- [56] Meneely, A., Williams, L.: Secure open source collaboration: an empirical study of Linus' Law. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 453–462. ACM, New York (2009)

Analysis and Support of Lifestyle via Emotions Using Social Media

Ward van Breda, Jan Treur, and Arlette van Wissen

VU University Amsterdam, Agent Systems Research Group, The Netherlands
{w.r.j.van.breda, j.treur, a.van.wissen}@vu.nl

Abstract. Using recent insights from Cognitive, Affective and Social Neuroscience, this paper addresses how affective states in social interactions can be used through social media to analyze and support lifestyle behaviour. A computational model is provided that integrates both mechanisms for the impact of one's emotions on behaviour, and for the impact of emotions of others on one's own emotion. The model is used to assess the state of a user with regard to a lifestyle goal (such as exercising frequently), based on extracted information of emotions exchanged in social interaction. Support is provided by proposing ways to affect these social interactions, which will indirectly influence the impact of the emotions of others. An ambient intelligent system based on this model has been implemented for the social medium Twitter.

Keywords: Social media, emotion, lifestyle support, Ambient Intelligence.

1 Introduction

In order to achieve a healthy lifestyle people have to adopt behaviours such as eating healthy or increasing their level of physical activity. However, developing and maintaining these healthy behaviours is for many a major challenge; e.g., [31]. Every New Year's Eve the same resolutions are made when it comes to going to the gym more frequently or starting that diet. One of the problems is that decisions for these behaviours are mainly made in an unconscious manner, and are closely related to emotions associated to them; e.g. [9]. These associated emotions are in turn influenced by social interactions and social norms. It turns out that people are more successful in complying with their healthy lifestyle when they receive positive social support than when they don't receive this support; e.g. [42]. Social influence is found to be fundamental for the maintenance of various health behaviors, such as smoking cessation, self-management for chronic patients, and weight loss; for example, see [11, 27, 41]. Also, the number of friends or family members and the frequency of social contact were found to be positively associated with higher physical activity levels [37]. As observed in [16], these findings indicate that social support can serve as a leveraging mechanism for achieving and maintaining a level of commitment and motivation for performing healthy behaviour (see also [19]).

A first step to improve adherence to healthy behaviours may be the formation of social networks or communities that provide possibilities for mutual monitoring and support. Technologies that enable sharing information about health-related activities

can provide powerful support for healthy habits; e.g., [7]. Online social media such as Facebook or Twitter allow people to give short real-time update messages concerning their current activities and emotions [4, 17]. This direct visibility makes these platforms very suitable for seeking expertise and social support [16].

However, potentially more consistent behaviour change can be achieved when a social network is provided with an automated intelligent system for monitoring, assessing and supporting the network, for example, of the type as addressed in Ambient Intelligence; e.g. [8]. The main characteristics of Ambient Intelligence applications are that (1) they are nonintrusive, hidden in a person's environment, (2) they are able to monitor using sensor systems, and (3) their interventions have a high extent of sensitivity to the context and state of the person. This perspective can be useful if the focus is not on individual persons but, as is the focus of this paper, on the social interactions they have. This means that both monitoring and interventions (mainly) target the interactions between people, and not individuals themselves. The aim such systems is to use analysis and promotion of social interactions between people in a network to stimulate healthy behaviour. To our knowledge, no earlier work exists that includes this approach of using support interventions in social media that are targeted on the interactions in a social network.

In the approach presented in this paper the first step is to obtain information by monitoring the social interactions. This information is then used to create model-based assessments about the states and processes of the individuals. Based on these assessments, interventions are generated. The underlying computational model for this approach makes use of recent insights from Cognitive, Affective and Social Neuroscience on emotion-related valuing in decision making and on mirroring of emotions; e.g., [21, 25, 38; 14, 30, 33]. The computational model addresses (1) how emotions affect behaviours, and (2) how these emotions can be affected by emotions of others. This computational model is used as a basis for monitoring, analysis and affecting the emotions expressed in the interactions in a social network.

In the remainder of this paper, Section 2 provides a brief overview of the background knowledge from Cognitive, Affective and Social Neuroscience on the interaction between emotions and behaviour, and on social contagion of emotions. Section 3 presents the computational model. In Section 4 the design of the ambient system is described. Section 5 addresses how the model was used to implement a support system using Twitter. In Section 6 an illustrative scenario is described which is generated by the system. Finally, Section 7 is a discussion.

2 The Role of Emotions in Contagion of Behaviour

Recent developments in Cognitive Neuroscience have revealed mechanisms behind the generation and contagion of affective states, and the roles they play in other mental processes. In this section they will be briefly reviewed.

In order to choose to perform a behaviour, usually unconsciously a number of options are considered, one of which is chosen based on some valuation. In a process of valuing these options, the predicted associated emotions are an important element. In recent neurological literature this has been studied in relation to a notion of value as represented in the amygdala; see, for example [24, 25, 35]. A role of amygdala

activation has been found in various processes involving emotional aspects. Usually emotional responses are triggered by stimuli for which a predictive association is made of a rewarding or aversive consequence, given the context, which may include certain goals. Feeling these emotions represents a way of experiencing the value of such a prediction, and to which extent it is positive or negative: this can be considered prior valuation of the option. Similarly, feelings of satisfaction are an important element of retrospective valuation of what is experienced after behaviour has been chosen. This idea of value is the basis of current work on the neural basis of decision making processes and economic choice in neuroeconomics; e.g., [24, 38].

In a social context emotions can play an even more important role, as their occurrence in one person can easily affect the same emotion in another person. This applies both to the emotions related to prior valuations of behavioural options, and to feelings of satisfaction about effects after behaviour was chosen. Therefore the idea of emotion-related valuing can be combined with recent neurological findings on the *mirroring function* of certain neurons (e.g., [14, 30, 33]). Mirror neurons are neurons that, in the context of the neural circuits in which they are embedded, show both a function to prepare for certain actions or bodily changes and a function to represent states of other persons. They are active not only when a person intends to perform a specific action or body change, but also when the person observes somebody else intending or performing this action or body change. Indeed, if states of others are affecting some of the person's own states that at the same time are connected via neural circuits to states that are crucial for the person's own feelings and actions, then this provides an effective mechanism for persons to fundamentally affect each other's actions and feelings. As mirror neurons make that some specific sensory input (an observed person) directly links to the relevant own preparation states, mirroring is a process that fully integrates mirror neuron activation states in ongoing internal processes. This includes expressing emotions in facial expressions, or in language expressions as, for example, used in social media; e.g., [1, 3]. Thus it is assumed that the mechanism of mirroring provides a neural basis for emotion contagion via text-based interaction.

Given the general principles described above, the mirroring function provides a mechanism by which emotions felt in different individuals about a considered behaviour mutually affect each other, and, assuming emotion-related valuing, consequently affect the valuation and choices for behaviour options.

3 The Computational Model

Following the conclusions from Section 2, to stimulate healthy behaviour, it will be fundamental to stimulate positive emotions and to minimize negative emotions people have related to these behaviours. In order to be able to reason about the emotional state of humans and about the consequences of social interactions on this state, a computational model has been developed. The model described here incorporates the elements as discussed in Section 2 in an abstracted manner. An overview is shown in Figure 1. In a particular context certain preparations for behaviour options *BO* are triggered. These options affect associated levels of certain emotions, which are also affected by emotions of others (by mirroring). These emotion levels in turn affect (as a way of valuing) the level of preparation for the option. In the remainder of this

section a more detailed specification of the computational model is discussed. Note that the model is described for a given, positive emotion. In the implementation the model has been applied as well in an independent manner to a negative emotion.

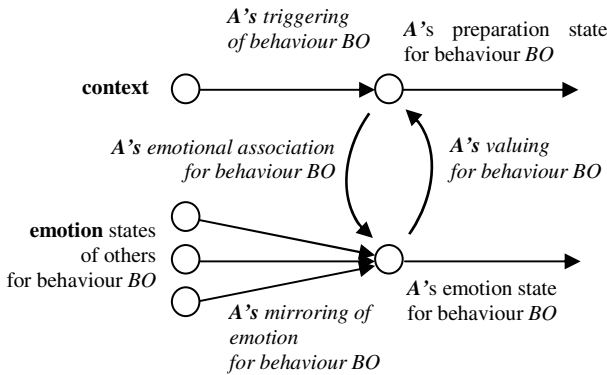


Fig. 1. Overview of the computational model

Some elements of the emotion contagion model used in this chapter were adopted from [10]. A number of aspects of the proposed computational model are distinguished that play a role in the contagion, varying from aspects related to an agent sending the emotion, an agent receiving the emotion, and the connection between sender and receiver; see Table 1.

Table 1. Parameters for personal and social characteristics

| | |
|--|---------------|
| level of A 's (internal) emotion | q_A |
| A 's emotion expression characteristic | ϵ_A |
| A 's openness for received emotion | δ_A |
| the strength of the connection from sender B to receiver A | α_{BA} |

Here q_A denotes the internal emotion level of person A , which is distinguished from the emotion level xq_A expressed by the person. In the expression of an emotion the personal characteristic ϵ_A plays a role; this depends on how introvert or extravert, expressive, and/or active or energetic person A is. These aspects correspond to the personality trait extraversion and sensation seeking. It represents to what extent a person transforms internal emotion into external expression: an introvert person will induce a weaker contagion of an emotion than an extraverted person. In the model it is assumed that the level of expressed emotion is proportional to the internal emotion level with proportion factor ϵ_A : it is taken as the product of the person's internal emotion level q_A and the person's expressiveness ϵ_A . The characteristic δ_A indicates the degree of susceptibility of A . This represents to what extent a receiver allows the emotions received from others to affect the own emotion, and how flexible/persistent the person is emotionally. The characteristic α_{BA} depends on the type and intensity of the contact between the two persons (e.g., distance vs. attachment). This α_{BA} may be related to a combination of more specific aspects such as the directness of the emotion

contagion, and the relations between sender and receiver. The stronger the connection, the higher α_{BA} and the more contagion will take place.

The characteristics shown in Table 1 have been formalized using numbers in the interval $[0, 1]$. In addition, the parameter γ_{BA} is used to represent the strength by which an emotion is received internally by A from sender B , modelled as:

$$\gamma_{BA} = \varepsilon_B \alpha_{BA} \delta_A \tag{1}$$

The model works as follows: if γ_{BA} is 0, there will be no contagion, if it is 1, there will be a maximum strength of contagion. If γ_{BA} is not 0, there will be contagion and the higher the value, the more contagion will take place. In a way γ_{BA} expresses the energy level with which an emotion is being expressed, transferred and received. The overall strength by which emotions from all others members in a social network N are received by A , indicated by γ_A , is defined as

$$\gamma_A = \sum_{B \in N \setminus \{A\}} \gamma_{BA} \tag{2}$$

For any $A \in N$, let

$$q_A^* = \sum_{B \in N \setminus \{A\}} w_{BA} q_B \tag{3}$$

be the weighted combined emotion levels from the other agents, where the weights w_{BA} are taken proportional to $\varepsilon_B \alpha_{BA} \delta_A$ and normalised, as defined by

$$w_{BA} = \varepsilon_B \alpha_{BA} \delta_A / \sum_{C \in N \setminus \{A\}} \varepsilon_C \alpha_{CA} \delta_A = \varepsilon_B \alpha_{BA} / \sum_{C \in N \setminus \{A\}} \varepsilon_C \alpha_{CA} \tag{4}$$

The set of differential equations for emotion contagion is then described by

$$\Delta q_A(t+\Delta t) = q_A(t) + \gamma_A [q_A^*(t) - q_A(t)] \Delta t \tag{5}$$

for all $A \in N$.

For each behaviour option BO the effect of emotion contagion for BO on A 's emotion level for BO is combined with the person's own emotion association to BO . Moreover, the emotion level and the context together affect A 's preparation for BO .

$$q_{A,BO}(t+\Delta t) = q_{A,BO}(t) + \gamma_{A,BO} [q_{A,BO}^*(t) + \omega_{A,BO}^{(2)} p_{A,BO}(t) - q_{A,BO}(t)] \Delta t \tag{6}$$

$$p_{A,BO}(t+\Delta t) = p_{A,BO}(t) + [\omega_{A,CA,BO}^{(1)} c_A(t) + \omega_{A,BO}^{(3)} q_{A,BO}(t) - p_{A,BO}(t)] \Delta t \tag{7}$$

Here the ω 's are the strengths of the relevant associations:

- $\omega_{A,CA,BO}^{(1)}$ strength of association from context c of A to $p_{A,BO}$
- $\omega_{A,BO}^{(2)}$ strength of association from $p_{A,BO}$ to emotion $q_{A,BO}$
- $\omega_{A,BO}^{(3)}$ strength of association from emotion $q_{A,BO}$ to $p_{A,BO}$

Note that for a positive emotion $\omega_{A,BO}^{(3)}$ is a positive number; for a negative emotion it is a negative number.

The parameters ε_A , α_{AB} , δ_B in this model define through the contagion strength γ_{BA} the impact of members of the network on each other through their connections. These contagion strengths could be kept constant over time, as was done in [10]. However, the model can describe more realistic scenarios if the connections also show development over time; some may become stronger, some weaker. The model here expands the existing model as described in [10] by considering a dynamical social network evolving over time based on such dynamic network characteristics. In particular, the connection

strength and expressiveness parameters α_{AB} and ε_A are assumed to change over time depending on the different social interactions and emotion levels. More specifically, it is assumed that connection strength will increase with more frequent communication, and with higher emotional intensity of the messages that are exchanged. These assumptions are based on literature studies that show that connection ('tie') strength in both online and offline social networks is influenced by (i) interaction frequency, (ii) emotional intensity of content, and (iii) emotional support and closeness [13, 18, 12]. The number of questions asked in a message also conveys information about connection strength. For example, in [22, 23] it was found that many participants' questions in online social networks were answered by friends they rated as close, and that closeness of a friendship was a motivator to answer questions. Asking questions also identifies the asker as someone who could use some guidance or support.

For the connection strength the impact is modeled as communication frequency value times average intensity over the last time unit, which can be considered as the overall intensity transferred per time unit:

$$\langle \text{connection impact} \rangle = \text{frequency value} \cdot \text{average intensity}$$

The frequency value is a function f of the frequency with values in the interval $[0, 1]$, and average intensity is a weighted sum of the emotion levels expressed in the messages and a value in $[0, 1]$ for the average number of questions in the messages during one time unit:

$$\begin{aligned} \text{average intensity} = & w_1 \cdot \text{average positive emotion level} + \\ & w_2 \cdot \text{average negative emotion level} + w_3 \cdot \text{average question level} \end{aligned}$$

So the connection impact becomes

$$\langle \text{connection impact} \rangle = f(\text{frequency}) * [w_1 \cdot \text{average positive emotion level} + w_2 \cdot \text{average negative emotion level} + w_3 \cdot \text{average question level}] \quad (8)$$

Given this impact, the dynamics of the connection strength is modelled as follows:

$$\alpha_{AB}(t+\Delta t) = \alpha_{AB}(t) + \eta [\langle \text{connection impact} \rangle - \alpha_{AB}(t)] \Delta t \quad (9)$$

Here η is an adaptation speed parameter. In the scenarios discussed in Section 6 the function f used is defined in a linear manner by:

$$f(V) = \begin{array}{ll} V/4 & \text{when } V \leq 4 \\ 1 & \text{when } V > 4 \end{array}$$

Adaptation of expressiveness over time is modelled in a similar manner as

$$\varepsilon_A(t+\Delta t) = \varepsilon_A(t) + \eta [\langle \text{expressivity impact} \rangle - \varepsilon_A(t)] \Delta t \quad (10)$$

with $\langle \text{expressivity impact} \rangle$ the expressed emotion level divided by the internal emotion level averaged (over the last time unit) for the positive and negative emotions:

$$\langle \text{expressivity impact} \rangle = \text{average of } w_4 xq\text{-pos}_A(t) / q\text{-pos}_A(t) + w_5 xq\text{-neg}_A(t) / q\text{-neg}_A(t) \quad (11)$$

Here $q\text{-pos}$ and $q\text{-neg}$ denote the emotion levels for positive and negative emotion, respectively, and $xq\text{-pos}$ is the expressed emotion level for the positive emotion and $xq\text{-neg}$ for the negative emotion.

4 System Design

The model presented in Section 3 can be incorporated by an intelligent system that is able to use the model to assess whether intervention is desirable. The system has been designed in an agent-based manner, using an *ambient agent model* adopted from [5], which also was used in [10]. The computational model is integrated as a *domain model* by embedding it within different components of the agent model. By incorporating a domain model within an agent model, the agent gets an understanding of the processes of its surrounding environment, which is a solid basis for knowledgeable human-aware intelligent behaviour. Three different ways to integrate domain models within an agent model are considered; see Fig. 2. The solid arrows indicate information exchange between processes (data flow) and the dotted arrows the integration process of the domain model within the agent model. In the current paper the adaptation model is left out of consideration. In future development this can be added, for example, to achieve on the fly parameter tuning.

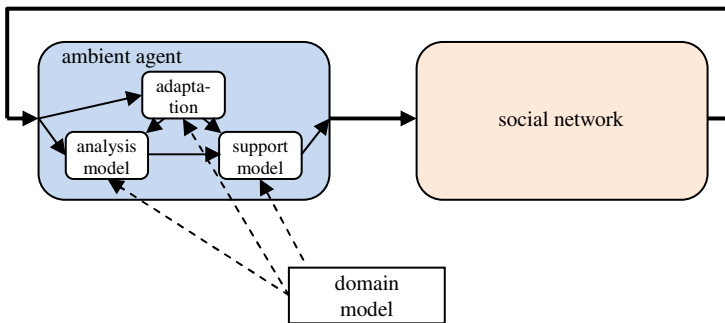


Fig. 2. Ambient support agent architecture and models used

Analysis Model. This model component is used for the analysis and assessment of human states and processes. Some aspects of the states and processes related to human functioning can be directly observed, but often many relevant aspects only can be indirectly derived from such observation information. For such derivations it is useful to have a domain model integrated within the analysis model, which can be used to estimate states of the human (for past, present and future time points). Given these estimated human states, an assessment is made to verify whether there is any reason to consider support. As an example, the level of positive emotions of a given person in the network may be assessed as too low.

The computational model described in Section 3 was integrated in the Analysis component of the ambient agent to analyze the (expected) dynamics of the humans in the network. Concepts needed in such a model for an ambient agent concern estimations of the human's states at different points in time; these estimations are described by the ambient agent's observations and beliefs; in addition an assessment of the (expected) emotion state of persons in the network is needed. Such an assessment expresses that the emotion level of a specific person at some (future) time point is expected to be too low, compared to some desired norm (EN). The emotion levels are estimated by simulation of the computational model described in Section 3, in

particular the dynamic specifications (6) and (7). The parameters used are adaptive over time, based on direct observations of the *expressed emotion levels* (x_{q_A}) and on the text and sentiment analysis, using (9) and (10) from Section 3.

Support Model. To generate intervention actions that fit with the assessment results of the analysis, a support model is used. An often-used approach in network interventions is to identify opinion leaders (i.e., persons who strongly influence the opinions, attitudes, beliefs, motivations, and behaviours of others [34]) to advocate healthy behaviours [40]. Opinion leaders can be identified for example by looking at the in-degree centrality, which is determined by the amount of incoming ties of the nodes in the network. Alternatively, rewiring the network by changing the strength of links could be an even more powerful, though challenging, approach. Intervention techniques based on personal characteristics also seem promising, as these characteristics determine a person's behaviour. Considering both network structure and attributes of persons in the network simultaneously seems therefore a lucrative approach to intervention, although little work has been done to explore how such an approach can harness the positive effects of social influence [40]. As an example of such an intervention, a person in the network with strong expressivity for positive emotions may be encouraged to communicate more to one of his friends with negative emotions, in order to stimulate more positive contagion.

The support model in this work uses a heuristic approach. The agent will reason about the proper actions that should be undertaken to achieve and maintain that the emotion levels of the members of the network are optimal for the desired behaviour. For example, it uses knowledge expressing that in case one's positive emotion level (e.g., pride or joy) is expected to be lower than a given norm, other members of the network are to be detected that can play a crucial role in a negative or positive sense. Next, it could be suggested to these members to increase or decrease their impact on the target network member. Three main intervention types are distinguished to optimize the emotion levels of a person A:

- interventions to change the ε_B
- interventions to change the α_{BA}
- interventions to change the δ_A

In the current scenario, the agent could ask a person with a positive emotion level to be more expressive (increase the person's expressiveness ε_B) or to interact more frequently with the target (increase the connection strength α_{BA}). As another example, if the target member A has many negative friends, he or she can be nudged to decrease his/her openness δ_A , ensuring a decrease of negative emotion contagion strengths.¹ A heuristic that is applied for positive emotion levels, viewed from the receiver's perspective and the sender's perspective, respectively, is the following:

- *Affect incoming contagion*
 - let for a given network member A with low expected emotion level the members connected to A with lower emotion levels have less impact on A, and

¹ In order to establish what is the most preferred intervention method, an estimate of the feasibility of these methods can be made.

- let the members connected to A with higher emotion levels have more impact on A
- *Affect outgoing contagion*
 - let a given network member A with high expected emotion level, have more impact on the other connected members, and
 - let a given network member A with low expected emotion level, have less impact on the other connected members

Members with ‘higher’ or ‘lower’ emotion levels can be defined as the members with the absolute highest or lowest emotion level, but also as members with emotion levels above or under certain thresholds. In general, two (low and high) *emotion thresholds* are assumed for this. For a network member with emotions levels under the low threshold, his or her impact on other members can be inhibited by (encouragement for) decreasing the person’s expressiveness, or by decreasing the connection strengths from this person to the other members. For negative emotion levels opposite heuristics are used.

5 Implementation of the System

This section describes some details of the implementation made.

The Twitter Environment. The system has been implemented in the Twitter context. Several studies indicate that Twitter can be very useful for purposes of monitoring and obtaining information relating to people’s health behaviour. For example, in [36] Twitter was identified as a way to gather and exchange important real-time health data. Also, as noted in [16], ‘Twitter presents an interesting yet underexplored tool for health-promoting activities’, and ‘Twitter might be a promising platform for leveraging social support to motivate health behaviour change’. Furthermore, the number of Twitter users is estimated at 75 million, making it one of the most popular social media [32].

In order to make sure that the ambient intelligent system is able to analyse the streams of communication between the participants of the experiment correctly, there are certain rules that need to be followed and explained to the users. First, the users need to add the support agent to their follower list. The support agent represents the ambient system on Twitter and applies the intervention actions by generating protected tweet messages, which are private messages intended only for the receiver(s), when needed; see [39]: About Public and Protected Tweets. After a user added the support agent to his or her follower list, the support agent adds the user to its follower list, making protected tweeting between them possible. Secondly, the users always have to add a specific hashtag to identify participation in a specific support domain, which, for the scenarios discussed in this paper, was #vusupport. Twitter enables the use of hashtags for the purpose of categorization and filtering; cf. [39]: What Are Hashtags?. Thirdly, the users need to use mention reply signs, or @ signs, for identifying the recipient of the tweet message; cf. [39]: What Are Replies And Mentions. This way the ambient system can recognize the sender and recipient of the message, containing possible emotional content. Messages that do not contain a mention or reply sign are ignored by the ambient system. Lastly, the users need to communicate publicly to each other for the ambient system to be able to analyse the messages.

For capturing the public stream of tweets of the participants of the experiment, the Twitter public streaming API was used; cf. [39]: The Streaming APIs. The API provides low latency delivery of public tweets that satisfy the chosen keywords, or chosen hashtags, that identify the specific support domain.²

The ambient system was implemented using PHP as a programming language [29], MySQL as a database [26], and JSON as a data format for storing the values, relating specific users and specific time points in the database [15].

Analyzing Transferred Emotions in Tweets. In the implementation the positive and negative emotions of messages are calculated in an independent manner, which means that the internal emotion q , the expressed emotion xq , the level of expressiveness ε , the contagion strength γ , and weights w relating to these variables, are represented separately for positive and negative emotion. In the scenarios discussed in Section 6, the connection strength α and openness for received emotion δ have been given the same values for positive and negative emotion, yet this could also be determined independently for both emotion types (for example, if it was assumed that the level of openness for received emotion can be different for positive and negative emotion).

For the initial input values of the level of internal emotion q , a self-assessment of each user is required. In this implementation a simple questionnaire was used containing two continuous scales, one for positive emotion and one for negative emotion. The participant can select a point on the continuous scale reflecting his or her internal emotion. Furthermore, the scale is divided in 5 sub-intervals, ranging from 0 to 4, where 0 represents no emotion and 4 represents the highest level of that type of emotion. The labels give the participant a sense where on the scale the internal emotion of the participant is reflected. The internal emotion q is then calculated by dividing the input from the questionnaire by 4, generating a value in the interval $[0, 1]$. The current expressiveness of emotion ε is calculated, by taking the expressed emotion xq , divided by the internal emotion q . The internal emotion q can never be zero, as it is taken as a value at least in the middle of the lowest subinterval: $q \geq 0.1$. Then this current expressiveness is used in the dynamic specification (10) in Section 3 to adapt the value for ε maintained over time.

For the initial values of the openness for received emotions δ , the same type of scale and questionnaire as for the internal emotion is used. For measuring the expressed emotion xq , a sentiment analysis tool is used that classifies the tweets in type of sentiment. As is described in [4, 28], the empirical analysis of sentiment and mood, and opinion mining through user-generated textual data are currently developing research fields. As the intended target group of this research is Dutch, it was chosen to use a sentiment mining tool that is currently in development, specifically designed for the Dutch language. This sentiment mining tool uses a bag-of-word approach combined with a rule-based system, which analyzes the surrounding semantic context of the found sentiment words. When a tweet is analyzed and positive or negative words are identified, the semantic context is searched for negation words, strengthening words and weakening words. For the total of found word sequences, a scoring method is applied, which eventually leads to a final independent score for positive emotion and for negative emotion, scaled to values in the interval $[0, 1]$. These values can be

² More information about the different types of Twitter APIs can be found at [38].

directly compared with the internal emotion q , making it possible to calculate the level of expressiveness ε . Several alternative state-of-the-art approaches on how to extract sentiment from text have recently been discussed in the literature, for example in [2]. Because of the modular nature of the ambient system, one could easily replace a sentiment analysis tool with another solution that measures the sentiment of the message.

As discussed in Section 3, the connection strength α between each sender and receiver for each time point is calculated by measuring the frequency of communication, the amount of questions in the tweet, and the intensity of emotion in the tweet. The frequency of communication is measured by counting the amount of tweets the sender sends to the receiver for that time unit, after which it is mapped to a value in the interval $[0, 1]$, where 0 represents no communication and 1 represents a high level of communication frequency. In this implementation a linear function f was used as indicated in Section 3 in relation to (9) with 4 or more tweets per time unit being the highest level of frequency, mapped to 1 . Also discussed in Section 3, the amount of questions is considered to be another indicative variable the connection strength. The amount of questions found in the tweets is averaged over the total frequency of tweets for that time unit, after which the variable is scaled to a value in the interval $[0, 1]$ by using a similar scaling formula. Furthermore, the intensity of emotion contained in the messages is also considered to be indicative for the connection strength. To determine the intensity, the absolute values of both the positive and negative emotions in the tweets were summed up, and the resulting value was averaged over the frequency of tweets for that time point, and was mapped to a value in the interval $[0, 1]$. Following the calculation of the obtained frequency value, question value and emotion value in $[0, 1]$, the connection strength α between each specific sender and receiver is calculated dynamically as described by (9) in Section 3.

As input for the support model, the calculated (using (6) from Section 3) internal emotion value q , and the calculated connection strength α were used. With these values different types of interventions can be generated to specific communication partners, for example focusing their connection strengths. It is easy to extend this type of intervention for the support model, for example, by also involving the level of expressiveness or openness for received emotion.

6 An Illustrative Scenario

In this section the proposed approach is illustrated by an example scenario. This scenario simulates a period of 10 days with four persons: Alice, Bob, Carol, and Dave. In the first scenario, the system was used to only perform analysis and was not allowed to intervene. In a second scenario the system was allowed to intervene in order to establish positive emotions of the user related to a particular healthy behaviour. In the scenario, daily generated tweets were used as input for the states of the persons. The target behaviour is running (exercising) and it is assumed that the tweets are selected based on their content of running or jogging, by selecting tweets using the hashtags '#jogging' or '#running'.

In the scenarios, Alice, Bob, Carol and Dave have different personalities and different levels of experience with running. Alice has been running for many years and she is very excited about running. Bob is just starting out with jogging and needs some tips and tricks to become more experienced. Carol and Dave are both more experienced, but they have trouble keeping up their motivation and commitment to exercise regularly. Alice and Bob both have a high expressiveness ϵ , while Carol and Dave have more introvert characters, expressed by low ϵ 's. Openness δ is constant in this scenario: Alice and Dave are not so receptive and have a low openness (0.1, 0.25) while Bob and Carol are easily influenced (0.8, 0.9).

Table 2. Initial emotion levels

| Initial emotion levels | Alice | Bob | Carol | Dave |
|----------------------------|-------|-----|-------|------|
| Positive emotion q_{pos} | 0.9 | 0.4 | 0.1 | 0.2 |
| Negative emotion q_{neg} | 0.1 | 0.6 | 0.8 | 0.8 |

The initial emotion levels for the persons are given in Table 2. As can be seen in Table 2, initially Alice has very positive emotions concerning running, while Carol and Dave are more negative, and Bob is more neutral with a slight bias to the negative side. The four persons are connected in a social network with different strengths of connections between them, as shown in Fig. 3.

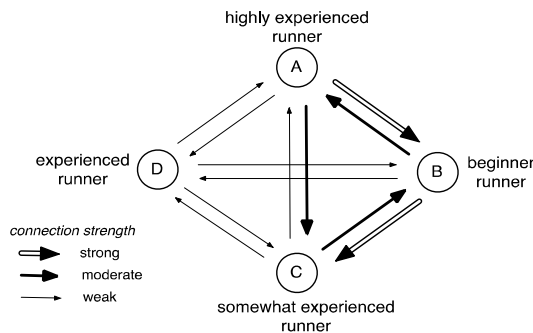


Fig. 3. Visualization of the social network containing Alice, Bob, Carol and Dave

The existence of a connection (a tie) between two persons represents personal contact in the Twitter environment (usage of the '@' sign to communicate with a specific person). A strong/moderate connection exists between Alice and Bob, and between Bob and Carol. Dave has weak connections to all of the others. For this scenario, a strong connection was taken as 3 or more tweets a day, a moderate connection as 1-2 tweets per day, and a weak connection as less than 1 tweet a day. Some examples of tweets that occurred in the scenario can be found in Table 3.³ The scenario resulted in the changes in emotions as shown in Fig. 4 for positive emotions and in Fig. 5 for negative

³ Note that these tweets are translations from the original Dutch tweets.

emotions. Here simulation using dynamic specification (6) from Section 3 was used as part of the analysis process, where the focus was on the social influence and for the sake of simplicity the connections to the preparation states have been left out of consideration. Moreover, (9) and (10) from Section 3 have been used in this simulation to make the network and its simulation adaptive to the real observed exchange of messages (in particular the connection strengths α and expressivity parameters ϵ).

Table 3. Example tweets

| sender | example tweet | sentiment analysis | |
|--------|--|--------------------|----------|
| | | positive | negative |
| Alice | @Bob yes man cmon you have to make it fun! you should go running in the park! :) letsdothis | 0.91 | 0 |
| | @Carol hows it goin girl!? all exited to go running? Where there's a will there's a way! ;) . . | 0.79 | 0 |
| Bob | @Dave yo! Tell me something, how's running treating you? Id like to hear from you! :) Are you a sportsman? | 0.74 | 0 |
| | @Dave Right! Could be a solution! Ill give it my best shot! | 0.72 | 0 |
| Carol | @Bob hey bob it's a bit tough, man. Calves often start to hurt and also my thighs... don't think i'm up for it | 0 | 0.68 |
| | @Bob, did it again, I went running, even stretched. Was ok. | 0.33 | 0 |
| Dave | @Alice hey. Have to admit...Running with a buddy is definitely better. | 0.75 | 0 |
| | @Bob I'm OK. Not very motivated though Just not very into it... | 0 | 0 |

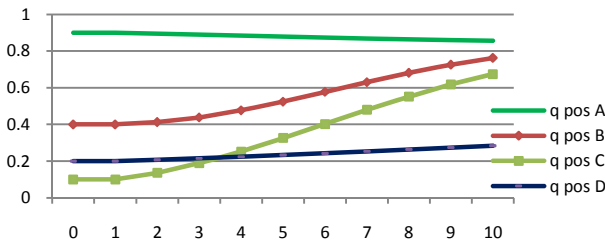


Fig. 4. Positive emotions over time without intervention

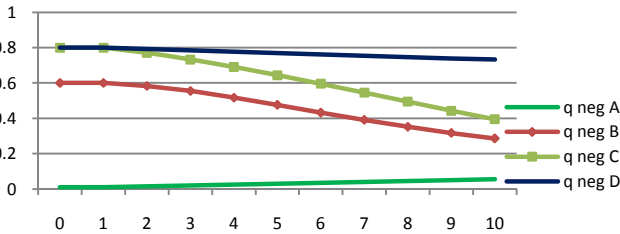


Fig. 5. Negative emotions over time without intervention

As can be seen in Fig. 4, the positive emotions of Bob and Carol increase to relatively high levels, but the level of positive emotions of Dave remain low. This can be explained by the fact that Dave has only weak connections to the others, whereas Bob has a strong connection from Alice which has a positive effect on his own emotions, and Carol has a strong connection from Bob which results in a positive effect for her. The opposite pattern is shown for the negative emotion levels in Fig. 5: they decrease for Bob and Carol. However, Dave’s negative emotions remain high, as he is not influenced much by the others (due to a low openness and few connections). As the scenario without intervention shows that the situation of Dave does not improve much, a next step is to see what happens when the ambient agent is allowed to intervene. Indeed in this second scenario the system detects that Dave needs some attention as his emotion level is estimated too low, and finds out that Alice and Bob are suitable network members to be encouraged to communicate more to Dave, as they have or develop high positive emotions and low negative emotion levels. In Figs 6 and 7 below it is shown that indeed this intervention is successful, as now also Dave’s positive emotion levels become higher and the negative emotions levels become lower.

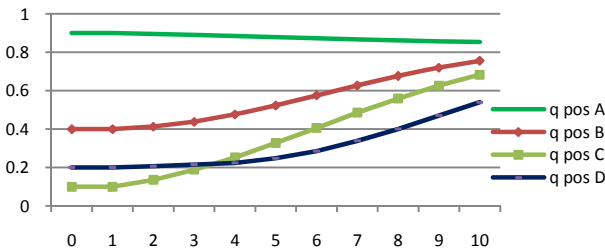


Fig. 6. Positive emotions over time with intervention

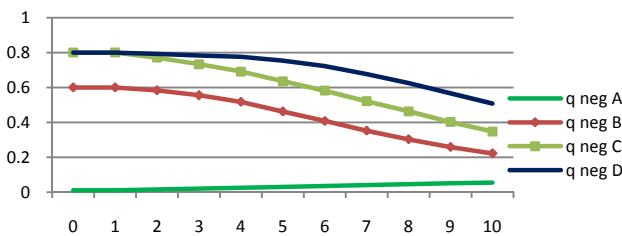


Fig. 7. Negative emotions over time with intervention

Fig. 8 shows the effect of the intervention on the social network evolution over time: the connection strength from Alice to Dave changes over time, based on dynamic specification (9) from Section 3 and the intervention after 3 time units. The connection strength from Bob to Dave is not shown, but follows a similar pattern, as the intervention also addressed this connection.

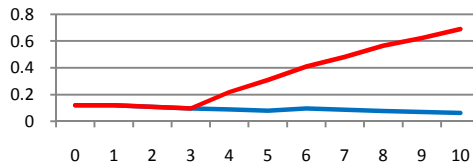


Fig. 8. Connection strength over time from A to D: without intervention (lower line) and with intervention (upper line)

7 Discussion

This paper addresses how a social network can be used to effectively support the development of healthy lifestyle behaviour, by combining an ambient intelligent system with social media. More specifically, it was shown how a social network on Twitter can be supported by an ambient system that monitors and analyzes the social interactions (the tweets exchanged) in the background, and also generates certain network interventions. Examples of these interventions are those encouraging to strengthen the connections between specific members of the social network, or suggesting some members to be more expressive in messages. The use of such an intelligent system in combination with a social network aims at affecting the dynamics of how the network functions and develops over time, using the potential in the network to positively affect members that need some extra attention from the network. It should be acknowledged however, that purposely changing networks could be challenging, as people may not easily be inclined to accept advice on how often they should communicate or on the intensity with which they do so. Yet as it seems a very promising venue for behaviour change, future work may want to explore how these interventions could be designed such that they are effective and acceptable to the user.

The proposed ambient intelligent system uses a neurologically inspired computational model for (1) how decisions for certain behaviour are affected by certain levels of emotions (based on emotion-related valuing), and (2) how such emotion levels are affected by social interactions (based on mirroring). This computational model is incorporated by an ambient system and is used as a basis for analysis of the social network and its members, in order to estimate how the emotional states of the network members develop over time. The ambient system has been implemented in a Twitter environment and determines, based on the analysis, when interventions are desirable and which interventions are suitable. Two scenarios generated with the system were discussed, illustrating how the system works.

In future work, more extensive experiments with Twitter users will be conducted to determine the effects and user evaluations of the model and the proposed interventions. Moreover, extensions of the system's possible interventions will be explored, as well as criteria under which they can be applied.

References

1. Arbib, M.A.: *How the Brain Got Language: The Mirror System Hypothesis*. Oxford University Press (2012)
2. Balahur, A., Montoyo, A., Martinez-Barco, P., Boldrini, E. (eds.): *Proc. of the Third International Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA 2012*. Association for Computational Linguistics. ACL (2012)
3. Bejarano, T.: *Becoming Human: From Pointing Gestures to Syntax*. John Benjamins (2011)
4. Bollen, J., Mao, H. and Pepe, A.: *Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena*. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011)
5. Bosse, T., Hoogendoorn, M., Klein, M.C.A., Treur, J.: *An Ambient Agent Model for Monitoring and Analysing Dynamics of Complex Human Behaviour*. *Journal of Ambient Intelligence and Smart Environments* 3, 283–303 (2011)
6. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: *A Language and Environment for Analysis of Dynamics by Simulation*. *International Journal of Artificial Intelligence Tools* 16, 435–464 (2007)
7. Consolvo, S., Everitt, K., Smith, I., Landay, J.A.: *Design Requirements for Technologies that Encourage Physical Activity*. In: *Proc. CHI 2006*, pp. 457–466. ACM Press (2006)
8. Cook, D.J., Augusto, J.C., Jakkula, V.R.: *Ambient Intelligence: Technologies, applications, and opportunities*. *Pervasive and Mobile Computing* 5, 277–298 (2009)
9. Damasio, A.: *Descartes' error: emotion, reason, and the human brain*. Grosset/Putnam, New York (1994)
10. Duell, R., Memon, Z.A., Treur, J., van der Wal, C.N.: *An Ambient Agent Model for Group Emotion Support*. In: Cohn, J., Nijholt, A., Pantic, M. (eds.) *Proceedings of the Third International Conference on Affective Computing and Intelligent Interaction, ACII 2009*, pp. 550–557. IEEE Computer Society Press (2009)
11. Gallant, M.P.: *The Influence of Social Support on Chronic Illness Self-management: A Review and Directions for Research*. *Health Education and Behavior* 30, 170–195 (2003)
12. Gilbert, E., Karahalios, K.: *Predicting Tie Strength With Social Media*. In: *Proc. CHI 2009* (2009)
13. Granovetter, M.: *The Strength of Weak Ties: A Network Theory Revisited*. *Sociological Theory* 1, 201–233 (1983)
14. Iacoboni, M.: *Mirroring People*. Farrar, Straus & Giroux, New York (2008)
15. JSON (2012), <http://www.json.org/>
16. Kendall, L., Hartzier, A., Klasnja, P.V., Pratt, W.: *Descriptive Analysis of Physical Activity Conversations on Twitter*. In: *Proc. CHIEA 2011*, pp. 1555–1560 (2011)
17. Kramer, A.: *The Spread of Emotion via Facebook*. In: *Proc. CHI 2012*, pp. 767–770 (2012)
18. Marsden, P.V., Campbell, K.E.: *Measuring Tie Strength*. *Social Forces* 63, 482–501 (1990)
19. McNeill, L.H., Kreuter, M.W., Subramanian, S.V.: *Social Environment and Physical Activity: a review of concepts and evidence*. *Soc. Sci. Med.* 63, 1011–1022 (2006)
20. Mitrovic, M., Paltoglou, G., Tadic, B.: *Quantitative Analysis of Bloggers Collective Behavior Powered by Emotions*. *J. Stat. Mech.* (2011)
21. Montague, P.R., Berns, G.S.: *Neural Economics and the Biological Substrates of Valuation*. *Neuron* 36, 265–284 (2002)

22. Morris, M.R., Teevan, J., Panovich, K.: A Comparison of Information Seeking Using Search Engines and Social Networks. In: Proc. ICWSM 2010 (2010)
23. Morris, M.R., Teevan, J., Panovich, K.: What Do People Ask Their Social Networks, and Why? A Survey Study of Status Message Q&A Behavior. In: Proc. CHI 2010, pp. 291–294 (2010)
24. Morrison, S.E., Salzman, C.D.: Revaluing the Amygdala. *Current Opinion in Neurobiology* 20, 221–230 (2010)
25. Murray, E.A.: The amygdala, reward and emotion. *Trends Cogn. Sci.* 11, 489–497 (2007)
26. MySQL (2012), <http://www.mysql.com/>
27. Palmer, C.A., Baucom, D.H., McBride, C.M.: Couple Approaches to Smoking Cessation. In: Schmalings, K.B., Sher, T.G. (eds.) *The Psychology of Couples and Illness: Theory, Research, and Practice*, pp. 311–336. American Psychological Association, Washington, DC (2000)
28. Paltoglou, G., Thelwall, M., Buckley, K.: Online Textual Communications Annotated with Grades of Emotion Strength. In: *Proceedings of the 3rd International Workshop of Emotion: Corpora for Research on Emotion and Affect*, pp. 25–31 (2010)
29. PHP (2012), <http://www.php.net/>
30. Pineda, J.A. (ed.): *Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition*. Humana Press Inc. (2009)
31. Quinn, J.A., Pascoe, A., Wood, W., Neal, D.T.: Can't Control Yourself? Monitor Those Bad Habits. *Pers. Soc. Psychol. Bull.* 36, 499–511 (2010)
32. Rjmetrics, New data on Twitter's Users and Engagement (2010), <http://themetricssystem.rjmetrics.com/2010/01/26/new-data-on-twitters-users-and-engagement/>
33. Rizzolatti, G., Sinigaglia, C.: *Mirrors in the Brain: How Our Minds Share Actions and Emotions*. Oxford University Press (2008)
34. Rogers, E.M., Cartano, D.G.: Methods of Measuring Opinion Leadership. *Public Opinion Quarterly* 26, 435–441 (1962)
35. Salzman, C.D., Fusi, S.: Emotion, Cognition, and Mental State Representation in Amygdala and Prefrontal Cortex. *Annu. Rev. Neurosci.* 33, 173–202 (2010)
36. Scanzfeld, D., Scanzfeld, V., Larson, E.L.: Dissemination of health information through social networks: Twitter and antibiotics. *Amer. J. of Infection Control* 38, 182–188 (2010)
37. Spanier, P.A., Allison, K.R.: General Social Support and Physical Activity: An Analysis of the Ontario Health Survey. *Canadian Journal of Public Health* 92, 210–213 (2001)
38. Sugrue, L.P., Corrado, G.S., Newsome, W.T.: Choosing the Greater of Two Goods: Neural Currencies for Valuation and Decision Making. *Nat. Rev. Neurosci.* 6, 363–375 (2005)
39. Twitter (2012), <http://support.twitter.com/>
40. Valente, T.W.: *Social Networks and Health: Models, Methods and Applications*. Oxford University Press (2010)
41. Wing, R.R., Jeffrey, R.W.: Benefits of Recruiting Participants with Friends and Increasing Social Support for Weight Loss and Maintenance. *Journal of Consulting and Clinical Psychology* 67, 132–138 (1999)
42. Zimmerman, R.S., Connor, C.: Health Promotion in Context: The Effects of Significant Others on Health Behavior Change. *Health Educ. Behav.* 16, 57–75 (1989)

A Computational Analysis of Joint Decision Making Processes

Rob Duell and Jan Treur

VU University Amsterdam, Agent Systems Research Group
De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands
rob@duell.net, treur@cs.vu.nl,
<http://www.cs.vu.nl/~treur>

Abstract. In this paper a computational analysis is made of the circumstances under which joint decisions are or are not reached. Joint decision making as considered does not only concern a choice for a common decision option, but also a good feeling about it, and mutually acknowledged empathic understanding. As a basis a computational social agent model for joint decision making is used. The model was inspired by principles from neurological theories on mirror neurons, internal simulation, and emotion-related valuing. The computational analysis determines the different possible outcomes of joint decision making processes, and the types of processes leading to these outcomes.

1 Introduction

Joint decision making may occur when different persons make a choice for a common decision option. However, joint decision making involves more than just making such a common choice. Is it really a joint decision when a common choice is made, but one of the persons does not feel good about it? And can it really be called a joint decision when one of the persons feels good with the chosen option, but does not experience any empathic understanding from the other person? For a genuine joint decision making processes as addressed in this paper the answer on such questions is ‘no’. For example, when a person does not feel good about a chosen option, probably any future occasion will be used to come to a different choice; decisions without a solid emotional grounding may not last long. Similarly, not experiencing empathic understanding from another person may also cast doubt on the chosen option. To take into account such realistic social phenomena, a joint decision as addressed here is considered to be characterised by three elements:

- A choice for a common decision option
- A good feeling about the chosen option
- Mutually acknowledged empathic understanding

Not all joint decision making processes may end up satisfying all three criteria. Maybe a common choice is made but one (or both) of the persons does not feel good about it. Or a common choice is made and both the persons feel good about it, but due to lack of verbal and/or nonverbal communication no mutual empathic understanding is acknowledged. Moreover, one type of outcome can be reached in different ways. Was one of the persons ahead in the process and affecting the other(s)? For a given

person, did the choice for the option come first and the good feeling later, or was it the other way around? Viewed from this perspective, joint decision making processes offer a complex landscape with a wide variety of possibilities to be explored.

Developments in social neuroscience indicate some of the mechanisms underlying the different elements in joint decision making processes (e.g., [7, 13, 18]). In [40] a computational social agent model was introduced incorporating such mechanisms. In the current paper this model is used as a point of departure to analyse computationally the different types of joint decision making processes that may occur.

In the paper, first in Section 2 some core concepts used are briefly reviewed. Next, in Section 3 the adopted social agent model is presented. Section 4 presents a classification of the different types of outcomes of joint decision making processes. In Section 5 the same is done for the different types of processes leading to such outcomes. Finally, Section 6 is a discussion.

2 Mirroring, Internal Simulation and Emotion-Related Valuing

Two concepts used here as a basis for joint decision making are mirror neurons and internal simulation; in combination they provide an individual's mental function of mirroring mental processes of another individual (see also [39]). Mirror neurons are not only firing when a subject is preparing an action, but also when somebody else is performing or preparing this action and the subject just observes that. They have first been found in monkeys (cf. [15, 34]), and after that it has been assumed that similar types of neurons also occur in humans, with empirical support, for example, in [25] based on fMRI, and [14, 30] based on single cell experiments with epilepsy patients (see also [23, 24, 27]). The effect of activation of mirror neurons is context-dependent. A specific type of neurons has been suggested to be able to indicate such a context. They are assumed to indicate self-other distinction and exert control by allowing or suppressing action execution; e.g., [6, 19, 24], and [23], pp. 196-203.

Activation states of mirror neurons play an important role in *mirroring* mental processes of other persons by *internal simulation*. In [26] the following causal chain for generation of felt emotions is suggested (see also [12], pp. 114-116):

sensory representation → preparation for bodily changes → expressed bodily changes
→ emotion felt = based on sensory representation of (sensed) bodily changes

As a further step *as-if body loops* were introduced bypassing actually expressed bodily changes (cf. [8], pp. 155-158; see also [10], pp. 79-80; [11, 12]):

sensory representation → preparation for bodily changes = emotional response →
emotion felt = based on sensory representation of (simulated) bodily changes

An as-if body loop describes an *internal simulation* of the bodily processes, without actually affecting the body, comparable to simulation in order to perform, for example, prediction, mindreading or imagination; e.g., [2], [16], [17], [20], [28]. The feelings generated in this way play an important role in valuing predicted or imagined effects of actions, in relation to amygdala activations; see, e.g., [29], [31]. The emotional response and feeling mutually affect each other in a bidirectional manner: an as-if body loop usually has a cyclic form (see, for example, [11], pp. 91-92; [12], pp. 119-122):

emotion felt = based on sensory representation of (simulated) bodily changes →
 preparation for bodily changes = emotional response

As mirror neurons make that some specific sensory input (an observed action of another person) directly links to related preparation states, they combine well with as-if body loops; see also [39], or [12], pp. 102-104. In this way states of other persons lead to activation of some of a person's corresponding own states that at the same time play a role in the person's own feelings and decisions for actions. This provides an effective mechanism for how observed actions and feelings and own actions and feelings are tuned to each other. Thus a mechanism is obtained which explains how in a social context persons fundamentally affect each other's individual decisions and states, including feelings. Moreover, it is also the basis for empathic understanding of other persons' preferences and feelings. Both the tuning and convergence of action tendencies and the mutual empathic understanding play a crucial role in joint decision making processes. Mutually acknowledged empathic understanding as used here is based on the following criteria (see also [36]): (1) showing the same state as the other agent (nonverbal part of the empathic response), and (2) acknowledging that the other agent has this state (verbal part of the empathic response).

In the area of decision making the role of emotions has been discussed for example, in [1, 8]. If you make a decision with a bad feeling it may be questioned how robust the decision is. The focus in decision making is on how to perform valuing of decision options. Feelings generated in relation to an observed situation and prepared action option play an important role in valuing predicted or imagined effects of such an action in the situation. Such valuations have been related to amygdala activations (see, e.g., [1, 8, 29, 31]). Although traditionally an important function attributed to the amygdala concerns the context of fear, in recent years much evidence on the amygdala in humans has been collected showing a function beyond this fear context. Stimuli trigger emotional responses for which (by internal simulation) a prediction is made of consequences. Feeling these emotions represents a way of experiencing the value of such a prediction: to which extent it is positive or negative. This valuation in turn affects the activation of the decision option.

3 The Adopted Social Agent Model

The issues and perspectives briefly reviewed in Section 2 have been used as a basis for the neurologically inspired social agent model presented in [40]; in summary:

- Decision making uses *emotion-related valuing of predicted effects* of action options
- Both the tendency to go for an action and the associated emotion are transferred between agents via *mirroring processes* using *internal simulation*
- Mirroring processes induce a process of mutually *tuning* the considered actions and their emotion-related valuations, and the development of *empathic understanding*
- The outcome of a joint decision process in principle involves three elements: a *common action* option, a *shared positive feeling* and *valuation* for the effect of this action option, and mutually *acknowledged empathic understanding* for both the action and feeling
- The mutually acknowledged empathic understanding is based on the following criteria:
 - (a) Showing the same state as the other agent (nonverbal part of the empathic response)
 - (b) Acknowledging that the other agent has this state (verbal part of empathic response)

For an overview, see Fig. 1. Here the circles denote states and the arrows temporal-causal connections between states. In the model *s* denotes a *stimulus*, *a* an *option* for an *action* to be decided about, and *e* a world state which is an *effect* of the action. The effect state *e* is *valued* by associating a *feeling* state *b* to it, which is considered to be positive for the agent (e.g., in accordance with a goal). The state properties used in the model are summarised in Table 1.

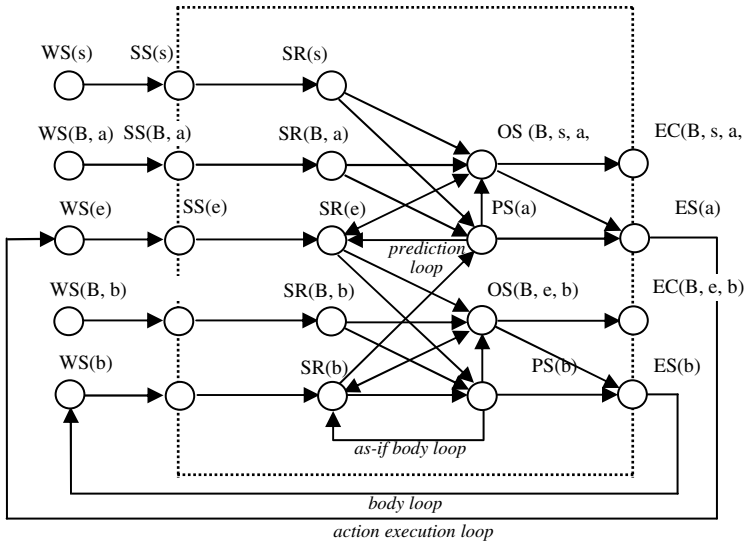


Fig. 1. Overview of the social agent model

The social agent model uses ownership states for actions *a* and their effects *e*, both for self and other agents, specified by $OS(B, s, a, e)$ with *B* another agent or self, respectively (see Fig. 1). Similarly, ownership states are used for emotions indicated by body state *b*, both for self and other agents, specified by $OS(B, e, b)$ with *B* another agent or self.

Table 1. State properties used

| notation | description |
|------------------|---|
| $WS(W)$ | world state <i>W</i> : for an action <i>a</i> of agent <i>B</i> , a feeling <i>b</i> of agent <i>B</i> , a stimulus <i>s</i> , effect <i>e</i> , or an emotion indicated by body state <i>b</i> |
| $SS(W)$ | sensor state for <i>W</i> |
| $SR(W)$ | sensory representation of <i>W</i> |
| $PS(X)$ | preparation state for <i>X</i> : action <i>a</i> or expressing emotion by body state <i>b</i> |
| $ES(X)$ | execution state for <i>X</i> : action <i>a</i> or expressing emotion by body state <i>b</i> |
| $OS(B, s, a, e)$ | ownership state for <i>B</i> of action <i>a</i> with effect <i>e</i> and stimulus <i>s</i> |
| $OS(B, e, b)$ | ownership state for <i>B</i> of emotion indicated by body state <i>b</i> and effect <i>e</i> |
| $EC(B, s, a, e)$ | communication to <i>B</i> of ownership for <i>B</i> of action <i>a</i> with effect <i>e</i> and stimulus <i>s</i> |
| $EC(B, e, b)$ | communication to <i>B</i> of ownership for <i>B</i> of emotion indicated by <i>b</i> and effect <i>e</i> |

As an example, the four arrows to $OS(B, s, a, e)$ in Fig. 1 show that an ownership state $OS(B, s, a, e)$ is affected by the preparation state $PS(a)$ for the action *a*, the sensory representation $SR(b)$ of the emotion-related value *b* for the predicted effect *e*, the sensory representation $SR(s)$ of the stimulus *s*, and the sensory representation $SR(B)$ of the

agent B. Note that $s, a, e, b,$ and B are parameters for stimuli, actions, effects, body states, and agents. In a given model multiple instances of each of them can occur.

Prediction of effects of prepared actions is modelled using the connection from the preparation $PS(a)$ of the action a to the sensory representation $SR(e)$ of the effect e . Suppression of the sensory representation of a predicted effect (according to, e.g., [3], [4], [28]) is modelled by the (inhibiting) connection from the ownership state $OS(B, s, a, e)$ to sensory representation $SR(e)$. The control exerted by the ownership state for action a is modelled by the connection from $OS(B, s, a, e)$ to $ES(a)$. Communicating ownership for an action (a way of expressing recognition of the other person’s states, as a verbal part of showing empathic understanding) is modelled by the connection from the ownership state $OS(B, s, a, e)$ to the communication effector state $EC(B, s, a, e)$. Similarly, communicating of ownership for an emotion for effect e indicated by b is modelled by the connection from the ownership state $OS(B, e, b)$ to the communication effector state $EC(B, e, b)$. Connections between states j and i (the arrows in Fig. 1) have strengths or weights, indicated by $\omega_{j,i}$. A weight usually has a value between -1 and 1 and may depend on the specific instance for agent B , stimulus s , action a and/or effect state b involved. Note that in general weights are assumed non-negative, except for inhibiting connections, which model suppression of the sensory representation of effect e , or of the sensory representation of body state b . In [40] the dynamics following the connections between the states in Fig. 1 are described in more detail. This is done for each state by a dynamic property specifying how the activation value for this state is updated based on the activation values of the states connected to it (the incoming arrows in Fig. 1). For a state i depending on multiple other states, to update its activation level, input values for incoming activation levels are to be combined to some aggregated input value $agginput_i$. This update itself then takes place according to a differential equation

$$dy/dt = \gamma [agginput_i - y_i]$$

where γ is the update speed for state i , $agginput_i$ is the aggregated input for i , and y_i is the activation level of state i . The aggregation is created from the individual inputs $\omega_{j,i} y_j$ for all states j connected toward state i , where $\omega_{j,i}$ is the strength of the connection from j to i (a number between -1 and 1). For this aggregation a combination function $f(V_1, \dots, V_k)$ is needed, applied to the different incoming values $V_j = \omega_{j,i} y_j$. Using this, the above differential equation can be expressed as:

$$dy/dt = \gamma [f(\omega_{1,i} y_1, \dots, \omega_{k,i} y_k) - y_i]$$

Here only for states j connected to state i the value of $\omega_{j,i}$ can be nonzero, for not connected states they are trivially set 0 ; for simplicity of notation, often the arguments for not connected states are left out of the function f . The combination function f is a function for which different choices can be made, for example, the identity function $f(W) = W$ or a combination function based on a continuous logistic threshold function of the form

$$th(\sigma, \tau, X) = \left(\frac{1}{1 + e^{-\sigma(X - \tau)}} - \frac{1}{1 + e^{\sigma\tau}} \right) (1 + e^{-\sigma\tau}) \quad \text{or} \quad th(\sigma, \tau, X) = \frac{1}{1 + e^{-\sigma(X - \tau)}}$$

with σ a steepness and τ a threshold value, when $X \geq 0$, and 0 when $X < 0$. Note that for higher values of $\sigma\tau$ (e.g., $\sigma > 20/\tau$) the right hand side threshold function can be used as an approximation. In the example simulations, for single connections, f is

taken the identity function $f(W) = W$, and for the other states f is a combination function based on the logistic threshold function: $f(X_1, X_2) = th(\sigma, \tau, X_1+X_2)$, and similarly for more arguments; other types of combination functions might be used as well.

Note that in the model s , a , e , b , and B are parameters for stimuli, actions, effects, body states, and agents, respectively; multiple instances for each of them can be used.

The agent model has been computationally formalised in differential equation format and using the hybrid modeling language LEADSTO (cf. [5]); see [40] for further details of the social agent model.

4 Different Types of Outcomes

The variety of possibilities for joint decision processes is explored in two steps. First, in this section the different possible outcomes are analysed (abstracting from the temporal dimension), and their dependence on the different possible contributions by the different agents. Abstracting from the temporal dimension means that the exact timing is left out of consideration in the current section, as, for example, is also done in a numerical equilibrium analysis. These temporal aspects will be addressed as a second step in the next section. It is very hard to explore in a systematic manner all different possibilities for a model with numerical values. Therefore both in this and in the next section the introduced approach abstracts from the quantitative aspects of (activation levels of) states; instead abstracted binary qualitative states are adopted, for which states either occur or do not occur; they can be related to numerical values by assuming some threshold value, for example, 0.5. To obtain a limited number of (qualitative) states some one-to-one dependencies of states are assumed. More specifically, it is assumed that within a given agent A faithful expression and communication takes place with respect to any other agent B , which is formulated as follows:

- A has an intention for option O if and only if A expresses this intention
- A has a positive feeling for option O if and only if A expresses this feeling
- A acknowledges understanding that another agent B has the intention for option O if and only if A has an ownership state for B for this intention
- A acknowledges understanding that another agent B has a positive feeling for option O if and only if A has an ownership state for B for this feeling

Given these assumptions the number of relevant states can be limited. A contribution of one of the agents A with respect to another agent B is then assumed to be represented as any subset of the set of the following four states that can be generated (at some point in time) by agent A or not:

- A has an intention for option O
- A has a positive feeling for option O
- A acknowledges understanding that B has an intention for option O
- A acknowledges understanding that B has a positive feeling for option O

Given these four states that each can occur or not occur for a given agent, theoretically 16 possibilities can be distinguished, as shown in Table 2. Note that it is assumed that both for feeling and for intention acknowledgements always occur, for feeling or no feeling, and for intention or no intention. For example, labels in Tables 2 and 3 such as ‘no intention acknowledgement’ are interpreted as acknowledgement for no intention.

Table 2. The 16 different possible outcomes for one agent

| | | | | | | | | | | | | | | | | |
|--|---------------------------|---------|--------------|---------|----------------------------|---------|--------------|---------|------------------------------|---------|--------------|---------|----------------------------|---------|--------------|---------|
| <i>A</i> acknowledges understanding of <i>B</i> 's intention for <i>O</i> | intention acknowledgement | | | | | | | | no intention acknowledgement | | | | | | | |
| <i>A</i> acknowledges understanding of <i>B</i> 's positive feeling for <i>O</i> | feeling acknowledgement | | | | no feeling acknowledgement | | | | feeling acknowledgement | | | | no feeling acknowledgement | | | |
| <i>A</i> has an intention for <i>O</i> | intention | | no intention | | intention | | no intention | | intention | | no intention | | intention | | no intention | |
| <i>A</i> has a positive feeling for <i>O</i> | feel | no feel | feel | no feel | feel | no feel | feel | no feel | feel | no feel | feel | no feel | feel | no feel | feel | no feel |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |

For example, the possibility indicated by 9 describes a case in which agent *A* has a positive feeling and intention for option *O*, and has acknowledged understanding that *B* has a positive feeling for *O* as well, but acknowledged understanding that *B* has an no intention for *O*. As another example, possibility 4 describes a case in which agent *A* has a no positive feeling and no intention for option *O*, but has acknowledged understanding that *B* has a positive feeling for *O*, and has acknowledged understanding that *B* has an intention for *O*. The possibility described by 1 is the most positive one: feeling, intention and acknowledgements all occur. The possibility described by 16 is the opposite of 1: an emotionally grounded choice for no intention to go for option *O*.

Such possible outcomes for one agent *A* have to be interpreted in the context of other agents *B*, which themselves also show one of these 16 possibilities. To be able to present a feasible systematic overview, the approach is illustrated for the case of two agents. In this case all theoretically possible pairings can be visualised in a two-dimensional form as shown in Fig. 2 for two agents *A* (vertical axis) and *B* (horizontal axis). This pairing leads to $16 \times 16 = 256$ possibilities, all shown in the matrix in Fig. 2. States in this matrix can be indicated by their coordinates (x, y) , where x is the column number referring to agent *A* and y the row number referring to agent *B*.

In this set of all combined states some subsets can be distinguished, indicated in Fig. 2 4 by different colours. First of all there is the subset of full joint decisions: decisions with full emotional grounding and full mutual acknowledged empathy. There are only two of such states (indicated in dark green); they are the full joint decision to go for the option, found in $(1, 1)$, and the opposite joint decision to not go for the option, depicted in $(16, 16)$. The other 254 possible outcomes are not fully joint decisions. However, there is a subset of 12 possibilities concerning at least a common choice with full emotional grounding for each of the agents, and acknowledged empathy by one of the agents (indicated in light green); these can be considered as almost fully joint decisions. Instances can be found at $(1, 5)$, $(1, 9)$, $(1, 13)$, $(4, 16)$, $(5, 1)$, $(8, 16)$, $(13, 1)$ and $(16, 4)$, $(16, 8)$, $(16, 12)$. The set of all possibilities with common choice with full emotional grounding and acknowledged empathy by none of the agents has 20 different states (indicated in light green with shading). This type of decisions can still be solid due to the individual emotional grounding at both sides, but there is no exchange of empathic understanding between the agents. The other states with a common choice have no full emotional grounding (indicated in light and dark blue), and for this reason can be considered as less solid.

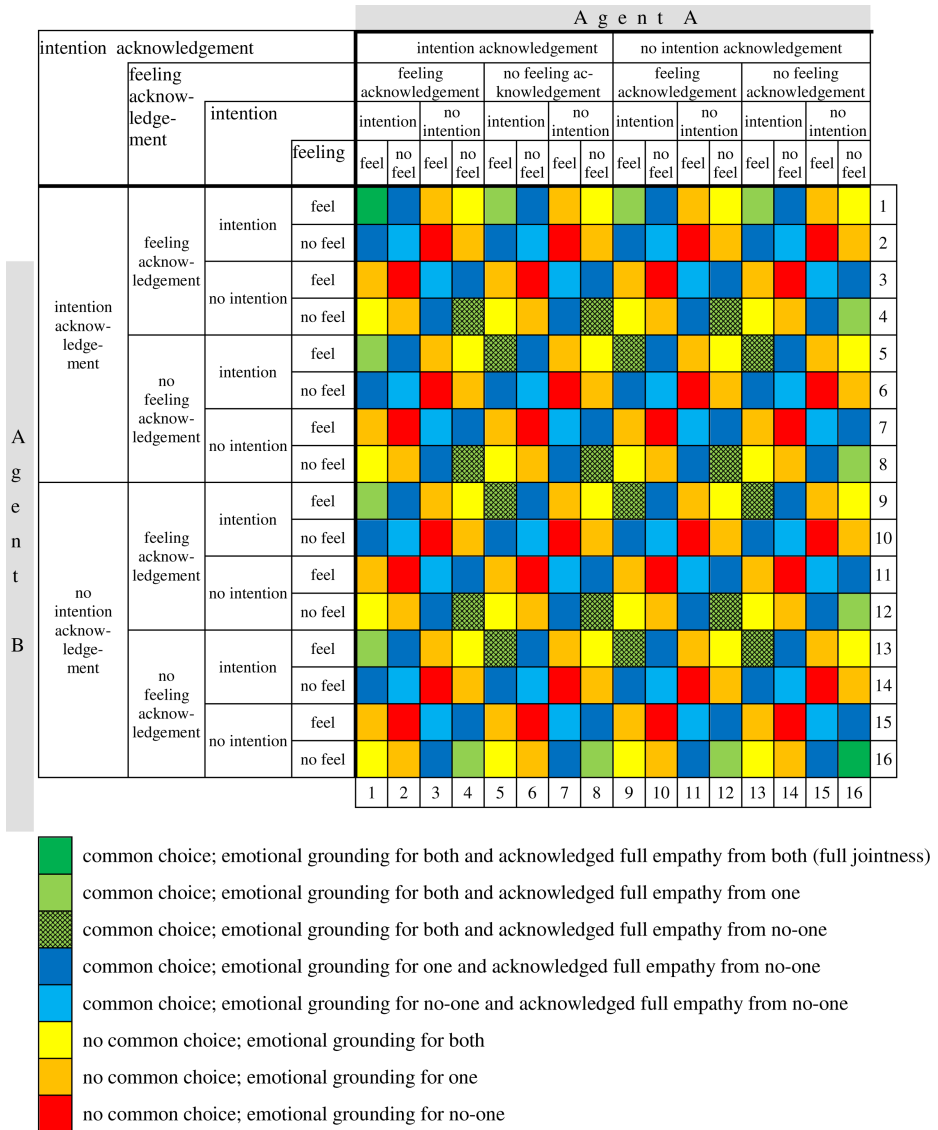


Fig. 2. The 256 different combined possible outcomes for two agents A and B

On the other end of the spectrum, there are also many possibilities of outcomes without a common choice (indicated in yellow, orange and red, depending on the emotional grounding). Note that the overview in Table 2 and Fig. 2 show the theoretically possible combinations; it is not claimed that all of these possibilities have the same extent of plausibility, or proper functioning. As an example, as discussed earlier possibility 4 describes a case in which agent A has a no positive feeling and no intention for option O, but has acknowledged understanding that B has a positive feeling for O, and has acknowledged understanding that B has an intention for O. However,

acknowledging understanding of an intention or a feeling without having (and showing) the same intention or feeling can be considered to be not grounded, and at least is not considered as fulfilling the criteria for showing empathic understanding. More in general, note that, for cases with opposite intentions, full empathy (which also involves expressing the intention of the other) is not feasible: for the set of outcomes without a common choice the expressed intentions are opposite. When in addition the opposite intentions each have emotional grounding (indicated in yellow), apparently not only the intentions are opposite, but also the feelings about them. In Section 5 it will be addressed in more detail how different theoretically possible options can (or cannot) develop over time.

5 Different Types of Processes

This section addresses the temporal aspects for joint decision processes, in a qualitative fashion. These temporal aspects relate to the main causal relationships in the social agent model as depicted in Fig. 1. In accordance with the social agent model, a single agent can be activated either by a world stimulus or by social interaction with other agents. More specifically, activation takes place either by observing a world stimulus $obs(s)$, or by observing the expression from an other agent for feeling by generating $obs(f)$ and/or for intention by generating $obs(i)$. Internally, an agent can develop an intention after:

- observing a world stimulus, or
- observing the intention expression from another agent, or
- as a result of developing feeling

Likewise, an agent can develop feeling after:

- observing the feeling expression from another agent, or
- as a result of developing intention.

Following the development of feeling, an agent will express its feeling by generating $expr(f)$, and acknowledge an observed feeling expression from another agent by generating $ack(f)$. Similarly, following the development of an intention, an agent will express its intention by generating $expr(i)$, and acknowledge an observed intention expression from another agent by generating $ack(i)$. These main causal relationships from the social agent model lead to in total 18 possible types of processes, as shown in a tree representation in Fig. 2. Here each path represents a specific type of single agent process. For example, the path indicated by 8 describes a case in which the agent is activated by a world stimulus and subsequently develops intention. After this, the agent expresses intention and also develops feeling. The developed feeling is also expressed. After another agent expresses feeling and intention, these expressions are acknowledged. As another example, the path indicated by 12 describes a case in which the agent's process is socially activated by a feeling expression from another agent. As a consequence the agent develops and expresses feeling.

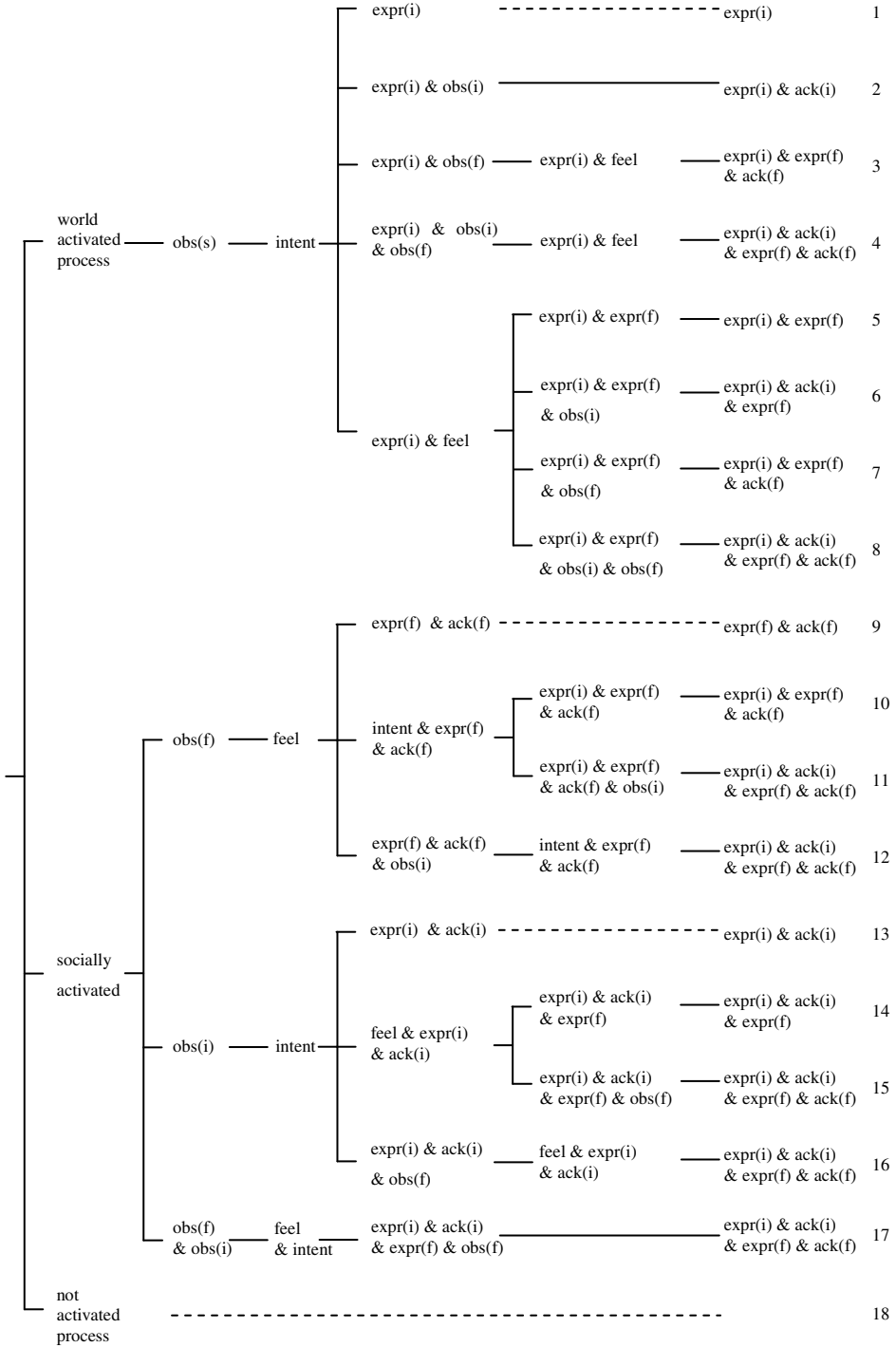


Fig. 3. The 18 different possible types of processes for one agent

In this case the agent also mirrors an observed intention expression from another agent and subsequently develops intention. This results in expressing and acknowledging intention. The single agent process type described by path 18 is a special case in which the agent's process is not activated at all. As a consequence, neither feeling nor intention is developed and therefore no expressions and acknowledgements are generated. The analysis of the process types for two interacting agents is based on combining two single agent process types and representing them as cells in an interaction matrix (Fig. 4) and an initiation matrix (Fig. 5).

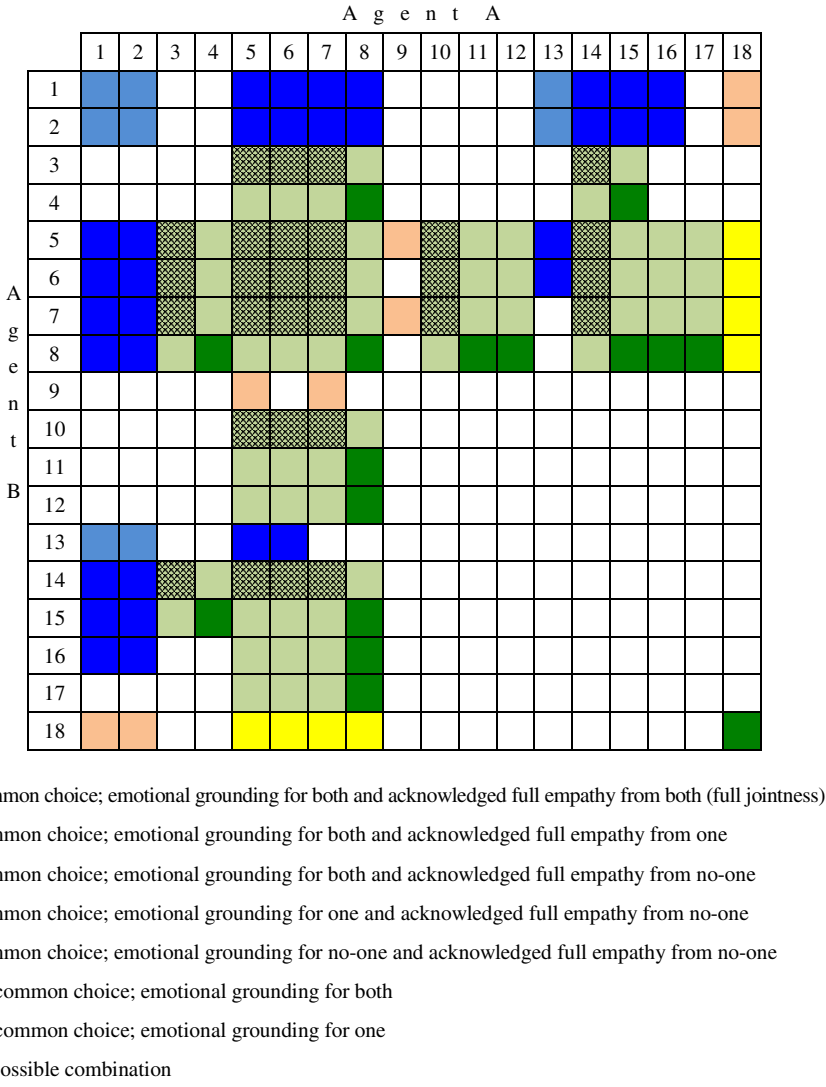


Fig. 4. The different possible combinations of types of processes for two agents

Each matrix dimension represents the 18 single agent process types corresponding to the different paths in the tree depicted in Fig. 3. Each matrix cell represents whether the two single agent process types can occur in combination, and if so, what is the outcome for this specific combination of single agent process types. The outcome of an interaction process between two agents can be classified according to the several outcome-types, as also discussed in Section 4.

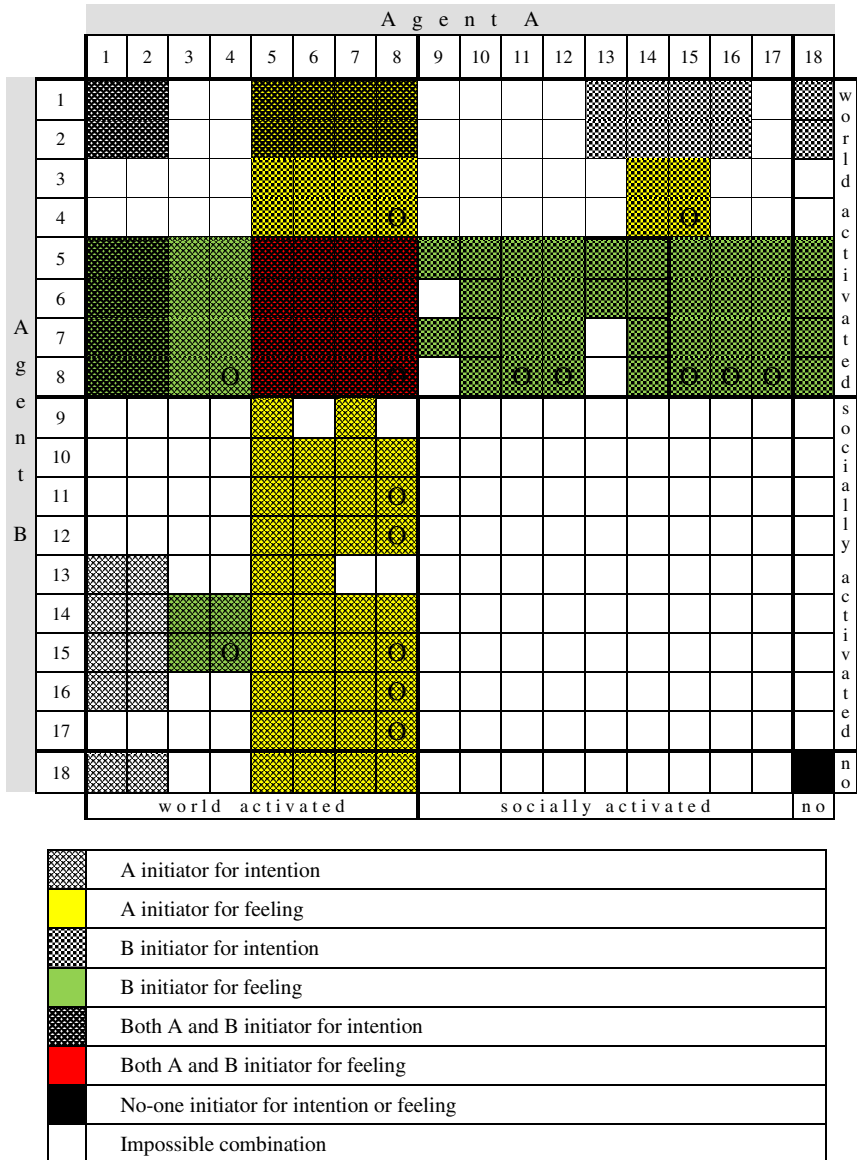


Fig. 5. The 324 different possibilities for initiation for two agents A and B; processes leading to full joint decisions to go for the option are marked by O

These outcome-types range from reaching a common choice with emotional grounding and acknowledged full empathy for both interacting agents, to not reaching a common choice and without emotional grounding for any of the interacting agents. In the matrix the relevant outcome-types are distinguished by different colors. The interaction matrix in Fig. 4 has a number of regions with impossible combinations (indicated by cells left blank), where the two types of single agent processes cannot co-occur. This is the case when one agent does not develop the intention or feeling that the other agent needs for activating its process; for example, agent A needs $obs(f)$, but agent B does not generate $expr(f)$, as is the case for (10, 1) and (10, 2). Another reason for impossibility of a combination is when such a combination would entail a circular mutual dependency. Cells of special interest show a full joint decision with emotional grounding and mutually acknowledged empathic understanding. Examples are cell (18,18) and cell (8,4).

In the initiative matrix (Fig. 5), columns 5 to 8 show processes in which agent A initiates both the intention and the feeling. Some of the types of processes in column 8 lead to a full joint decision to go for the option (marked by an O) for example the one depicted at (8, 17). This represents a process achieving a full joint decision where one person fully develops the decision to go for the option first and then persuades or contagates the other person to go for the option too. The same applies to (17, 8) in which the initiative is from the other agent. In the processes depicted in (8, 4), (8, 11), (8, 12), (8, 15) and (8, 16) (and (4, 8), (11, 8), (12, 8), (15, 8) and (16, 8)) more overlap takes place between the development in one person and the contagion of the other person. In the red shaded area the type of processes are depicted where both agents initiate both the intention and the feeling. For (8, 8) these processes lead to a full joint decision to go for the option. As another example, representing a more complex interaction, the cells (3, 14), (3,15), (4, 14) and (4, 15) depict processes where agent A initiates and expresses the intention which is observed by agent B, who in turn develops the intention as well, and based on that initiates and expresses the feeling which in turn is observed by agent A. For (4, 15) these processes lead to a full joint decision. A similar but opposite process can be found in (15, 4). This shows more types of processes leading to a full joint decision.

6 Simulation Examples

Various simulation experiments have been performed to generate examples of the different types of processes that have been identified. In this section some of them are discussed. As a first example, Fig. 6 (a) shows two agents A and B that reach full jointness illustrating cell (8,15) in the process-type matrices. In this scenario agent A is world-activated and first develops intention and subsequently develops feeling, both represented by their respective preparation states. Agent B is socially activated and follows agent A in first developing intention and then feeling. Both agents express intention and feeling and acknowledge the expressions from the other agent. As another example, Fig. 6 (b) shows an agent A with reduced observation capabilities. In this situation agent B still follows process-type 15, but agent A cannot fully observe the expressions from agent B and therefore does not generate acknowledgements $ack(i)$ and $ack(f)$.

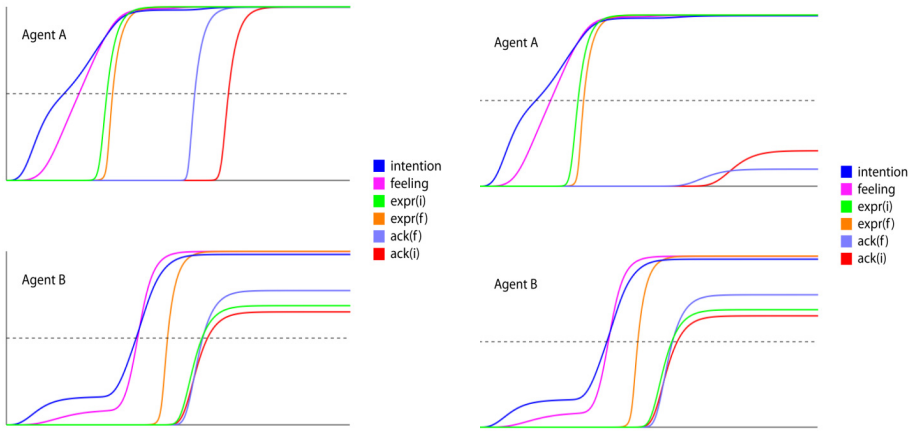


Fig. 6. Example simulations (a) Left hand side showing a full joint decision, illustrating cell (8, 15), (b) right hand side showing no acknowledgements, illustrating cell (5, 15)

Agent A follows process-type 5, the scenario illustrating matrix cell (5,15). In the example depicted in Fig. 7, agent B shows reduced mirroring capabilities for intention. Because of the reduced intention mirroring, feeling mirroring takes over and agent B first develops feeling, followed by developing intention. Agent B neither expresses intention nor acknowledges the intention expression from agent A. Because agent B does not express intention, agent A does not acknowledge intention. This scenario illustrates matrix cell (7,10).

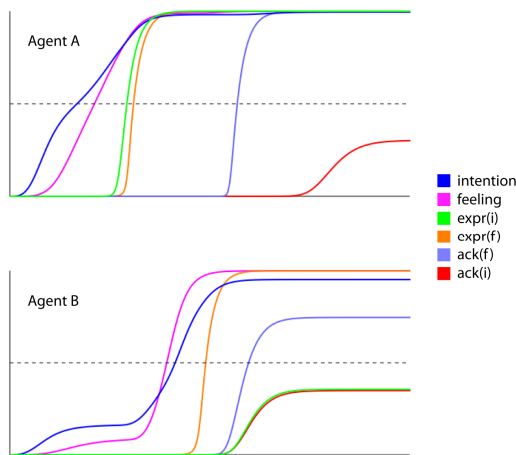


Fig. 7. Example simulation showing reduced mirroring, illustrating cell (7, 10)

Table 1 provides an overview of the connections and their weights as used in the example simulation experiments discussed here. The world stimulus for agent A is 1.0 and for B 0.6 in all scenarios. The context is 1.0 for both agents in all scenarios. All other settings are in accordance with the original social agent model.

Table 3. Overview of setting for the example simulations

| Connection | | Weight values | | | | | |
|-------------|-------------|---------------|---------|------------|---------|------------|---------|
| From | To | Scenario 1 | | Scenario 2 | | Scenario 3 | |
| | | Agent A | Agent B | Agent A | Agent B | Agent A | Agent B |
| SR(B,a) | PS(a) | 1.00 | 0.60 | 1.00 | 0.60 | 1.00 | 0.35 |
| SR(B,b) | PS(b) | 1.00 | 0.50 | 1.00 | 0.50 | 1.00 | 0.50 |
| SR(B,a) | OS(B,s,a,e) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(s) | OS(B,s,a,e) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(s) | OS(A,s,a,e) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(B,b) | OS(B,e,b) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(e) | OS(B,e,b) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(e) | OS(A,e,b) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SR(b) | PS(a) | 1.00 | 0.80 | 1.00 | 0.80 | 1.00 | 0.80 |
| SR(b) | OS(B,e,b) | 1.00 | 0.70 | 1.00 | 0.70 | 1.00 | 0.70 |
| SR(b) | OS(A,e,b) | 1.00 | 0.70 | 1.00 | 0.70 | 1.00 | 0.70 |
| OS(B,s,a,e) | EC(B,s,a,e) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| OS(B,e,b) | EC(B,e,b) | 1.00 | 0.80 | 1.00 | 0.80 | 1.00 | 0.80 |
| OS(A,s,a,e) | ES(a) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| PS(a) | ES(a) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| OS(A,e,b) | ES(b) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| PS(b) | ES(b) | 1.00 | 0.90 | 1.00 | 0.90 | 1.00 | 0.90 |
| SS(B,a) | SR(B,a) | 1.00 | 1.00 | 0.50 | 1.00 | 1.00 | 1.00 |
| SS(B,b) | SR(B,b) | 1.00 | 1.00 | 0.32 | 1.00 | 1.00 | 1.00 |

7 Discussion

This paper presented a computational analysis of different types of processes to reach a common decision. A genuine joint decision does not only concern a choice for a common decision option, but also a good feeling about it, and mutually acknowledged empathic understanding. As a basis for the computational analysis a numerical computational social agent model for joint decision making is used, adopted from [40]. This model was inspired by principles from neurological theories on mirror neurons, internal simulation, and emotion-related valuing. For the analysis, this model was abstracted to a qualitative form.

The analysis provided on the one hand a systematic overview of the different possible outcomes of fully successful and less successful joint decision making processes, abstracting from the temporal dimension of the processes involved. On the other hand it provided a systematic overview of the possible types of processes leading to these outcomes. The different types of outcomes and processes may relate to specific cognitive and social neurological characteristics of the persons. For example, persons with a not well-functioning mirror system may experience difficulties both in reaching a common choice and affective and empathic states in a decision process; e.g., [23, 32, 35]. On the other hand, persons who have a not well-functioning system for emotion-related valuing turn out to experience often problems in decision making in general; e.g., [1, 8, 9, 10, 11, 31]. The computational analysis contributed in this paper may provide a basis to further explore such relationships in the context of joint decision making. From a wider perspective the presented model-based analysis of joint decision making may provide a basis for further work aiming at development of support for such decision making processes, for example, in the form of a mediation

assistant. Such an assistant may provide analyses and give advices in order to develop a joint decision, and take care that no escalating conflicts arise. This will be a direction of future research.

References

1. Bechara, A., Damasio, H., Damasio, A.R.: Role of the Amygdala in Decision-Making. *Ann. N.Y. Acad. Sci.* 985, 356–369 (2003)
2. Becker, W., Fuchs, A.F.: Prediction in the Oculomotor System: Smooth Pursuit During Transient Disappearance of a Visual Target. *Experimental Brain Res.* 57, 562–575 (1985)
3. Blakemore, S.-J., Frith, C.D., Wolpert, D.M.: Spatio-Temporal Prediction Modulates the Perception of Self-Produced Stimuli. *J. of Cognitive Neuroscience* 11, 551–559 (1999)
4. Blakemore, S.-J., Wolpert, D.M., Frith, C.D.: Why can't you tickle yourself? *Neuroreport* 11, 11–16 (2000)
5. Bosse, T., Jonker, C.M., van der Meij, L., Treur, J.: A Language and Environment for Analysis of Dynamics by Simulation. *Intern. J. of AI Tools* 16, 435–464 (2007)
6. Brass, M., Spengler, S.: The Inhibition of Imitative Behaviour and Attribution of Mental States. In: Striano, T., Reid, V. (eds.) *Social Cognition: Development, Neuroscience, and Autism*, pp. 52–66. Wiley-Blackwell (2009)
7. Cacioppo, J.T., Berntson, G.G.: *Social neuroscience*. Psychology Press (2005)
8. Damasio, A.R.: *Descartes' Error: Emotion, Reason and the Human Brain*. Papermac, London (1994)
9. Damasio, A.R.: The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. *Philosophical Transactions of the Royal Society: Biological Sciences* 351, 1413–1420 (1996)
10. Damasio, A.R.: *The Feeling of What Happens. Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York (1999)
11. Damasio, A.R.: *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Vintage Books, London (2003)
12. Damasio, A.R.: *Self comes to mind: constructing the conscious brain*. Pantheon Books, NY (2010)
13. Decety, J., Cacioppo, J.T. (eds.): *Handbook of Social Neuroscience*. Oxford University Press (2010)
14. Fried, I., Mukamel, R., Kreiman, G.: Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition. *Neuron* 69, 548–562 (2011)
15. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action Recognition in the Premotor Cortex. *Brain* 119, 593–609 (1996)
16. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mindreading. *Trends in Cognitive Sciences* 2, 493–501 (1998)
17. Goldman, A.I.: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford Univ. Press, New York (2006)
18. Harmon-Jones, E., Winkielman, P. (eds.): *Social neuroscience: Integrating biological and psychological explanations of social behavior*. Guilford, New York (2007)
19. Hendriks, M., Treur, J.: Modeling Super Mirroring Functionality in Action Execution, Imagination, Mirroring, and Imitation. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) *ICCCI 2010, Part I. LNCS (LNAI)*, vol. 6421, pp. 330–342. Springer, Heidelberg (2010)
20. Hesslow, G.: Conscious thought as simulation of behaviour and perception. *Trends Cogn. Sci.* 6, 242–247 (2002)

21. Hoogendoorn, M., Treur, J., van der Wal, C.N., van Wissen, A.: Agent-Based Modelling of the Emergence of Collective States Based on Contagion of Individual States in Groups. *Transactions on Computational Collective Intelligence* 3, 152–179 (2011)
22. Hoogendoorn, M., Treur, J., van der Wal, C.N., van Wissen, A.: Modelling the Interplay of Emotions, Beliefs and Intentions within Collective Decision Making Based on Insights from Social Neuroscience. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) *ICONIP 2010, Part I. LNCS (LNAI)*, vol. 6443, pp. 196–206. Springer, Heidelberg (2010)
23. Iacoboni, M.: *Mirroring People: the New Science of How We Connect with Others*. Farrar, Straus & Giroux, New York (2008)
24. Iacoboni, M.: Mesial frontal cortex and super mirror neurons. *Behavioral and Brain Sciences* 31, 30–30 (2008)
25. Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J.C., Rizzolatti, G.: Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology* 3, e79 (2005)
26. James, W.: What is an emotion. *Mind* 9, 188–205 (1884)
27. Keysers, C., Gazzola, V.: Social Neuroscience: Mirror Neurons Recorded in Humans. *Current Biology* 20, 253–254 (2010)
28. Moore, J., Haggard, P.: Awareness of action: Inference and prediction. *Consciousness and Cognition* 17, 136–144 (2008)
29. Morrison, S.E., Salzman, C.D.: Re-valuing the amygdala. *Current Opinion in Neurobiology* 20, 221–230 (2010)
30. Mukamel, R., Ekstrom, A.D., Kaplan, J., Iacoboni, M., Fried, I.: Single-Neuron Responses in Humans during Execution and Observation of Actions. *Current Biology* 20, 750–756 (2010)
31. Murray, E.A.: The amygdala, reward and emotion. *Trends Cogn. Sci.* 11, 489–497 (2007)
32. Pineda, J.A. (ed.): *Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition*. Humana Press Inc (2009)
33. Preston, S.D., de Waal, F.B.M.: Empathy: its ultimate and proximate bases. *Behav. Brain Sci.* 25, 1–72 (2002)
34. Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L.: Premotor Cortex and the Recognition of Motor Actions. *Cognitive Brain Research* 3, 131–141 (1996)
35. Rizzolatti, G., Sinigaglia, C.: *Mirrors in the Brain: How Our Minds Share Actions and Emotions*. Oxford University Press (2008)
36. Singer, T., Leiberg, S.: Sharing the Emotions of Others: The Neural Bases of Empathy. In: Gazzaniga, M.S. (ed.) *The Cognitive Neurosciences*, 4th edn., pp. 973–986. MIT Press (2009)
37. Treur, J.: A Cognitive Agent Model Displaying and Regulating Different Social Response Patterns. In: Walsh, T. (ed.) *Proc. IJCAI 2011*, pp. 1735–1742 (2011)
38. Treur, J.: A Cognitive Agent Model Incorporating Prior and Retrospective Ownership States for Actions. In: Walsh, T. (ed.) *Proc. IJCAI 2011*, pp. 1743–1749 (2011)
39. Treur, J.: From Mirroring to the Emergence of Shared Understanding and Collective Power. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) *ICCCI 2011, Part I. LNCS (LNAI)*, vol. 6922, pp. 1–16. Springer, Heidelberg (2011)
40. Treur, J.: Modelling Joint Decision Making Processes Involving Emotion-Related Valuing and Empathic Understanding. In: Kinny, D., Hsu, J.Y.-j., Governatori, G., Ghose, A.K. (eds.) *PRIMA 2011. LNCS (LNAI)*, vol. 7047, pp. 410–423. Springer, Heidelberg (2011)

Collaboratively Constructing a VDL-Based Icon System for Knowledge Tagging

Xiaoyue Ma and Jean-Pierre Cahier

ICD/Tech-CICO Lab, Université de Technologie de Troyes, BP 2060, 10010 Troyes, France
{xiaoyue.ma, cahier}@utt.fr

Abstract. Tag system for a knowledge organization system centralizes and provides the tags that can be employed in classifying, sharing and seeking knowledge for personal or organizational use within a social community. Considering current constraints of textual tag system and developing iconic tag system, VDL-based iconic tag system has been built and validated to improve knowledge tagging with symbolic interpretation and graphical organization of tag structure. In this paper, we are proposing cooperative creation of such special icon system where VDL-based icons will be applied for social knowledge tagging and sharing. This VDL-based icon system could also serve as a visual knowledge organization system to facilitate icon searching in a given context.

Keywords: icon, tag system, knowledge tagging, knowledge organization system, Hypertopic, social community.

1 Introduction

Tag system for a Knowledge Organization System (KOS) [1] centralizes and provides the tags that can be employed in classifying, sharing and seeking knowledge for personal or organizational use in a social community. Understandable tags and a clear structure (semantic relation) are both essential to establish an effective tag system. However, textual tags, general form of tags appearing in current tag systems, may create problems on comprehending tags or identifying the structure of tags. An increased variety of vocabularies and languages cause connections between tags and documents marked by these textual tags to become less and less distinctive, making the use and reuse of tag system even harder [2] (see Figure 1(a)).

Studies on cognitive psychology like Dual-coding Theory [3] have gradually postulated that both visual and verbal codes are used to organize incoming information into knowledge that can be acted upon, stored, and retrieved for subsequent use. In addition, previous work has theoretically shown that an icon could act as an active visual representation of knowledge by considering graphic characters and symbolic characters [4]. Empirical researches have reflected the notion of "Icons System" like road signs, symbols of fire safety [5] and medical icon system [6]. Once icons in an icon system are applied for knowledge tagging, icon system is equally called as an iconic tag system.

As mentioned in the first paragraph, a tag system is not only interested in representing each tag, but also the tag structure which is increasingly essential to find, and

be able to find again later, a proper tag for knowledge sharing, especially when more diverse knowledge is concerned. If the categorization of knowledge is represented by icons without an explicit structure, users may experience disorientation when faced with too many isolated symbols (see Figure 1(b)).

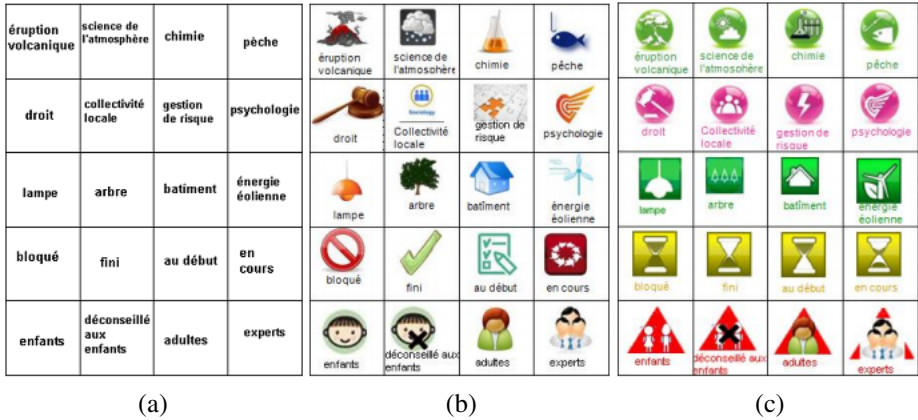


Fig. 1. Three types of tag system compared: (a) textual tag system (b) iconic tag system without explicit structure (c) well-structured iconic tag system

Although it is well known that icons are legible and universal as a visual representation, they have a limit for large use in knowledge tagging [7]. Previous attempts on icon system like symbols of fire safety [5] and medical icon system [6] required lots of designers to confirm system structure and create hundreds of specific icons in which little users' suggestion will be considered. This huge, high standard work makes construction of icon system so inconvenient that majority of practitioners still prefer to use textual tags in knowledge tagging, sharing and other activities on knowledge management.

Moreover, icon comprehension is a complex cognitive task which varies depending on users' different backgrounds, cognitive levels and information goals. This is also one of the reasons that knowledge builders express the categorization in KOS with texts rather than icons because it is hard to point out the very tag with merely an icon. For example, the icon representing a tree may be explained as textual tag "nature" or "plant". Although these possible textual tags are usually sorted in a common unit, iconic tag is useless when the exact meaning needs indicated.

Last but not least, new icons need to be designed again if shared domain changes. Such as in medical icon system [6], each icon is composed of several graphical components representing particular categories. However, when applied field is changed from medicine to sustainable development, the former creation rule will be no longer usable. The toughest thing is not to propose new symbols but to confirm the structure for actual icon system with the least modification. Since usually more than one sharing KOSs are required within a social community, complicate creation rule does restraint communication from one icon system to another. Designers have to think of

sustainable construction of icon system to adapt different cases while simultaneously users have also to get used to new graphical regularity which arises the problem of tiring learning.

VDL-based iconic tag system has been previously described and evaluated [8] [9] in which iconic tags were organized under graphical regularity (see Figure 1(c)). Assessment has proved that these iconic structured tags developed tagging efficiency taking advantage of explicit tag organization. Here tagging efficiency refers to quick tag finding and accurate tag choosing within a tag system. However, icons existed in this tag system were purely suggested by specialists without users' participation and only one icon corresponds to each textual tag.

Our objective is to co-establish a VDL-based icon system in a social community, in use of which the KOS may be constructed visually (iconic tags) and verbally (textual tags) at the same time (see figure 2). There is a cross-fertilization between the icon system (e.g., for its advantages in terms of semiotics, memorization) and the textual system (e.g., for its advantages in terms of disambiguation and lexical precision). The symbolic interpretation of iconic tags will enrich the comprehension of tagged resources; as well the graphical codes of VDL (pre-icons) enhance the connection between tags and resources tagged by them. Both these benefits of VDL-based icon system are assumed to ameliorate the efficiency of knowledge tagging and sharing. It is also meaningful to adopt cooperative activities in this construction where participants from different domains are allowed to contribute to social icon creation for a common sharing context.

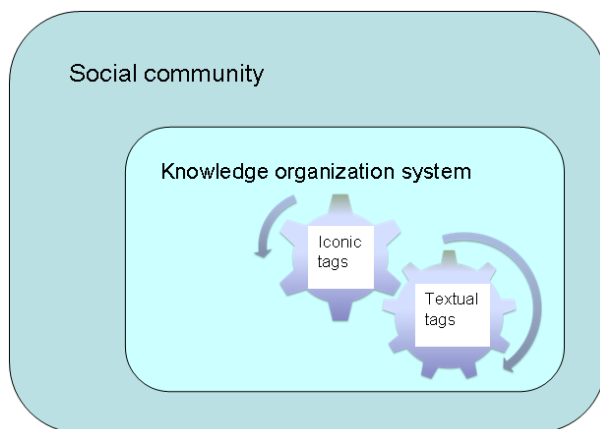


Fig. 2. Constructing KOS within a social community using iconic tags and textual tags

In this paper, cooperative building of a VDL-based icon system will be proposed for knowledge tagging within a social community. Before turning into the introduction of principal concept, several details on VDL-based iconic tags will be reminded. In the section 3, details on how to co-construct such icon system will be presented through three parts: categorization of icons, cooperative activities of participants and some supplementary collaborative functions.

2 Modelling a VDL-Based Iconic Tag System

In iconic tag systems, it has to firstly confirm the way to structure the tags, and then iconize them as well as their structure. Tags in a tag system can be regarded as key words to specify a possible categorization in a KOS. Confirming tag structure is, in fact, recommending a method to organize information and knowledge.

Due to the cooperative dimension of our research, a knowledge model respecting the principles of Social Semantic Web [10] is chosen as the method to organize the tag structure: "Hypertopic" [11]. It proposes classifying items by topics, attributes, and resources. Particularly, Hypertopic indicates that all the topics are from different viewpoints in correspondence with the various kinds of information goals.

Hypertopic suggests a method to categorize knowledge emphasizing the concept of viewpoint which is significant in collaborative knowledge categorization [12]. This categorization typically provides a meaningful structure to manage textual tags brought from topics or values of attributes. All the topics are catalogued under the tree structure considering the common viewpoint as the "parent" node. An actor with a particular view on the items is allowed to add new units of topics by creating a parent node named "my viewpoint". This convenience encourages collaborative knowledge management and social tagging. Attributes and their values also enrich items with additional information. They are joined in pairs with the name of attribute and the value of attribute as a facet [13]. A topic is a heuristic attribute. It can be regarded as a "special" attribute, considering "topic" as the implicit attribute name.

The idea is to benefit from the categorization made by Hypertopic (from topics in viewpoints and values of attributes) and iconize it for a better visualization of separate tags and their structure [8] [9]. The symbolic characters of icons will convey explicitly the represented objects, while graphical characters will help visualizing relations within the tag system. In particular, a special group of icons called "pre-icons" function to signify the categories of tags: the same viewpoint, the same branch of topic or the same name of attribute. Pre-icons act as the common base of iconic tags. The tags in each category will be specified by combining symbols with this corresponding iconic base. Nevertheless pre-icon for the name of attribute is useless in some cases. For example, when iconizing the values of the attribute "language", it is clear enough to represent them independently with national flags.

The approach [9] can be looked upon as a "graphical organizer" named Visual Distinctive Language (VDL) which aims to visually characterize the objects classified according to the protocol Hypertopic. VDL is more interested in the visual tag structure than in the symbol of each icon. A study on graphical semiotics of Bertin [14] proposed six basic visual variables: size, colour, shape, texture, value and orientation. These variables do not have the same ability to express the same information. Among these six visual variables, three of them are in less accordance with the purpose of structure representing: size, orientation and value. It is primarily due to the difficulty to distinguish two iconic tags from different sizes, different orientation or different value depending on the conditions of the computer screen. Moreover, icons are preferred to be designed in a unified size for aesthetic reasons. Limited choice of orientation and value makes it less possible to design iconic categories for a large scale tag system.

By contrast, shape, colour and texture are useful to build VDL with satisfying the visual structure of iconic tag system. All the tags under each viewpoint (pre-icon) are firstly designed into a uniform shape, and then topics classified under different categories from a common viewpoint continue to be added with another visual variable colour as the updated pre-icons. Since tags for topics are catalogued in the tree structure, new visual variables would have been still added to iconize the tags for topics at the following levels. However, on one hand the number of visual variables is limited; on the other hand excessive visual variables reduce the readability of the structure of iconic tags. To simplify VDL of iconic tag system, iconic tags for topics from the second level will always keep the same pre-icon without being distinguished by a new visual variable.

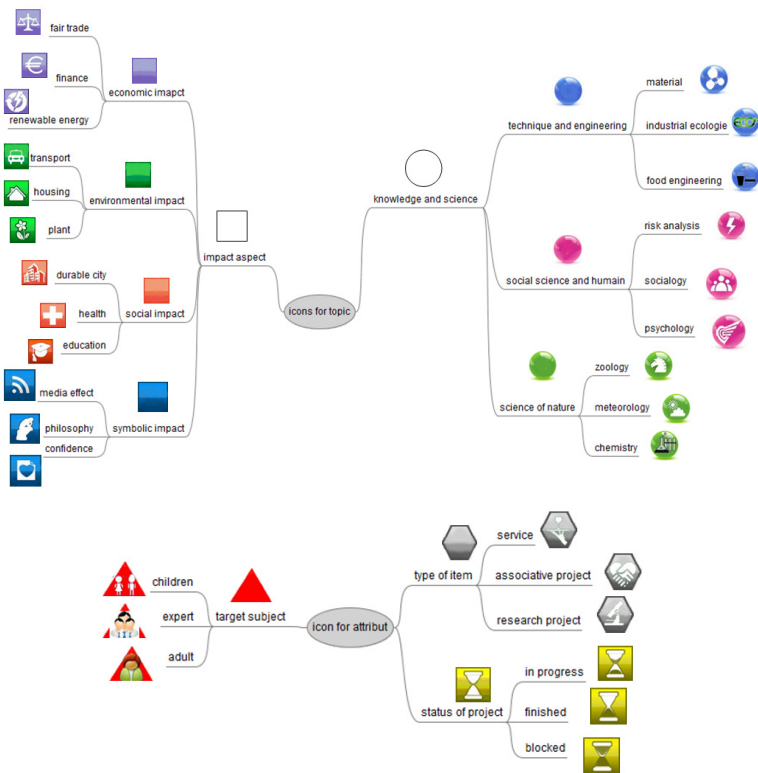


Fig. 3. Examples of Hypertopic-based iconic tags in two viewpoints and three attributes [9]

The graphical rule is similarly applied to attributes, another part of VDL. The name of attribute is directly iconized into coloured shapes and then the value is detailed by joining a symbol onto the pre-icon (except special cases mentioned in the previous paragraph, such as “language”). In case of monochrome, visual variable colour will be replaced by texture preserving all the other rules for the coloured version. The final version of both iconic tags for topics and iconic tags for attributes makes no visual difference unless specifically marked as a tag for topic or a tag for attribute.

Here is an example of VDL-based icons in the context of sustainable development (see figure 3). They are associated with seven categories of topics and three names of attributes. These structured icons provide a global view on how pre-icons visualize the connections within iconic tags.

3 How to Co-create a VDL-Based Icon System for Knowledge Tagging

3.1 Four Roles of Participants in the Co-Construction

In view of the reasons why former icon systems were not largely exploited in KOS, we propose a fresh conception on the co-construction of icon system for knowledge tagging. Collaboration from all users is assumed to facilitate icon design and icon understanding. Here we are considering the case where the icon system exists independently to KOS for icon searching. That is to say the entities in this icon system are purely icons instead of items (documents). However they will be connected automatically when these icons are used to tag the item. Meanwhile, icon system could be also embedded into a KOS where those icons are just iconic labels for item searching.

The cooperative creation of VDL-based icon system aims to achieve a link between three scientific subjects: Knowledge Engineering / Knowledge Management, Human Computer Interaction (HCI) and Computer Supported Collaborative Work (CSCW) [15]. Seeing from figure 4 presented in SeeMe¹, four essential roles participate in the co-construction: experts on Knowledge Management, designers, users and administrators. Experts on Knowledge Management work on categorizing icons by their representing objects - corresponding textual tags - in given context. As stated before [9], VDL-based iconic tags are iconized from textual tags in a KOS. VDL reflects the structure of textual tags which implies the categorization of knowledge while symbols of icons represent textual knowledge classified according to Hyper-topic [11]. As a result, experts on Knowledge Management have to define a categorization of knowledge at the first step, each element of which is textual tag, and then the VDL for this categorization, such as pre-icon for each category will be proposed by designers later. In the meantime, designers create as well symbols of represented objects and supplement fresh icons in time following users' demands.. They will recommend icons for each textual tag, in other words, they actually propose only the symbols for each textual tag and final icons are produced by a combination between symbols and pre-icons. Particularly, all the icons will be sub-titled with the textual tag proposed by experts. The purpose is to set up a unit of representing objects for each icon. Once designers and users recognize the text paired with visual representation, they will confirm the symbolic meaning corresponding to each tag and propose a more suitable icon.

Users are able to take cooperative activities like proposing symbols of icons or commenting on icons in the co-creation of VDL-based icon system. These authorized operations will be precisely explained in the next section. However, all the propositions and modifications are required to be validated by administrators. They act as

¹ Tool for analyzing roles in cooperative activities.

supervisors to make sure that all icons always keep the predefined graphical regularity and that system runs well even with diverse contribution from each role. This mode of collaboration among four roles enhances the effectiveness of each partner on the construction of icon system.

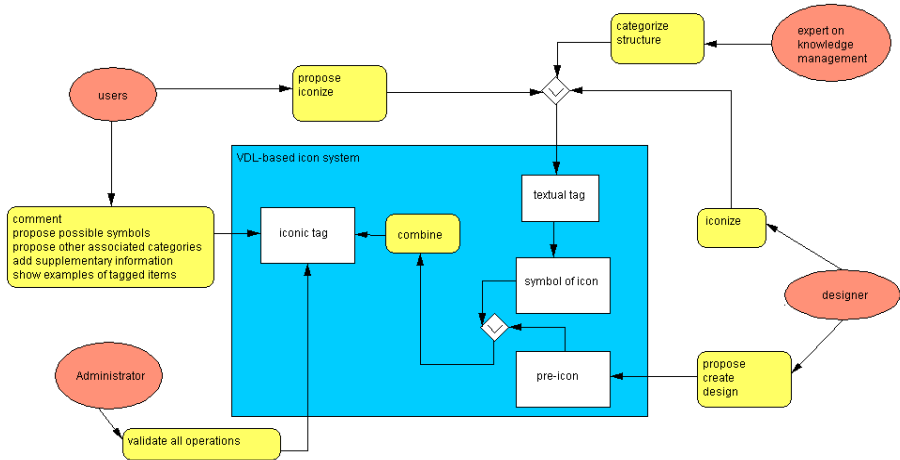


Fig. 4. The proposed participation architecture: four essential roles cooperate in the construction of VDL-based icon system

The icon system we want to create is a semi-participative system in which it is experts on knowledge management and designers that co-define the graphical regularity of icons. Once such icon system is established, with a clear categorization, users are not permitted to alter pre-icons of VDL except for adding their suggested icons under each category. This semi-participative notion provides a common standard for icon design and icon updating and makes sure that the structure of icon system will be solid even though kinds of proposition occur.

What's more, VDL-based icon system facilitates the modification of icon structure. On one hand, it needs frequently revising KOS in a social community, such as inserting a new category or changing some elements in a certain category. Corresponding modifications on iconic tags will be easily applied to VDL-based icon system by designing new pre-icon and attaching symbols onto it. This action will be carried out independently to other existing categories, which assures a relatively solid system structure. On the other hand, it may occur that one community needs more than one sub-icon systems to represent and tag the resources in different KOSs from different departments. Graphical regularity of VDL will be duplicated only by adding or cancelling pre-icons when shared context differs. And then, combination of new symbols with these pre-icons transforms former icon system to present one successfully. Users from different departments are assumed to quickly get used to others' icon system without learning a new rule, which facilitates communication between them.

Reflected those, to co-construct a VDL-based icon system involves firstly a categorization of representing objects of icons to provide a framework of system. Then, VDL is supposed to be proposed and confirmed to define pre-icons for each category.

After this work on informatics, designers and users propose symbols of icons to enrich the entities in the system. At this step, mechanism of proposition has to be drawn up to standard the activities in this co-construction. Finally, as what having been done with documents in a KOS, icons could as well be similarly managed by topics, attributes and resources through Hypertopic model. Each icon is able to be described with supplementary information as its attributes. In the next section, we will illustrate the co-construction of VDL-based icon system: categorization, cooperative activities and discussion on the management of icons as documents.

3.2 Categorization of Icons in an Icon System

3.2.1 Previous Empirical Categorization of Icons According to Symbolic Characters or Graphical Characters

When a number of icons are involved in an icon system, it is more effective to build up a categorization of icons for icon searching. As mentioned in the section before, categorization is essential in an icon system because it facilitates icon seeking and sharing with an explicit structure. Users can easily find a target icon by entering directly into the right category. Similarly, ones who recommend new icons into icon system will also find out the unit to put their icons quickly and accurately. Resulting from this, we need to firstly build up an appropriate categorization of VDL-based icons which will be equally regarded as iconic tags for further social tagging.

Previous researches have studied the categorization of icons. In Wangs' work, researchers have listed and analyzed nine icon taxonomies from 1983 to 2003 [16]. These studies relied on different classifying criterions but always focusing on the relation between symbols of icons and represented objects. The results were summarized that icon taxonomy done before was interested in physical form of icons. Icons were classified according to the cognitive distance with the reality.

A deeper work on icon taxonomy [17] has been recently carried out and demonstrated three methods to classify icons. *Lexical categorization* concerned the pictographical categorization initially divided into lexical words (or content words) and function words (or grammatical words). *Semantic categorization* aims at classifying icons into events and entities, or we can say into actions and objects. *Categorization by representation strategy* was used to convert concepts into pictographs: visual similarity, arbitrary convention and semantic association. Although this work added new factors into icon categorization, it was still conceptually close to former taxonomy emphasizing the sign theory.

Indeed, practitioners of Google Image have shown another method of icon categorization with more attention on graphical features (see figure 5). This function allows quickly selecting the icons of common graphical component without interest in symbolic representation unless we require something in the search bar. However, this time, accordant responses will be displayed all in one size or in one colour to implant side conditions on external appearance.

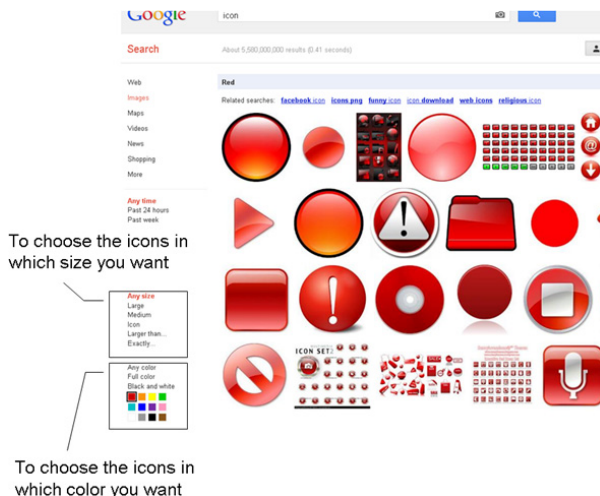


Fig. 5. Interface of icon categorizing and searching by graphical options in Google Image

We can get a primary conclusion that classic categorization of icons focused on symbolic characters while categorization based on graphical qualities is likewise adoptable. However, one unit of icons classified by two methods may not be permanently identical such as the example showed in figure 6. Four icons are catalogued in the same category considering their symbolic characters because they represent all the objects on energy. To the contrary, they are divided into two separate classes when relying on graphical feature: two in red and others in blue. The potential conflict between two standard of categorization complicates the construction of icon system. Experts have to prepare two kinds of categorization of icons to meet different searching goals since normally the categorization through symbolic characters is distinct from that through graphical characters. Consequently, VDL-based icon system aims to avoid the bias towards symbolic-relied icon categorization and simplify the categorization of icons.

3.2.2 Categorization of Icons in VDL-Based Icon System

No matter which icon will be added into VDL-based icon system, it has to obey the creation rule: pre-icon plus symbol. Even original icon was designed randomly in several colors or in a special form, further graphical operations will transform it in accordance with other VDL-based icons. What has to emphasize is that categorization stated below is just a function of system in given context, but not in purpose to serve universal icon taxonomy.

When several VDL-based icons are involved in a common knowledge context, they are supposed to be firstly classified according to the categorization of representing objects. Although this method of categorization is similar with those previous studies on icon taxonomy paying attention to symbols of icons, the criterion is no longer on representing strategy like arbitrary or similarity but real meaning of

the icon. As mentioned in the introduction of VDL-based icons, these structured icons are produced from a group of textual tags. We have firstly a unit of textual tags (sub-titles of icons) which are catalogued by Hypertopic model, and then iconize them as well as their structure (categorization of textual tags). This time, the categorization of icons can be regarded as an opposite process: we have a unit of VDL-based icons, and then the structure of corresponding textual tags will contribute to their categorization. That is to say, iconic tags whose textual prototypes are catalogued in one category will always be kept in the same class of icons.





| Classification of icons | | |
|----------------------------------|---|---|
| Classifying method | Through symbolic characters | Through graphical characters |
| Icons without explicit structure |  |  |
| VDL-based icons |  |  |

Fig. 6. Two criterions of categorization for icons without explicit structure vs VDL-based icons

On one hand, categorization of VDL-based icons brings out a high consistency between categorization through symbolic characters and graphical characters due to the graphical regularity of VDL. For example user who wants to choose an iconic tag “plant” may search from category “nature” or category “green icons”. With a former icon categorization, he has risk to get quite distinct results from two categories. However, in the case of VDL-based icon system, we define that green square icons, namely pre-icon, represent the topic of nature. Icons under category “nature” or category “green icons” are entirely the same. Consequently, VDL-based icon system integrates two separate categorizations into one which simplifies construction and practice.

On the other hand, pre-icons deal better with the situation where one iconic tag is concerned in more than one category, which frequently occurs in KOS. For example, the tag “renewable energy” is a multidisciplinary topic. It is associated with environmental aspect interested in energy and economic aspect respecting reduction on energy consumption. In the categorization of textual tags, renewable energy will appear both under the category “environmental aspect” and category “economic aspect”. However, textual form cannot express explicitly these two categories when tagging resources. On the contrary, VDL-based icons are capable to adapt multi-topic-concerned cases making use of pre-icon. The symbol of “renewable energy” will be attached onto two respective pre-icons corresponding to “environmental aspect” and “economic aspect”, by means of which, one tag can be catalogued under more than one categories. Similarly, if we encounter several icons with the same representing objects, their pre-icons determine which category they are in.

Seen from these evidences, the icon categorization in VDL-based icon system provides an effectual way to find a target icon. This categorization depends initially on Hypertopic-based categorization of represented objects. Simultaneously, the categorization originating from symbolic character of icons coheres with that through graphical character, taking advantage of VDL and pre-icons. The categorization of VDL-based icons also simplifies the construction of icon system and the management of multi-topic-concerned tags.

3.3 Cooperative Activities in VDL-Based Icon System

Graphical regularity of VDL allows easy participation of users and designers. They have just to learn the creation rule and then cooperate in the co-construction. Four types of operative operations will be presented in this section that may take place in VDL-based icon system.

Firstly, users and designers can propose a symbol to an existing icon. In this case, the sub-title (textual tag) oriented from topic or attribute value has already at least one corresponding icon. Users always find another symbol with a better comprehensibility and designers have also the fresh idea of a more artistic representation. Instead of the fear that new icons will not adapt to the actual system, VDL-based icon system provides a simple way. All what to do is joining proposed symbol onto fixed pre-icon to create the new icon. Even if symbol changes, predefined pre-icon makes it remind in the same category and the categorization of iconic tags maintains. This operation allows more alternatives for a topic or for an attribute value to respond to large participation of icon proposition.

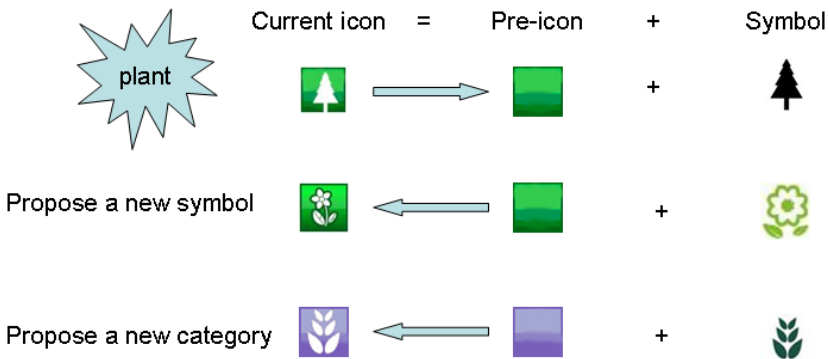


Fig. 7. Cooperative operations on the combination of pre-icon and symbol to propose new VDL-based icons

Secondly, an existing icon could be proposed to be associated with other categories. When an icon represents an object with multi-topics, it may be correlated with more than one category. In predefined icon categorization, experts cannot recognize all the multi-topic-icons like "renewable energy", resulting from which, users will be recommended later to re-produce an icon from another pre-icon with actual symbol. This operation encourages a collective recognition of icons. Each user is able to recognize a

reasonable concerned category for an existing icon. This multi-topic-icons proposition is assumed to make more accurate knowledge tagging because it is easy to realize which domain is highlighted seeing from the pre-icon of multi-topic-iconic tag.

Thirdly, it is meaningful to suggest an entirely new entity to VDL-based icon system. This operation can be explained as inserting represented object along with a corresponding icon which is similar to adding topics and attribute values in a KOS. Experts on knowledge management and users could also create a new iconic tag with merely proposing its textual form and calling for corresponding iconic representation from those being good at designing. However, these textual tags are rather the new items in each category than the possible subtitles of existing icons. For sure that new icon will always keep the fixed pre-icon. This type of textual proposition will not be accepted by administrator until an icon matches with.

Finally, users are able to give comments to icons principally about their design and applicable area. Users may remark on the understandability of symbol referring to the quality of iconic interpretation, or give advice on the possible textual explanation of icons. This proposition on textual meaning implies the type of knowledge to be tagged. For example, an icon of a symbolic factory may have several corresponding textual tags like "industry", "factory", even "pollution". If someone gets these subtitles proposed by others, he will probably use this icon to tag knowledge interested in these topics. Although each person has his own understanding of icon and will apply the same icon into varied domains, others' propositions constrains somehow a tendency of applicable area since there is semantic imitation in social tagging [18].

These four operations are dedicated to the co-construction of VDL-based icon system, which will enhance its functions in both top-down and bottom-up way. They avoid the limitation of expert-recommended icon system by absorbing as much as possible ideas from all participants. Meanwhile, by means of VDL, new joint icons will not disturb the actual system structure.

3.4 Discussion: From Icon System to KOS

Although VDL-based icon system acts as a supplementary system of KOS where icons are treated as future iconic tags, it is also able to function like an independent KOS in which icons are entities. Here Hypertopic is proposed as the protocol to manage icons by its topics, attributes and resources, as what is done to documents.

Above all, the topics of icons reveal the icon categorization of representing objects. This categorization completely employs the tag structure proposed by Hypertopic in which each category of symbolic interpretations of icons refers to a group of viewpoint or attribute name. As introduced before, icon categorization based on graphical characters and symbolic characters are unified in the case of VDL-based icon system. Adding or searching an icon is similar to the procedure of a target textual item in KOS.

Besides, attributes and resources are also attachable for an icon. A list of attributes of an icon may include its created time, pixel information, designer, proposed applying field and other useful information. Resources for an icon will relate to the link for uploading or the documents tagged by this icon before to show previously applied examples. Practical evidence of icon benefit from cooperative activities like proposing icons representing close objects with their link in the resource, or putting some comments in the attributes to suggest applicable domain.

As iconic tags are able to tag an item, they are as well used as labels for an icon. In this way, icon to be tagged is treated as an item in KOS while icons used to tag are iconic tags in another icon system. However, once a VDL-based icon system is implemented for a social sharing with a fixed knowledge context, it is clearer to tag an icon with textual tags to avoid confusion. Tagging icons with iconic tags is a conceptually achievable task but still cause other problems on readability.

Taking evidence from successful samples of KOS based on Hypertopic protocol, VDL-based icon system constructed by the same theoretical principle is also able to be managed as a KOS with special focus on knowledge tagging in social communities.

4 Conclusion

After validating that VDL-based iconic tag system improves the effectiveness of tagging process, we presented how to co-construct it for social tagging in a given knowledge context. This system underlines an integrated Hypertopic-based icon categorization in both symbolic and graphical ways, attributes list for each icon and four cooperative operations of users. What's more, VDL-based icon system is easily adaptable for kinds of knowledge tagging and sharing with a simple creation rule on pre-icon plus symbol. Its advantage caters for current social information goal and diverse knowledge contexts. To summarize our contribution is that the cooperative construction provides an easier and more efficient way to create VDL-based icon system, in which structured iconic tags make it possible to visually and verbally organize the knowledge within social communities.

References

1. Hodge, G.: Systems of Knowledge Organization for Digital Libraries. Beyond Traditional Authority Files. The Council on Library and Information Resources, Washington, DC (2000)
2. Furnas, G.W., Fake, C., Von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C., Naaman, M.: Why do Tag systems Works? In: SIGCHI Conference on Human Factor in Computing Systems (2006)
3. Paivio, A.: Mental representations: a dual coding approach. Oxford University Press, Oxford (1986)
4. Lohse, J., Rueter, H., Biolsi, K., Walker, N.: Classifying Visual Knowledge Representations: a Foundation for Visualization Research. In: IEEE Computer Society Technical Committee on Computer Graphics, Siggraph: ACM Special Interest Group on Computer Graphics and Interactive Techniques, pp. 131–138 (1990)
5. Collins, B.L., Lerner, N.: Assessment of Fire Safety Symbols. *Human Factors* 24(1), 75–84 (1982)
6. Lamy, J.B., Duclos, C., Bar-Hen, A., Ouvrard, P., Venot, A.: An Iconic Language for the Graphical Representation of Medical Concepts. In: *BMC Medical Informatics and Decision Making* (2008)
7. King, A.J.: On the Possibility and Impossibility of a Universal Iconic Communication System. In: Yazdani, M., Barker, P. (eds.) *Iconic Communication*, pp. 17–28. Intellect, Bristol (2000)

8. Ma, X., Cahier, J.P.: Iconic Categorization with Knowledge-Based Icon Systems can Improve Collaborative KM. In: CTS 2011, pp. 216–223. IEEE Conference Publications (2011)
9. Ma, X., Cahier, J.P.: Visual Distinctive Language: using a Hypertopic-based Iconic Tag system for Knowledge Sharing. In: IEEE 21st International WETICE, pp. 456–461 (2012)
10. Béné, A., Zhou, C., Cahier, J.P.: Beyond Web 2.0...and Beyond the Semantic Web. Design of Cooperative Systems, ch. 1. Springer (2009)
11. Zhou, C., Lejeune, C.H., Béné, A.: Towards a Standard Protocol for Community Driven Organizations of Knowledge. In: ISPE Conference on Concurrent Engineering, pp. 338–349. IOS Press (2006)
12. Cahier, J.P., Ma, X., Zaher, L.H.: Document and Item-based Modeling: a Hybrid Method for Socio-Semantic Web. In: ACM Symposium on Document Engineering, DocEng, pp. 243–246 (2010)
13. Mas, S., Marleau, Y.: Proposition of a Faceted Categorization Model to Support Corporate Information Organization and Digital Records Management. In: 42th Hawaii International Conference on System Sciences, pp. 1–10 (2009)
14. Bertin, J.: *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press (1983); W.J. Berg (Translator)
15. Herrmann, T., Hoffmann, M., Loser, K.-U., Moysich, K.: Semistructured models are surprisingly useful for user-centered design. In: Dieng, R., Giboin, A., Karsenty, L., De Michelis, G. (hrsg.): *Designing Cooperative Systems*, pp 159–174. IOC Press, Amsterdam (2000)
16. Wang, H., Hung, S., Liao, C.: A Survey of Icon Taxonomy Used in the Interface design. In: 14th European Conference on Cognitive Ergonomics, pp. 203–206 (2007)
17. Nakamura, C., Zeng-Treitler, Q.: A Taxonomy of Representation Strategies in Iconic Communication. *J. Human-Computer Studies* 70, 535–551 (2012)
18. Fu, W., Kannampallil, T., Kang, R., He, J.: Semantic Imitation in Social Tagging. *J. ACM Transaction on Computer-Human Interaction* 17(3), Article 12, 1–37 (2010)

A Multi-dimensional and Event-Based Model for Trust Computation in the Social Web

Barbara Carminati, Elena Ferrari, and Marco Viviani

Università degli Studi dell'Insubria
Dipartimento di Scienze Teoriche e Applicate (DiSTA)
Via J. H. Dunant, 3 – 21100 Varese – Italia
{barbara.carminati,elena.ferrari,marco.viviani}@uninsubria.it

Abstract. In this paper, we propose a general-purpose Trust Layer that fits and exploits the emerging concept of Social Web. Key features of our proposal are the consideration of several dimensions for trust computation and the exploitation of social interaction dynamics over the Web, through the definition and the evaluation of event patterns and trust rules. Besides presenting our trust model, we discuss a case study on the ACM Digital Library social environment.

Keywords: Trust, Multi-dimensional Social Trust, Social Web.

1 Introduction

Estimating trust of Web users is today one of the most challenging research issues. Although many proposals have so far emerged to capture trust as the extent to which one party is willing to depend on something or somebody (see Section 5 for an overview), considering connected benefits and risks [18,29], most of them are focused on specific scenarios (e.g. recommender systems, e-commerce, social networks), or they focus only on some of the issues related to trust computation (e.g., the computation of transitive trust). However, we believe that this vision of domain-dependent trust computation does not fit anymore into the current vision of the Web, which is rapidly evolving into the concept of *Social Web*, that is, the set of multiple types of relationships that link together people over the Internet [4], crossing the boundaries of the specific services they are using and their related technologies. This evolution is also witnessed by the fact that major online social networks such as Facebook, Google, Myspace and, in a different way, Twitter, as well as e-commerce websites and mobile applications, are following the idea that users and their resources can be represented on complex graphs [6,21,36] connecting different entities with different kinds of relationships. This is instrumental to provide users with the possibility to increase both their interactions with other users, as well as resource sharing in the most possible *personalized* and *semantic-meaningful* way.

Technically speaking, Facebook has developed Open Graph, an extension of its social graph via the Open Graph protocol¹. This is an RDFa-based protocol

¹ <http://ogp.me>

enabling any Web page to become a rich object in a social graph by adding to it basic RDFa metadata. In the same way, Google and Myspace (together with a number of other social networks) are following the OpenSocial² public specification, which defines a component hosting environment (container) and a set of common application programming interfaces for web-based applications. In this scenario, intuition and literature [22,24] suggest us that a key dimension for trust computation is to keep into account the social graph dynamics, that is how social graphs evolve over the time, according to “social events” (e.g., *add, modify, delete* a relationship) generated by the social community.

For these reasons, in this paper we propose our *Multi-dimensional and Event-based Trust Layer*. Multi-dimensional, in the sense that we exploit for trust computation *multiple types of relationships* representing diverse interactions among users and resources in a social scenario. Event-based, because a certain trust relationship holds in a certain dimension when an *event* or some *event patterns*, meaningful for that specific dimension, occur. Besides presenting the main components of the Trust Layer, we also present some preliminary performance results, showing the effectiveness of our proposal on top of the social environment we extract from the ACM Digital Library dataset.³

The rest of the paper is organized as follows. Section 2 overviews the architecture of the Trust Layer and discusses its properties. Section 3 provides a formalization of the Trust Layer components. In Section 4 we show the use of the Trust Layer applied to the AMC Digital Library social environment. Section 5 surveys related work. Finally, Section 6 concludes the paper.

2 A Multi-dimensional and Event-Based Trust Layer

Our idea is to build a Multi-dimensional and Event-based Trust Layer on top of any social environment via an *augmented social graph* [8] able to aggregate all information gathered from the Social Web concerning users and their resources (e.g., actions, opinions, user profile attributes), in order to evaluate users’ trust relationships. As introduced before, the *evolution* of the augmented social graph can be exploited to compute trust between users. To keep trace of the augmented graph evolution and to evaluate trust accordingly, we take inspiration from *Event-Driven Architectures* (EDA) [34], where certain actions are triggered as a response to the detection of particular *events* or *event patterns*. *Complex Event Processing* (CEP) [26], which detect interesting events or event patterns in data streams and react to them in presence of critical situations, can be exploited for the purpose of trust computation. The idea is, therefore, to (i) gather from the augmented social graph all the events that change the social interactions on the graph (i.e., edges creation/deletion/modification), (ii) encode them into streams, and (iii) evaluate over them a set of meaningful event patterns. Finally, (iv) these patterns are evaluated by some trust rules, that associate with involved users a given trust value when some meaningful event patterns occur.

² <http://docs.opensocial.org>

³ <http://dl.acm.org/>

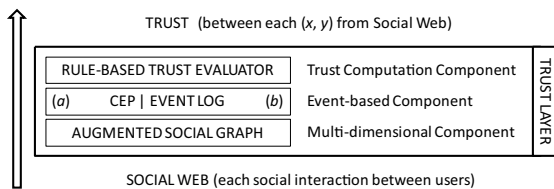


Fig. 1. The architecture of the Trust Layer

Trust rules are monitored in the Trust Layer by a *Complex Event Processing engine* (see Figure 1, component (a)), to immediately detect changes on the augmented social graph that implies a new trust value for involved users. This architecture has the strong benefit of a real-time estimation of users trust values. This might be fundamental in some scenarios where trust is a key parameter in the decision process. This is, as an example, the case of relationship-based access control [14], where information releasing in a social network is regulated by the existing trust relationships (e.g., a user authorizes the access to a given resource only to friends with a trust value greater than 0.6). However, if we consider the huge amount of possible changes in a social environment, this architecture might imply an high overhead due to the continuous event monitoring and evaluation of trust rules. As such, as an alternative, we also consider an architecture where events are collected into an *event log* (see Figure 1, component (b)), over which trust rules are periodically evaluated.

It is relevant to note that the selection of the CEP-based or log-based architecture does not change the formal representation of the augmented social graph, event patterns and trust rules. The only difference between the two architectures is indeed in terms of trust rules evaluation. The choice of an architecture with respect to the other, depends on a trade-off between efficiency and the risk of not having up-to-date trust information. Since the formalization is the same, and our experiments have been conducted on a log-based architecture, hereafter we refer to trust rules evaluated over event logs.

For the sake of clarity and for evaluation purposes, we will refer in this paper to the ACM Digital Library dataset. The ACM Digital Library is a database created by the Association for Computing Machinery (ACM) containing every article ever published by ACM, information about its authors, and bibliographic citations from major publishers in computing. We will consider social aspects connected to this database, in particular the augmented social graph that it is possible to extract from coauthoring information and other information such as users' affiliation, conference venues, and publication years.

3 Trust Layer Modeling

3.1 Multi-dimensional Component

We start introducing the basic concepts necessary to formalize the augmented social graph. Hereafter, we denote with $U = \{u_1, u_2, \dots, u_n\}$ and $R = \{r_1, r_2, \dots, r_n\}$

the *set of users* and the *set of resources* in a given social environment, respectively. We consider as resources both ‘classical’ user profile information, such as name, surname, affiliation, etc. and ‘concrete objects’, as photos, videos, posts, etc.

Each user $u_i \in U$ and each resource $r_k \in R$ can be connected to other users, or to other resources, via a certain *relationship*. The *set of relationship names* supported in a scenario, that we denote as $RN = \{\eta_1, \eta_2, \dots, \eta_\infty\}$, is potentially infinite, since in our model it is given the possibility to introduce new kinds of relationships. Being RN be the set of relationship names, a *relationship type* ρ_p is formally defined as:

$$\rho_p = \langle \eta_p, \sigma_p, \varsigma_p \rightarrow \tau_p \rangle \quad (1)$$

where $\eta_p \in RN$ is the *relationship name*, σ_p is a (facultative, assigned by a domain expert) *trust judgement* connected to the semantics describing the relationship,⁴ ς_p represents the *source entity type* and τ_p represents the *target entity type*. A source and target entity type can be both a *user type* (\mathbf{u}) or a *resource type* (\mathbf{r}).

The creation of a new relationship in the augmented social graph implies the generation of an instance of the corresponding relationship type. Being RT the *set of relationship types* and $\rho_p \in RT$ be a relationship type, we define ρ_p^r as *instance* of ρ_p when we associate with the relationship ρ_p a *concrete source entity* s_p^r and a *concrete target entity* t_p^r . Formally:

$$\rho_p^r = \langle \eta_p, \sigma_p, s_p^r \rightarrow t_p^r \rangle \quad (2)$$

where, if $\varsigma_p = \mathbf{u}$, then $s_p^r \in U$ (if $\varsigma_p = \mathbf{r}$, then $s_p^r \in R$); if $\tau_p = \mathbf{u}$, then $t_p^r \in U$ (if $\tau_p = \mathbf{r}$, then $t_p^r \in R$). We refer to IR as the *set of instantiated relationships*.

Once defined the concepts of user, resource, relationship type and instance of a relationship type, we are now ready to define the *augmented social graph* as

$$G = \langle V, E \rangle \quad (3)$$

where: (i) $V = V_U \cup V_R$ is the *set of vertexes*, where $V_U = U$ and $V_R = R$; (ii) $E = IR$ is the *set of edges* connecting vertexes between them.

Example 1. Figure 2 illustrates an example of the graph G instantiated on the ACM Digital Library social environment, where $RN = \{\eta_1 : \text{collaborate}, \eta_2 : \text{publish}, \eta_3 : \text{venue}, \eta_4 : \text{year}\}$ and $RT = \{\rho_1 : \langle \text{collaborate}, \text{positive}, \mathbf{u} \rightarrow \mathbf{u} \rangle, \rho_2 : \langle \text{publish}, \text{positive}, \mathbf{u} \rightarrow \mathbf{r} \rangle, \rho_3 : \langle \text{venue}, \text{neutral}, \mathbf{r} \rightarrow \mathbf{r} \rangle, \rho_4 : \langle \text{year}, \text{neutral}, \mathbf{r} \rightarrow \mathbf{r} \rangle\}$.⁵ The set of user vertexes V_U consists of four users: *authors* a, b, c and d . The set or resource vertexes V_R consists of: (i) eight *papers* p_1, \dots, p_8 ; (ii) four *conferences* c_1, \dots, c_4 ; (iii) five *years of publication* y_1 to y_5 . The set E of edges contains the instances of $\rho_1, \rho_2, \rho_3, \rho_4$, as emerges from the figure.

⁴ $\sigma_p \in \{\text{positive}, \text{negative}, \text{neutral}\}$, where to *positive*, *negative* and *neutral* can be assigned ranges of values in the interval $[-1, 1]$, depending on the specific domain.

⁵ A *positive* judgement is assigned to the *collaborate* relationship since it represents mutual and strict interaction between users, clearly based on trust [9, 32]. We leave as *neutral* the fact of having published in a certain year or in a specific conference since we can not statically evaluate this information in terms of trust.

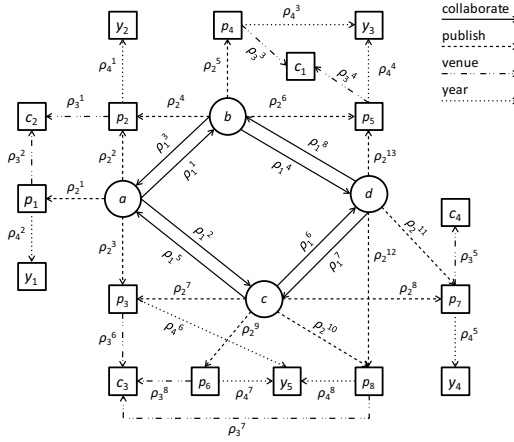


Fig. 2. An example of augmented social graph for the ACM Digital Library scenario

3.2 Event-Based Component

Literature offers several options for events formalization, in particular the work described in [3, 11], which are however not related to trust computation. For this reason, building on these, we define the specific concepts underlying the event-based component of our Trust Layer.

An *atomic event* \hat{e} is an instance of a type \hat{E} formalized as follows:

$$\hat{E}(t, \nu, \varrho, s_p^r, t_p^r, \eta_p, \sigma_p) \tag{4}$$

where \hat{e} is a predicate symbol (the name of the atomic event). The *set of properties* $\hat{P} = \{t, \nu, \varrho, s_p^r, t_p^r, \eta_p, \sigma_p\}$ associated with the event type \hat{E} represents a sort of *intensional description* of the atomic event \hat{e} itself. In addition to the parameters introduced before, it contains: t , the timestamp associated with the atomic event; ν , the *type* of the atomic event, i.e., $\nu \in \{add, delete, modify\}$; and ϱ , the *responsible* of the atomic event, i.e., the subject having generated the relationship. We assume an *event log*⁶ containing the set $\hat{\mathcal{E}} = \{\hat{e}_1, \hat{e}_2, \dots, \hat{e}_m\}$ of atomic events involving users and their resources. In order to have a direct access to each single parameter belonging to an atomic event \hat{e}_l , we use the notation $\hat{e}_l.parameter$.

Atomic events can be monitored through *simple* and *complex event patterns*. A *simple event pattern* \dot{e} is formalized as:

$$\dot{e}(v_t|v_\nu|v_\varrho|v_{s_p^r}|v_{t_p^r}|v_{\eta_p}|v_{\sigma_p}) \tag{5}$$

where $v_{\hat{p}}$ denotes the value of the parameter $\hat{p} \in \hat{P}$ associated with \hat{E} . The specification of $v_{\hat{p}}$, for any \hat{p} , is facultative. A simple event pattern allows to

⁶ We recall that we focus on the log-based architecture of the Trust Layer.

define a sort of query able to ‘capture’ each atomic event $\hat{e}_l \in \hat{\mathcal{E}}$ whose parameters values match with the ones expressed in the query.

Being $\dot{E} = \{\dot{e}_1, \dot{e}_2, \dots, \dot{e}_{n_e}\}$ the *set of simple event patterns*, a *complex event pattern* \ddot{e} is formalized as:

$$\ddot{e}(\varphi_1 \star \varphi_2 \star \dots \star \varphi_n) \quad (6)$$

where $\varphi_i \in \dot{E}$, $i \in \mathbb{N}$, and $\star \in \{\wedge, \vee, \neg, \in, \notin, \subset, \supset, \subseteq, \supseteq, \triangleright, \triangleleft\}$ ⁷

Example 2. Considering the ACM Digital Library scenario, if we are interested in capturing events where author a has established collaborations in each year between 1996 and 1998, we have to define the simple event patterns: $\dot{e}_1(1996|add|a|a| \star |collaborate| \star)$, $\dot{e}_2(1997|add|a|a| \star |collaborate| \star)$, $\dot{e}_3(1998|add|a|a| \star |collaborate| \star)$ and the complex event pattern: $\ddot{e}(\dot{e}_1 \wedge \dot{e}_2 \wedge \dot{e}_3)$.

Before introducing *trust event patterns*, we introduce the notion of *condition pattern* $C(e)$ on e as a boolean expression of *atomic conditions* $c(e)$. An *atomic condition* is formalized as:

$$c(e) : exp_1 \diamond exp_2 \quad (7)$$

where $\diamond \in \{=, \neq, >, <, \geq, \leq, \subset, \supset, \subseteq, \supseteq, \in, \notin\}$. Furthermore:

- if e is a simple event pattern: exp_1 is a parameter $\hat{p} \in \hat{P}$, whereas exp_2 can be a **value**, another parameter in \hat{P} , or a set of atomic events;⁸
- if e is a complex event pattern:
 - exp_1 can be a parameter in \hat{P} acting *globally* on the complex event pattern;⁹ or a parameter in \hat{P} acting *locally* on a component simple event pattern \dot{e} ;¹⁰
 - exp_2 can be a **value**, or a parameter $\hat{p} \in \hat{P}$ (global or local), or a set of atomic events.

A *trust event pattern* \tilde{e} is a formula of the following form:

$$\tilde{e}(e, C(e)) \quad (8)$$

where e can be: (i) an atomic event, (ii) a simple event pattern, or (iii) a *complex event pattern*, whereas $C(e)$ is an optional condition pattern acting on e . The condition pattern *is always unspecified*, if e is an atomic event (i.e., $e = \hat{e}$). Associating conditions with trust event patterns, allows us to evaluate further constraints acting on simple and complex event patterns. Conditions act at different levels of granularity and generality depending on if we are dealing with a simple or a complex event pattern.

⁷ Where $\triangleright = \textit{before}$, and $\triangleleft = \textit{after}$.

⁸ The set of atomic events can be given explicitly or it can be the result of a projection operation, that will be described later.

⁹ Acting on each simple event pattern constituting the complex event pattern.

¹⁰ Via the *e.parameter* notation.

Example 3. We can capture events where author a has established collaborations between 1996 and 1998 adding a condition acting on a simple event pattern defining: $\dot{e}(*|add|a|a|*|collaborate|*)$, $c_1(\dot{e}) : t \geq 1996$, $c_2(\dot{e}) : t \leq 1998$. The resulting trust event pattern is defined as: $\tilde{\epsilon}(\dot{e}, c_1(\dot{e}) \wedge c_2(\dot{e}))$.

If we are interested in capturing collaborations of a and b with others after 1998, we specify a condition acting globally on the time parameter associated with all the corresponding simple event patterns, i.e., $\dot{e}_1(*|add|a|a|*|collaborate|*)$, $\dot{e}_2(*|add|b|b|*|collaborate|*)$, $\ddot{e}(\dot{e}_1 \wedge \dot{e}_2)$, $c(\ddot{e}) : t > 1998$. The resulting trust event pattern is defined as: $\tilde{\epsilon}(\ddot{e}, c(\ddot{e}))$.

Finally, we can capture events where a has collaborated with c after that he/she has already collaborated with b , adding a condition acting locally on the time parameters associated with the single event patterns constituting a complex event pattern, i.e., $\dot{e}_1(*|add|a|a|b|collaborate|*)$, $\dot{e}_2(*|add|a|a|c|collaborate|*)$, $\ddot{e}(\dot{e}_1 \wedge \dot{e}_2)$, $c(\ddot{e}) : \dot{e}_1.t < \dot{e}_2.t$. The resulting trust event pattern is defined as: $\tilde{\epsilon}(\ddot{e}, c(\ddot{e}))$.

Given a trust event pattern $\tilde{\epsilon}(e, C(e))$, we denote via the *projection* notation:

$$\Pi(\tilde{\epsilon}(e, C(e))) \quad (9)$$

the operator retrieving the set of atomic events satisfying the event pattern, and with the *cardinality* notation:

$$||\tilde{\epsilon}(e, C(e))|| \quad (10)$$

the operator retrieving the number of atomic events in $\Pi(\tilde{\epsilon}(e, C(e)))$. Projection can also be applied to each parameter, i.e., $\Pi.t, \Pi.\nu, \Pi.\rho, \Pi.s_p^r, \Pi.t_p^r, \Pi.\eta_p, \Pi.\sigma_p$, obtaining the set of the values associated with the single parameter satisfying the event pattern.

Example 4. Let us consider the subset of atomic events involving instantiated collaborate relationships of Figure 2 (where $y_2 = 1996$, $y_3 = 1998$, $y_4 = 2000$, $y_5 = 2004$), i.e., $\hat{\mathcal{E}} = \{\hat{e}_1(1996, add, a, a, b, collaborate, positive), \hat{e}_2(1996, add, b, b, a, collaborate, positive), \hat{e}_3(1998, add, b, b, d, collaborate, positive), \hat{e}_4(1998, add, d, d, b, collaborate, positive), \hat{e}_5(2000, add, c, c, d, collaborate, positive), \hat{e}_6(2000, add, d, d, c, collaborate, positive), \hat{e}_7(2004, add, a, a, c, collaborate, positive), \hat{e}_8(2004, add, c, c, a, collaborate, positive)\}$.

Via the projection notation, we show the set of atomic events satisfying the event patterns: $\Pi_1(\dot{e}(1996|add|*|*|*|collaborate|*)) = \{\hat{e}_1, \hat{e}_2\}$, $\Pi_2(\dot{e}(*|add|a|*|*|collaborate|*)) = \{\hat{e}_1, \hat{e}_7\}$, $\Pi_3(\tilde{\epsilon}(\dot{e}(*|add|a|*|*|collaborate|*), t > 2000)) = \{\hat{e}_7\}$.

The following is an example of the use of the cardinality notation applied to a simple event pattern: $||\dot{e}(1996|add|*|*|*|collaborate|*)|| = 2$.

If we apply projection to single parameters we obtain: $\Pi_1.\nu = \{a, b\}$, $\Pi_2.t = \{1996, 2004\}$, $\Pi_3.t = \{2004\}$. If we apply cardinality, we obtain: $||\Pi_1.\nu|| = 2$, $||\Pi_2.t|| = 2$, $||\Pi_3.t|| = 1$.

We will see in Section 4 the utility of the introduction of the concepts of projection and cardinality for the definition of trust event patterns and condition patterns.

3.3 Trust Computation Component

The last component of our model is composed of *trust rules* that, applied to trust event patterns, returns the *trust values* to associate to users involved in it. Being $\tilde{e}(e, C(e))$ a trust event pattern and $x, y \in U$, a *trust rule* \tilde{r}_u is defined as:

$$\tilde{r}_u(\tilde{e}(e, C(e))) \rightarrow tv(x, y) \quad (11)$$

where $tv(x, y)$ is the *trust value* associated with (x, y) via the evaluation of the event pattern \tilde{e} by \tilde{r}_u . We denote with $\tilde{\mathfrak{R}}$ the *set of trust rules* defined in the system. The trust value $tv(x, y)$ is computed differently depending on whether the event pattern is composed by (i) an atomic event or (ii) a simple or complex event pattern. In case (i), the trust value of the atomic event is equal to the trust judgement (if it exists) associated with the semantics describing the involved action. Formally:

$$\tilde{r}(\tilde{e}(\hat{e}, C(\hat{e}))) = \tilde{r}(\hat{e}) \rightarrow tv(x, y) = \hat{e}.o \quad (12)$$

since $C(\hat{e})$ is unspecified for an atomic event.

Computing trust values associated with trust rules evaluating simple or complex event patterns (case (ii)) is a strictly domain-dependent task. In fact, due to the multi-dimensionality of users relationships in a social scenario, trust rules can extract trust relationships among users only if associated to event patterns that are meaningful in a specific context. For this reason, we detail this aspect in the next section on the concrete scenario of the ACM Digital Library dataset.

4 Trust Layer and the ACM Digital Library Dataset

We focus on the ACM Digital Library social environment, which models the evolution of collaborations between authors based on the year of their first joint publication. In such a scenario, we expect that new established collaborations are based on trust among authors [9, 32]. To catch this correlation in our experiments, our idea is to analyze the evolution of the augmented social graph (i.e., of new established collaborations built from the ACM Digital Library dataset), to see how this depends on trust. More precisely, we perform several experiments so as to associate a trust value with each couple of authors (a, b) that have not yet collaborated at a given time T ¹¹ so as to verify if they later on have established a collaboration. We expect a new collaboration if the trust value between not yet collaborating authors is greater than a given threshold α . If this happens, we say that our trust rule *captures the new collaboration*. To this purpose, we defined three trust rules whose trust event patterns take into account different events involving coauthors of a and b in a certain period of time. The ratio underlying the experimented trust rules is explained in the next sections. It is important to note that the experimented trust rules have been defined based on a set of assumptions. Here, we do not want to show that the proposed trust rules are the best possible ones. Instead, we intend

¹¹ In the ACM Digital Library, the time T of creation of a new co-authorship relationship is given by the year of publication of the co-authored paper.

to show that, given a scenario based on some properties that an expert judges relevant for trust computation on it (e.g., the number of mutual collaborators), the components of our Trust Layer are expressive enough to capture these properties and to associate proper trust values with the involved users.

1. First of all, we expect that for each couple of authors (a, b) not having collaborated yet, the higher is the number of coauthors that a has not in common with b , the lower is the level of trust that a has in b . This because we expect that a could decide not to write a paper with b , in the case b is an author having collaborated with very few of a 's (trusted) coauthors.
2. Furthermore, we expect that, being equal the number of mutual collaborators between a and b , the higher is the number of those, among the mutual collaborators of a and b , having established a collaboration relationship, the higher is the trust between a and b . We correlate, this way, a high social connectivity with a high level of trust.
3. As a last property, we expect that the higher is the difference between a and b in terms of number of publications with the mutual collaborators (a has published more with the mutual collaborators with respect to b), the lower is the level of trust that a has in b . Similarly to property 1, even in this case, a could decide not to collaborate with b , if b is an author that, in general, publishes very few papers with a 's trusted coauthors.

4.1 Trust Computation

We develop three formulas (each of which respecting the corresponding property described before) to compute the trust value $tv(a, b)$ between each couple of authors (a, b) , which will be then associated with three different event patterns:

$$\text{F1. } tv(a, b) = \frac{\sigma(a, b) \cdot cc(a, b)}{c(a)} \quad (13)$$

$$\text{F2. } tv(a, b) = \begin{cases} \left(\frac{\sigma(a, b) \cdot cc(a, b)}{c(a)} \right)^{\left(1 - \frac{\sigma(a, b) \cdot cccc(a, b)}{cc(a, b)} \right)} & \text{if } cccc(a, b) < cc(a, b) \\ \left(\frac{\sigma(a, b) \cdot cc(a, b)}{c(a)} \right)^{\left(\frac{\sigma(a, b) \cdot cc(a, b)}{cccc(a, b)} \right)} & \text{otherwise} \end{cases} \quad (14)$$

$$\text{F3. } tv(a, b) = \begin{cases} \left(\frac{\sigma(a, b) \cdot cc(a, b)}{c(a)} \right)^{\left(1 - \frac{cp(b, a)}{cp(a, b)} \right)} & \text{if } cp(b, a) < cp(a, b) \\ \left(\frac{\sigma(a, b) \cdot cc(a, b)}{c(a)} \right)^{\left(\frac{cp(a, b)}{cp(b, a)} \right)} & \text{otherwise} \end{cases} \quad (15)$$

where

- $\sigma(a, b)$ is the *trust judgement* associated with the semantics describing the relationship between a and b ;
- $cc(a, b)$ is the *number of mutual collaborators* between a and b ;
- $cccc(a, b)$ is the *number of authors, among the mutual collaborators* of a and b , *having established a collaboration relationship*;

- $c(a)/c(b)$ is the *number of coauthors of a/b*;
- $cp(a,b)/cp(b,a)$ is the *number of publications of a/b with collaborators common to b/a*.

4.2 Events and Rules Specification

In order to compute the above defined formulas for collaboration prediction based on trust, we have (i) to associate a specific rule with them; and (ii) to express them in terms of event patterns. The rule must, at a given time, associate a trust value with each couple of authors (not having collaborated yet) on the basis of the addition of collaborations between other authors in the data set. This means to formally define:

- the simple event pattern capturing all the collaborations:
 $\dot{e}_1(*|add|*|*|*|collaborate|*)$;
- the simple event patterns capturing collaborations between a and b :
 $\dot{e}_2(*|add|a|a|b|collaborate|*)$, $\dot{e}_3(*|add|b|b|a|collaborate|*)$;
- the complex event patterns verifying the nonexistence of collaborations between a and b : $\ddot{e}_1(\dot{e}_2 \notin \dot{e}_1)$, $\ddot{e}_2(\dot{e}_3 \notin \dot{e}_1)$ in a certain period of time: $c_1(\ddot{e}) : t_1 \leq t \leq t_2$;
- the trust event pattern: $\tilde{e}_1(\ddot{e}_1 \wedge \ddot{e}_2, c_1(\ddot{e}))$;
- the trust rule: $\bar{r}_1(\tilde{e}_1) \rightarrow tv(a, b)$.

With \bar{r}_1 we can now associate one of the three formulas (13), (14) and (15). This means to express the components of the formulas in terms of event patterns.

- To compute $c(a)$ and $c(b)$:
 $\dot{e}_4(*|add|a|a|*|collaborate|*)$, $\dot{e}_5(*|add|b|b|*|collaborate|*)$, $\tilde{e}_2(\dot{e}_4, \emptyset)$, $c(a) = \|\tilde{e}_2\|$, $\tilde{e}_3(\dot{e}_5, \emptyset)$, $c(b) = \|\tilde{e}_3\|$.
- To compute $cc(a, b)$:
 $\ddot{e}_3(\dot{e}_4 \wedge \dot{e}_5)$, $c_2(\ddot{e}) : \dot{e}_4.t_p^r = \dot{e}_5.t_p^r$, $\tilde{e}_4(\ddot{e}_3, c_2(\ddot{e}))$, $cc(a, b) = \|\tilde{e}_4\|$.
- To compute the trust value associated with the action ‘collaborate’:
 $\sigma(a, b) = \Pi.\dot{e}_1.\sigma$.
- To compute the number of common collaborators of common collaborators between a and b : $\tilde{e}_5(\dot{e}_1, \dot{e}_1.t_p^r \in \Pi.\tilde{e}_4.t_p^r)$, $cccc(a, b) = \|\tilde{e}_5\|$.
- To compute the number of publications of a with collaborators common to b , and viceversa: $\dot{e}_6(*|add|a|a|*|publish|*)$, $\dot{e}_7(*|add|b|b|*|publish|*)$, $\tilde{e}_6(\dot{e}_6, \dot{e}_6.t_p^r \in \Pi.\tilde{e}_3.t_p^r)$, $\tilde{e}_7(\dot{e}_7, \dot{e}_7.t_p^r \in \Pi.\tilde{e}_2.t_p^r)$, $cp(a) = \|\tilde{e}_6\|$, $cp(b) = \|\tilde{e}_7\|$.

4.3 Evaluation

From the ACM Digital Library dataset, we have extracted a subset of data in a thirteen years interval, from 1996 to 2010. In this time interval, we dispose of 283 authors and 6,477 collaborations between them. As introduced before, it is a static a posteriori dataset. Since it is our aim to evaluate the number of collaborations which has been established based on trust, we have, for each year, considered authors not having collaborated before. For each couple of them, we evaluate their trust level using the trust rule introduced in Section 4.2, associated with the three formulas for trust computation described in Section 4.1. We

used, as a trust threshold α , a value of 0.6, representing a ‘sufficient’ amount of trustworthiness between users with a $\sigma = 0.9$ value connected to the semantics of the ‘collaborate’ relationship, since this kind of relationship represents a high degree of correlation between users. Figure 3 illustrate the correlation between the computed trust values and the generation of new collaborations. As Figures

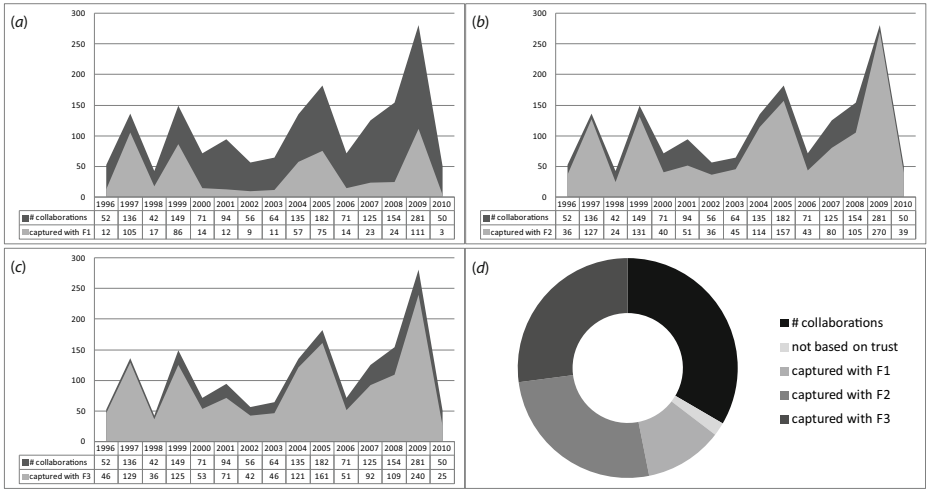


Fig. 3. Correlation between trust and new collaborations with F1. (a), F2. (b), F3. (c) and global correlation from 1996 to 2010 (d).

3 (a), (b) and (c) clearly show, by using F2 or F3, we have better results with respect to F1. This means that, by considering trustworthiness between users based on common collaborators, a clearer correlation between trust and new collaborations emerges as we exploit additional information connected to users with respect to F1, such as the number of common collaborators of the common collaborators, or the exact number of publications between common collaborators. Independently from the chosen formula, if we consider the complete time interval from 1996 to 2010, we can see in Figure 3 (d) how the number of collaborations created in absence of trust (i.e., collaborations that have not been captured by our trust rules) is extremely low. On the contrary, it clearly emerges the correlation (lower or higher, depending on the chosen formula) between trust and the generation of new collaborations.

5 Related Work

Trust has been widely investigated in different fields and for different aims, such as security and access control [7,20,30], peer-to-peer systems [1,5,10,37], agent systems [15,28,35], online service provision and recommender systems [17,18,31].

Nowadays, in the scenario of the Social Web, relationships between social behavior and trust, have to be taken into consideration in the process of trust management. In this sense, trust modeling and computation has been explored in the context of online social networks [12, 16, 23, 27]. Most of these techniques are based on probabilistic approaches, and on the concept of trust transitivity among users [19]. However, as discussed in [13], trust propagation is far from being a trivial task, and it is not always applicable. It is fundamental to understand which are the assumptions under which trust can propagate transitively, and in which context. Research in this direction has been done in [25], a work proposing a complex social network structure taking into account trust, social relationships, recommendation roles and preference similarity. Authors use trust transitivity connected to these aspects based on the topology of the complex social network. Another technique in the same field, has been proposed in [2], where authors describe a high-level model to evaluate how much a person or computational agent trusts another in a special decision context. Unfortunately, they leave as topic of future study how social norms contribute to trust and how this notion of social trust can be inferred from network structure, network flows and the evolution of the network structure. A more concrete approach for multi-dimensional social trust computation has been described in [33]. Here, authors capture multi-faceted trust relationships considering multi-faceted interests similarity among users in product review sites. This work does not illustrate a general model to capture social trust independently from the domain, and does not address the issue of the dynamic evolution of social scenarios.

The Trust Layer we have proposed in the paper deals with these open issues, and overreach trust transitivity problems connected to context and to the topology of a specific social network. Taking into account the bigger and general scenario of social interactions on the Web, and multiple evolving relationships among users, we are able to capture different context-based trustworthiness among them, without having to deal with, at least in this phase, the concept of transitivity.

6 Conclusions and Future Work

In this paper, we presented a new way to estimate trust values among users in the Social Web. The key idea is to gather as much as possible information from different social environments so as to form an augmented social graph, consisting of multiple types of relationships among users as well as resources. By elaborating this information and analyzing the evolution of the augmented social graph, we can estimate the trust relationships between involved users. At this purpose, we introduced event-based trust rules, associating a certain trust value with users performing a given pattern of events. We studied the effectiveness of the proposed Trust Layer on the social environment extracted from the ACM Digital Library dataset. We plan to extend this work following several directions. The first concerns the implementation of the CEP-based architecture in addition to the log-based one illustrated in the paper. Another relevant future work implies the generation of other experiments on different datasets, among which data from online social networks (e.g., Facebook) and other social environments.

References

1. Aberer, K., Despotovic, Z.: Managing trust in a peer-2-peer information system. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM 2001, pp. 310–317. ACM, New York (2001)
2. Adali, S., Wallace, W.A., Qian, Y., Vijayakumar, P., Singh, M.: A Unified Framework for Trust in Composite Networks. In: Proceedings of the 13th AAMAS Workshop on Trust in Agent Societies (Trust), pp. 1–12 (2011)
3. Anicic, D., Stojanovic, N.: Expressive logical framework for reasoning about complex events and situations. In: Intelligent Event Processing, Papers from the 2009 AAAI Spring Symposium, pp. 14–20 (2009)
4. Appelquist, D., Brickley, D., Carvahlo, M., Iannella, R., Passant, A., Perey, C., Story, H.: A standards-based, open and privacy-aware social web. Tech. rep., W3C (December 2010)
5. Aringhieri, R., Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: Fuzzy techniques for trust and reputation management in anonymous peer-to-peer systems. *J. Am. Soc. Inf. Sci. Technol.* 57(4), 528–537 (2006)
6. Atzori, L., Iera, A., Morabito, G.: SIoT: Giving a Social Structure to the Internet of Things. *IEEE Communications Letters* 15(11), 1193–1195 (2011)
7. Blaze, M., Feigenbaum, J., Ioannidis, J., Keromytis, A.D.: The Role of Trust Management in Distributed Systems Security. In: Dell’Acqua, P. (ed.) *Secure Internet Programming*. LNCS, vol. 1603, pp. 185–210. Springer, Heidelberg (1999)
8. Breslin, J., Decker, S.: The Future of Social Networks on the Internet: The Need for Semantics. *IEEE Internet Computing* 11, 86–90 (2007)
9. Castaneda, A., da Silva, P.P.: Extracting Trust Network Information from Scientific Web Portals. Tech. Rep. UTEP-CS-08-32, University of Texas at El Paso (August 2008)
10. Chu, X., Chen, X., Zhao, K., Liu, J.: Reputation and trust management in heterogeneous peer-to-peer networks. *Telecommunication Systems* 44, 191–203 (2010)
11. Ding, L., Chen, S., Rundensteiner, E.A., Tatemura, J., Hsiung, W.-P., Candan, K.S.: Runtime semantic query optimization for event stream processing. In: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, pp. 676–685. IEEE Computer Society Press, Washington, DC, USA (2008)
12. DuBois, T., Golbeck, J., Srinivasan, A.: Predicting Trust and Distrust in Social Networks. In: Proceedings of the Third IEEE International Conference on Social Computing (SocialCom 2011). IEEE (October 2011)
13. Falcone, R., Castelfranchi, C.: Trust and Transitivity: A Complex Deceptive Relationship. In: Proceedings of the 12th AAMAS Workshop on Trust in Agent Societies (Trust), pp. 43–54 (2010)
14. Fong, P.W.L.: Relationship-based access control: protection model and policy language. In: Proceedings of the First ACM Conference on Data and Application Security and Privacy, CODASPY 2011, pp. 191–202. ACM, New York (2011)
15. Ghiselli Ricci, R., Viviani, M.: Asymptotically idempotent aggregation operators for trust management in multi-agent systems. In: IPMU 2008: Proceedings, Malaga, Spain, June 22–27, pp. 129–137 (2008)
16. Golbeck, J., Hendler, J.: Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol.* 6, 497–529 (2006)
17. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22(1), 5–53 (2004)
18. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. *Decis. Support Syst.* 43, 618–644 (2007)

19. Jøsang, A., Marsh, S., Pope, S.: Exploring different types of trust propagation. In: Stølen, K., Winsborough, W.H., Martinelli, F., Massacci, F. (eds.) *iTrust 2006*. LNCS, vol. 3986, pp. 179–192. Springer, Heidelberg (2006)
20. Kagal, L., Finin, T., Joshi, A.: Trust-based security in pervasive computing environments. *Computer* 34(12), 154–157 (2001)
21. Kazienko, P., Musial, K., Kajdanowicz, T.: Multidimensional Social Network in the Social Recommender System. *IEEE Transactions on Systems, Man, and Cybernetics, Part A* 41(4), 746–759 (2011)
22. Kumar, R., Novak, J., Tomkins, A.: Structure and Evolution of Online Social Networks. In: Yu, P.S.S., Han, J., Faloutsos, C. (eds.) *Link Mining: Models, Algorithms, and Applications*, pp. 337–357. Springer (2010)
23. Kuter, U., Golbeck, J.: Using probabilistic confidence models for trust inference in web-based social networks. *ACM Trans. Internet Technol.* 10, 8:1–8:23 (2010)
24. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* 1(1) (March 2007)
25. Liu, G., Wang, Y., Orgun, M.A.: Trust Transitivity in Complex Social Networks. In: Burgard, W., Roth, D. (eds.) *AAAI*. AAAI Press (2011)
26. Luckham, D.C.: *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley Longman Publishing Co., Inc, Boston (2001)
27. Maheswaran, M., Tang, H.C., Ghunaim, A.: Towards a gravity-based trust model for social networking systems. In: *Proceedings of the 27th ICDCS Workshops*, p. 24. IEEE Computer Society Press, Washington, DC, USA (2007)
28. Mass, Y., Shehory, O.: Distributed Trust in Open Multi-agent Systems. In: Falcone, R., Singh, M., Tan, Y.-H. (eds.) *AA-WS 2000*. LNCS (LNAI), vol. 2246, pp. 159–174. Springer, Heidelberg (2001)
29. McKnight, D.H., Chervany, N.L.: *The Meanings of Trust*. Tech. Rep. MISRC Working Paper Series 96-04, University of Minnesota (1996)
30. Nin, J., Carminati, B., Ferrari, E., Torra, V.: Computing Reputation for Collaborative Private Networks. In: *Proceedings of COMPSAC 2009*, pp. 246–253. IEEE Computer Society Press, Washington, DC, USA (2009)
31. O'Donovan, J., Smyth, B.: Trust in recommender systems. In: *Proceedings of the 10th International Conference on Intelligent User Interfaces, IUI 2005*, pp. 167–174. ACM, New York (2005)
32. Rana, O.F., Hinze, A.: Trust and reputation in dynamic scientific communities. *IEEE Distributed Systems Online* 5(1) (2004)
33. Tang, J., Gao, H., Liu, H.: mTrust: Discerning Multi-faceted Trust in a Connected World. In: Adar, E., Teevan, J., Agichtein, E., Maarek, Y. (eds.) *WSDM*, pp. 93–102. ACM (2012)
34. Taylor, H., Yochem, A., Phillips, L., Martinez, F.: *Event-Driven Architecture: How SOA Enables the Real-Time Enterprise*, 1st edn. Addison-Wesley Professional (2009)
35. Wang, Y., Singh, M.P.: Trust representation and aggregation in a distributed agent system. In: *Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006*, pp. 1425–1430. AAAI Press (2006)
36. Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pp. 981–990. ACM, New York (2010)
37. Xiong, L., Liu, L.: PeerTrust: supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering* 16(7), 843–857 (2004)

On Recommending Hashtags in Twitter Networks

Su Mon Kywe, Tuan-Anh Hoang, Ee-Peng Lim, and Feida Zhu

Singapore Management University, Singapore

{monkywe.su.2011,tahoang.2011,eplim,fdzhu}@smu.edu.sg

Abstract. Twitter network is currently overwhelmed by massive amount of tweets generated by its users. To effectively organize and search tweets, users have to depend on appropriate hashtags inserted into tweets. We begin our research on hashtags by first analyzing a Twitter dataset generated by more than 150,000 Singapore users over a three-month period. Among several interesting findings about hashtag usage by this user community, we have found a consistent and significant use of new hashtags on a daily basis. This suggests that most hashtags have very short life span. We further propose a novel hashtag recommendation method based on collaborative filtering and the method recommends hashtags found in the previous month's data. Our method considers both user preferences and tweet content in selecting hashtags to be recommended. Our experiments show that our method yields better performance than recommendation based only on tweet content, even by considering the hashtags adopted by a small number (1 to 3) of users who share similar user preferences.

Keywords: Twitter, hashtag, recommendation systems.

1 Introduction

Motivation. Recommendation systems have been widely used at e-commerce websites to identify from a huge range of products the most appropriate ones for each user. Various recommendation methods have been proposed to predict the wanted products based on users' preferences as well as preferences of similar users [12]. A rising star of social information networks, Twitter presents to its users the same challenge of finding the most appropriate people to follow, tweets to read, as well as hashtags to use in their tweets [7]. Recommendation systems are therefore pertinent in these scenarios [12].

In Twitter, users write tweets which are short messages containing no more than 140 characters. A hashtag is a word prefixed by a # symbol and one or more hashtag can be inserted into a tweet. Past empirical research shows that hashtags have been used for different purposes. Some people use hashtags to categorize their tweets. Others use hashtags to tag content related to disasters or special events such as elections. Hashtags are also used for brand promotion or micro-meme discussions [6]. Hashtags make tweets easily searchable by other relevant users and this facilitates conversations among the users. Moreover,

hashtags make tweets more accessible by hashtag-based search engines such as [hashtags.org](http://www.hashtags.org)¹. In [4], hashtags have been used to help users tag other social media sites. Since hashtags are neither registered nor controlled by any user or group, it will be hard for some users to find appropriate hashtags for their tweets.

Research Objectives and Contributions. In this paper, we therefore address the personalized hashtag recommendation task in Twitter. The objective is to recommend a list of hashtags appropriate for a given user who has just written a new tweet. We do not consider hashtag recommendation for retweets (i.e., “forwarded” tweets) as they often contain the same hashtags as the corresponding original tweets. It is therefore relatively easy to derive hashtags for retweets. Hashtag recommendation should be personalized as we would like to consider the user preferences in the choice of hashtags. Twitter users adopt different styles and preferences in writing tweets. For example, users from UK may prefer hashtags in British spellings. Classical music lovers may prefer using composer names as hashtags for musical pieces. Knowing their personal preferences will help to predict the appropriate hashtags.

We begin this research by first analyzing a Twitter dataset consisting of tweets written by more than 150,000 Singapore users over a three-month period from October 2011 to December 2011. This is a reasonably large user community with 44M tweets. We examine the usage of hashtags among these users and their tweets, and highlight several interesting findings about the dataset.

The second part of the paper focuses on our proposed hashtag recommendation method. Our proposed method selects hashtags from both similar tweets (of the target tweet) and similar users (of the user who writes the target tweet). The selected hashtags are ranked and the top ranked hashtags are then recommended to the target user. We evaluate our proposed method and compare it with the recommendation method which only considers the hashtags from the most similar tweets. The results show that our method outperforms the latter approach by about 20 percent.

On the whole, this paper makes a number of contributions to hashtag analysis and recommendation as shown below:

- For the first time, a very large user community and its tweets have been used in a study on hashtag usage and recommendation. We have observed in this dataset that less than 8% of tweets contain hashtags and 40% of users ever use hashtags in our three-month data.
- Our study shows that hashtag usage by a user community is very skewed. Very few hashtags enjoy high popularity in tweets and users, while the vast majority of them are used in one tweet or by one user. This observation is consistent with the earlier studies.
- For any given day, we observe that 40% of the hashtags are not used by the user community in the last 30 days. This suggests that a lot of hashtags used are new to the users. This observation is only possible as we track the tweets from the same user community over time.

¹ <http://www.hashtags.org>

- We have developed a personalized hashtag recommendation method considering both user preferences and tweet content. The former has not been used in the previous methods.
- Our experiments show that user preferences from very few similar users can help to improve recommendation accuracy significantly.

Paper Outline. Our paper is structured as follows. Section 2 provides a quick summary of the related recommendation research in Twitter. We describe the Singapore’s user community and Twitter data collected from its users in Section 3. We also present our analysis results in this section. Section 4 describes our proposed hashtag recommendation method and its experiment results. We conclude the paper in Section 5.

2 Related Works

In this section, we give a brief overview of the traditional recommendation systems followed by previous recommendation research on Twitter network.

2.1 Traditional Recommendation Systems

Recommendation systems are information filtering systems which predict the preference of a user towards items (such as songs, books, or movies) or social elements (e.g. people or groups) that she has not considered before [5,12]. There are two types of recommendation systems – *personalized* and *non-personalized*. Non-personalized recommendation systems rank the items without considering the individual user’s preferences. For example, one may recommend the top ten popular songs of the current month. On the other hand, personalized recommendation systems consider the preferences of an individual user. The focus of our paper is on personalized recommendation. There are essentially two major approaches to perform personalized recommendation, namely collaborative filtering and content-based recommendation.

Collaborative Filtering Approach. The underlying assumption of the collaborative filtering approach is that if a person X has adopted several common items as adopted by another person Y previously, X is more likely to adopt other Y ’s items than the items of a random person. In the context of product rating recommendation, collaborative filtering has been used to predict the rating a target user assigns to an item using the ratings on the item assigned by other users who share similar rating preferences as the target user. This type of collaborative filtering is referred as the “user-to-user” collaborative filtering [13].

Another type of collaborative filtering approach is “item-to-item” collaborative filtering. In this approach, we first derive the correlation between two items which is measured by the portion of common users who purchase both items. We then recommend a new item to a target user using its correlation with other items already adopted by the target user.

Beyond user and item level information, collaborative filtering approach can also be performed in the latent factor space as a user or item can be represented by a set of latent factors through matrix factorization techniques [8]. It has been shown that matrix factorization techniques can yield very accurate recommendation results albeit higher algorithmic complexity.

Content Based Approach. Content-based recommendation approach measures similarity between items by comparing their features and characteristics. The recommendation of an item is made to a targeted user if the item is similar to other items adopted by the user before. Unlike item-to-item collaborative filtering, the content based approach makes use of item content only to determine similarity between items.

Other Approaches. More recently, community-based recommendation systems have been introduced to recommend items based on the preferences of a user's friends. Such a recommendation approach is only possible when users are connected with one another by friendship links or other forms of social relations [3]. There are many other recommendation systems using demographic information, such as age, profession, country, language, etc., to predict the user's preferences. Such systems use domain specific knowledge about how item features meet the user's needs and preferences or how the items are useful for the user. There are also hybrid systems which combine the above approaches.

2.2 Hashtag Recommendation for Twitter Users

Currently, Twitter has not implemented any hashtag recommendation system which suggests appropriate hashtags for the users' tweets. In the research literature, there are works related to hashtag recommendation and hashtag prediction. We found two hashtag recommendation approaches that are relevant and both of them use *only* tweet content [15, 11]. They will be described below in greater detail. Hashtag prediction refers to predicting the hashtag to be used by a user in the future. In [14], Yang et. al proposed to solve hashtag prediction by training a SVM classifier using a variety of features. Note that this task does not involve any target tweet.

Tweet Similarity Approach. Zangerle et al. [15] assumed that the primary purpose of the hashtags is to categorize the tweets and facilitate the search. The paper recommends suitable hashtags to the a target user, depending on the content that the user enters without considering user's preference for specific hashtags. Preliminary analysis of hashtags usage in a Twitter data collection obtained by a set of search queries shows that 86% of unique hashtags are used less than five times within 3,209,281 tweets with hashtags. The five most popular hashtags (#jobs, #nowplaying, #zodiacfacts, #news and #fb) appear in 8% of all tweets with hashtags. In other words, a few popular hashtags are used intensively while most of the other hashtags are used very sparsely. The paper also finds out the use of hashtags by spammers (e.g. assigning 17 hashtags to a single spammed tweet).

Zangerle et al. proposed a hashtag recommendation system that retrieves a set of tweets similar to a user given tweet. Similarity score is calculated by TF-IDF scheme. Then, the hashtags are extracted from the retrieved similar tweets and are ranked using one of the proposed score functions: (a) OverallPopularityRank score: number of hashtag occurrences in the whole dataset; (b) RecommendationPopularityRank score: number of hashtag occurrences in the retrieved similar tweet dataset; or (c) SimilarityRank score: similarity score of the most similar tweets containing the hashtag. Experiments showed that SimilarityRank score is the best among them and the performance of the recommendation system is the best when only five hashtags are recommended.

Naive Bayes Method. In [11], Mazzia et al. recommended hashtags by observing the content produced by the target user. Instead of TF-IDF to find similar tweets, the method proposes to use Bayes model to estimate the probabilities of using different hashtags. In the experiments, the Twitter dataset used is first cleaned by removing micro-memes and spams. Micro-memes are detected by identifying tweets which use the same hashtags but are very dissimilar. Spams are filtered by removing users who have too many tweets using the same hashtag. The Bayes model used in this paper is represented by the following formula.

$$p(C_i|x_1, \dots, x_n) = p(C_i)p(x_1|C_i)...p(C_i)p(x_n|C_i)/p(x_1...x_n)$$

where C_i represents the i^{th} hashtag and x_1, \dots, x_n represent the words. $p(C_i|x_1, \dots, x_n)$ is the probability of using hashtag C_i given the words that the user generates and the hashtags with the highest probabilities are recommended to the user. $p(C_i)$ is the ratio of the number of times hashtag C_i is used to the total number of tweets with hashtags. $p(x_1|C_i)...p(x_n|C_i)$ is calculated from the existing data of tweets.

3 Hashtag Usage Analysis

3.1 Twitter Data for Usage Analysis

In our study, we collect the Twitter data generated by a community of Singapore users. A complete analysis of hashtag usage in the entire Twitter network is not possible as such a dataset is not publicly available. Most researchers in the past chose to analyze Twitter data collected using some forms of data sampling on the stream of Twitter data returned by the APIs provided by the company. For example, Zangerle et. al used a set of query keywords to gather tweets [15]. Inevitably, the analysis results will be biased by the query relevant tweets.

As there is yet a comprehensive analysis of hashtag usage in the tweets generated by user communities, and how the usage patterns may affect hashtag recommendation, we first perform a detailed analysis on the three-month data (October 2011 to December 2011) generated by this Singapore user community. Our analysis aims to answer the following research questions: (a) How often are hashtags used in tweets? (b) How many hashtags do we expect in a tweet? (c)

How familiar are users in using hashtags? (d) Do the hashtags assigned already appear in earlier tweets? Providing answers to the above questions will give us a good understanding of the hashtag usage patterns of a user community and their changes over time.

We collect the Twitter data generated by more than 150,000 Singapore users who are identified by the location field in their user profiles. Every user is at least directly or indirectly connected to a small set of carefully selected seed users so as to prevent spammers to be included. The seed users are popular political bloggers, commentators, election candidates and news media during Singapore Election 2011. Since election is a big socio-political event, we believe that we cover the majority of Singapore Twitter users. We crawl all tweets of these Singapore users on a daily basis. In this manner, we are assured that almost all tweets from this user community have been completely downloaded for our study. Table 1 shows the important statistics found in this dataset. There are more 65,000 users who have written some original tweets during the three-month period. The remaining users (nearly 60% of total user population) do not write original tweets. They could perform retweeting or simply reading tweets from others. Our dataset also contains nearly 450,000 distinct hashtags and 45M original tweets.

Table 1. Data Statistics

| | |
|---------------------------------------|------------|
| # users | 65,410 |
| # users using hashtags | 46,244 |
| # distinct hashtags | 449,206 |
| # original tweets | 44,997,784 |
| # original tweets containing hashtags | 3,534,869 |

3.2 Hashtag Usage Analysis

There are substantial fraction of users (about 39%) using hashtags in their original tweets, and very small fraction of original tweets containing hashtags (<8%) as shown in Table 1. This suggests that many users know how to use hashtags but very few actually tweet a lot using hashtags. Figure 2 shows that the fraction of users using hashtags and the fraction of tweets containing hashtag over the three-month period remain very stable for this user community.

We define *tweet popularity* of a hashtag by the number of tweets containing the hashtag. We show the scatterplot of tweet popularity of hashtags in Figure 2(a). Each point in the figure represents the number of hashtags with the same tweet popularity. The distribution is power law-like showing that most hashtags appear in one tweet each and very few tweets enjoy very high tweet popularity. In a similar way, we define *user popularity* of a hashtag by the number of users using the hashtag. Figure 2(b) shows that the user popularity distribution of hashtags also follows the power law distribution. This suggests that only a few hashtags enjoy high popularity while most hashtags are used by a single user.

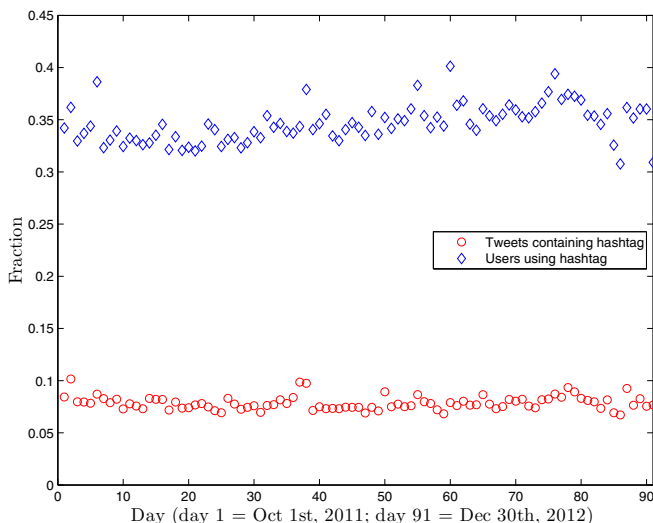
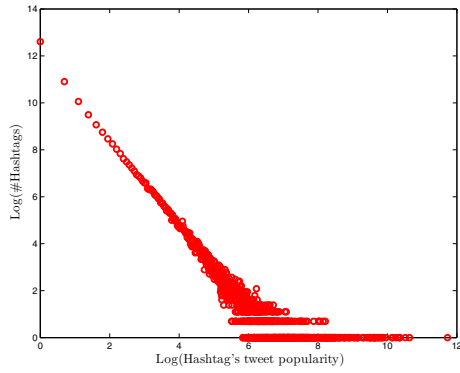


Fig. 1. Hashtag usage

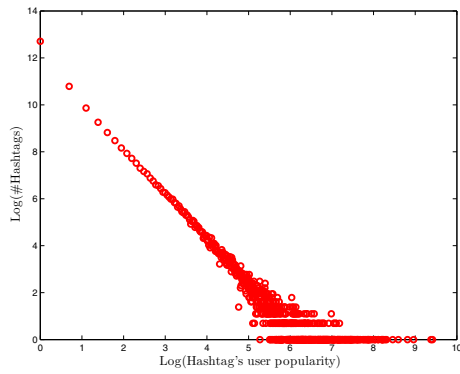
Next, we analyze how frequently users write tweets with hashtags. As shown in Figure 2(c), most users write only one tweet containing hashtag(s) during the observed period. Very few users write many tweets that contain hashtags. Finally, we found out most tweets with hashtag(s) contain only one hashtag as shown in Figure 3. There are very few tweets containing more than one hashtag. This is not a surprise given the short tweet length.

Finally, we want to know if the hashtags are new as users assign them to tweets. Unfortunately, the verification of new hashtags is very costly and may not be viable due to the lack of all historical twitter data. We therefore introduce the definition of “fresh hashtag”. A hashtag is said to be fresh to a user community if it has not been used by any user in the community in the last k months. This definition constrains the freshness verification to only k previous months of data generated by a user community. To reduce the verification cost, we have $k = 1$ in our current study.

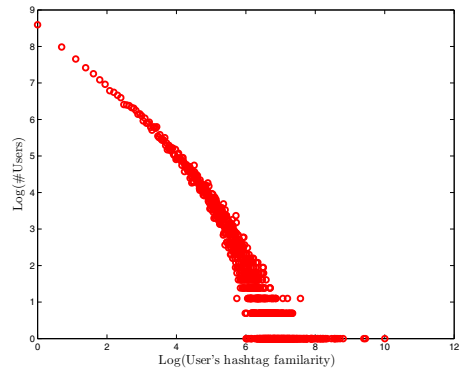
Figure 4 depicts the fraction of fresh hashtags, the fraction of tweets containing fresh hashtags and the fraction of users using fresh hashtags for each day. It is interesting to find 40% fresh hashtags are introduced each day. This suggests that another 40% hashtags are replaced each day. The life expectancy of many hashtags are therefore very short. Less than 30% of tweets contain fresh hashtags and around 40% of users use fresh hashtags each day. These observations lead us to believe that hashtag recommendation is an important task as it helps users to adopt more hashtags and makes their tweets easily searchable by other relevant users. The recommendation should also involve recent past data so as to recommend fresher hashtags.



(a) Tweet Popularity



(b) User Popularity



(c) User Hashtag Familiarity

Fig. 2. Data Distribution

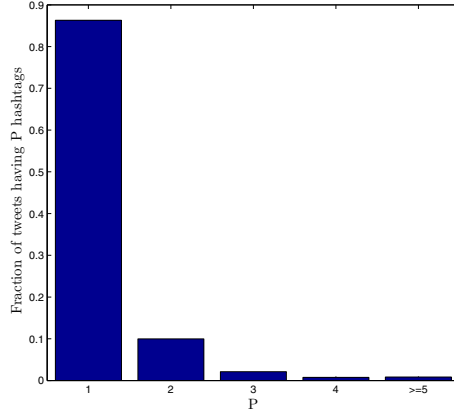


Fig. 3. Number of Hashtags in Each Tweet

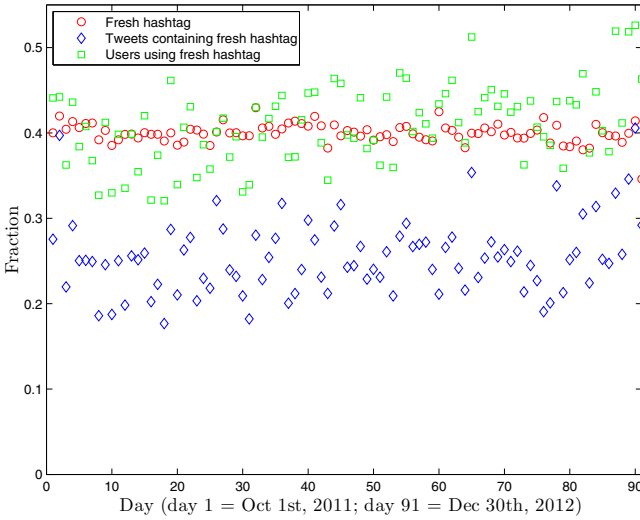


Fig. 4. Fresh hashtag usage

4 Personalized Hashtag Recommendation

Unlike previous methods that recommend hashtags found in similar tweets [15,11,10], we propose a new recommendation method that recommends hashtags which are not only appropriate for the tweet but also match the target user’s taste. In other words, given a user-tweet pair, we would like to find other similar user-tweet pairs and recommend the hashtags from those user-tweet pairs. We believe that this approach will be able to personalize the recommended hashtags to the user’s preferences. In the following, we first describe our proposed method followed by its evaluation.

4.1 Our Proposed Method

Finding similar user-tweet pairs involves three subtasks: (a) selecting hashtags from users with preferences similar to the target user, (b) selecting hashtags from tweets that are similar to the target tweet, and (c) deriving ranking scores for the selected hashtags. In both subtasks (a) and (b), we adopt a TF-IDF scheme to find similar users and tweets as described below.

Selecting Hashtags from Similar Users. We represent a user by her preference weights for each hashtag in our hashtag dictionary H . Formally, a user u_j is represented by a weight vector:

$$u_j = \{w_{1j}, w_{2j}, w_{3j}, \dots, w_{i|H}\}$$

where w_{ij} is the preference weight of user u_j towards hashtag h_i and can be defined by the TF-IDF scheme.

$$w_{ij} = TF_{ij}.IDF_i$$

$$TF_{ij} = \frac{Freq_{ij}}{Max_j}, IDF_i = \log\left(\frac{N_u}{n_i}\right)$$

where $Freq_{ij}$ = usage frequency of hashtag h_i by user u_j , Max_j = maximum hashtag usage frequency by u_j , N_u = total number of users, and n_i = number of users who use h_i before.

The intuition of TF_{ij} is that if a user uses a hashtag a lot, more preference weight is given to the hashtag. At the same time, this weight is normalized by the maximum hashtag frequency of the user. IDF_i assigns higher weight to a hashtag if the latter is rarely used by other users.

Given a target user u and another user u_i , we can measure the cosine similarity between them as follows.

$$Sim(u, u_i) = \frac{u \cdot u_i}{\|u\| \cdot \|u_i\|}$$

We then rank the users by similarity score and the most similar X users are selected. Let $TopXUsers(u)$ denote the X users most similar to u , and $Hashtags(u_i)$ be the set of hashtags previously used by u_i . We combine the hashtags from these top- X users to be our candidate hashtag set $HTofUsers(u)$.

$$HTofUsers(u) = \cup_{u_i \in TopXUsers(u)} Hashtags(u_i)$$

Selecting Hashtags from Similar Tweets. In a similar manner, we represent a tweet t_k can be represented by a weighted vector of words in a word vocabulary W .

$$t_k = \{w_{k1}, w_{k2}, w_{k3}, \dots, w_{k|W}\}$$

where

$$w_{kl} = TF_{kl}.IDF_l$$

$$TF_{kl} = \frac{Freq_{kl}}{Max_k}, IDF_l = \log \left(\frac{N_t}{n_l} \right)$$

where $Freq_{kl}$ = frequency of word w_l in tweet t_k , Max_k = maximum word frequency in t_k , N_t = total number of tweets, and n_l = number tweets in which w_l appears.

The similarity score between the target tweet t and another tweet t_k is defined by:

$$Sim(t, t_k) = \frac{t \cdot t_k}{||t|| \cdot ||t_k||}$$

We now select the top- Y tweets most similar to the target tweet t , denoted by $TopYTweets(t)$. Let $Hashtags(t_k)$ denote the set of hashtags in tweet t_k . We derive a second set of candidate hashtags $HTofTweets(t)$ from $TopYTweets(t)$ as follows.

$$HTofTweets(t) = \cup_{t_k \in TopYTweets(t)} Hashtags(t_k)$$

Ranking Candidate Hashtags. The candidate hashtags to be recommended for the target user u and tweet t can be obtained by the union of hashtags from top- X similar users and top- Y similar tweets.

$$SuggestedHashtags(u, t) = HTofUsers(u) \cup HTofTweets(t)$$

After that, hashtags in $SuggestedHashtags(u, t)$ are ranked by frequency. The hashtag frequency is defined by adding the number of times the hashtag is used by top- X users with the number of times it appears in top- Y tweets. Finally, the top ranked hashtags are finally recommended to the user u .

4.2 Experiment

To evaluate our hashtag recommendation method, we conduct experiments using the tweets generated by Singapore users in November and December of 2011. Tweets that do not contain hashtags are removed from the dataset. The remaining dataset in November contains 2,264,801 tweets and 37,617 unique users and is used as training data. To evaluate the recommendation results, we randomly selected 5606 original tweets from the December data with authors in the training set. These tweets form our target tweet set. The hashtags actually used in the target tweets serve as the ground truth. Since the hashtags to be recommended are from November, we expect that they are still relatively fresh.

Since other previous methods recommend hashtags purely based on similar tweets, our experiment varies the number of similar users (i.e., X) used in our method. When $X = 0$, our method will recommend only hashtags from similar tweets. We also want to evaluate the different number of similar tweets Y used in recommendation.

For each target user-tweet pair, we consider the top *five* and top *ten* recommended hashtags and measure the performance of our method using **hit rate** as defined below.

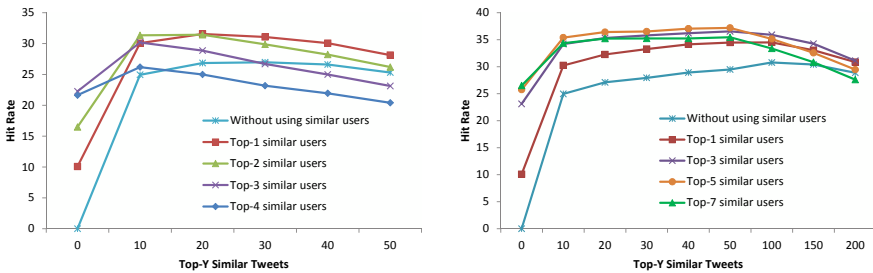
$$Hit\ Rate = \frac{\text{Number of Hits}}{\text{Number of Target User-Tweet Pairs}} \tag{1}$$

A hit occurs when the recommended hashtags for a target tweet *t* include at least one of the ground truth hashtags. Although multiple hashtags may be used in a target tweet, such cases are rare. Hence, it is reasonable to use the above hit rate measure.

We use Apache Lucene² to derive the similarity scores and retrieve the hashtags of the top-*X* similar users and hashtags of the top-*Y* similar tweets as Lucene is very efficient in such computation and retrieval.

4.3 Results

Figure 5(a) shows the hit rate (in percentage) of top five recommended hashtags. We vary the number of top similar tweets *Y* used from 0 to 50, and measure the performance of our method with top *X* = 0 to 4 similar users. The figure shows that as we increase the number of similar tweets from 0 to 10, the hit rate improves significantly. The improvement beyond 10 similar tweets is however very small or even negative. We can also observe that considering top 1 to 3 similar users can help to further improve the hit rate when the number of similar tweets are small, i.e., 10 and 20. The improvement percentage of recommendation using top 1 similar user over recommendation without similar user at *Y* = 10 is about 20%. Our method performs best with hit rate = 31.56% when *X* = 1 and *Y* = 20.



(a) Hit Rate (%) for Five Recommended Hashtags (b) Hit Rate (%) for Ten Recommended Hashtags

Fig. 5. Hit Rate

Figure 5(b) shows the hit rate (in percentage) of top ten recommended hashtags. We vary the number of top similar tweets *Y* used from 0 to 200, and measure the performance of our method with top *X* = 0, 1, 3, 5 and 7 similar

² <http://lucene.apache.org/core/>

users. On the whole, the hit rate has improved as we recommend more hashtags. Again, most significant improvement in hit rate occurs between $Y = 0$ and $Y = 10$. Beyond $Y = 10$, the improvement is small. On the other hand, using similar users is almost always better than not using similar users. The improvement margin of recommendation using top 1 similar user over recommendation without similar user at $Y = 10$, i.e., 21%, is similar to that observed for top 5 recommended hashtags. This time, our method performs best with hit rate = 37.19% when $X = 5$ and $Y = 50$.

5 Conclusions

Hashtag recommendation is a novel problem in Twitter. It is also important as most tweets do not carry hashtags and most hashtags do not have long life span. Our hashtag usage study on a three-month Twitter data generated by over 150,000 users in Singapore confirms the above observations. Our study shows that 40% of the hashtags in any day are fresh, i.e, not used in the last 30 days. We also observe that the usage patterns are stable over the period.

Our paper also proposes a personalized hashtag recommendation method that considers both target user preferences and target tweet content. Given a user and a tweet, our method selects the top most similar users and top most similar tweets. Hashtags are then selected from the most similar tweets and users and assigned some ranking scores. Experiment results show that using user preferences and tweet content will give us better recommendation than just using tweet content alone.

Beyond this early and promising results, there are several other interesting future directions to explore for hashtag recommendation. We can further divide hashtags into different categories, e.g., by freshness or by topic, and study their recommendation accuracies. In [9], popular hashtags have been clustered into four categories by their before-peak, after-peak, and during-peak popularity. For each hashtag category, it will be interesting to propose different recommendation methods that work well. So far, our proposed method is based on simple collaborative filtering. More sophisticated methods such as matrix factorization can also be used in the future.

Acknowledgements. This research/project is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

1. Armentano, M.G., Godoy, D.L., Amandi, A.A.: Recommending information sources to information seekers in twitter. In: International Workshop on Social Web Mining (2011)
2. Armentano, M.G., Godoy, D.L.: A topology-based approach for followees recommendation in twitter. In: The 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, ITWP 2011 (2011)

3. Chua, F.C.T., Lauw, H.W., Lim, E.-P.: Predicting item adoption using social correlation. In: SIAM Conference on Data Mining, pp. 367–378 (2011)
4. Correa, D., Sureka, A.: Mining tweets for tag recommendation on social media. In: The 3rd International Workshop on Search and Mining User-generated Contents (2011)
5. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22(1), 5–53 (2004)
6. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational tagging in twitter. In: The 21st ACM Conference on Hypertext and hypermedia, pp. 173–178 (2010)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: The 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (2007)
8. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* 42(8), 30–37 (2009)
9. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in twitter. In: The 21st International Conference on World Wide Web, pp. 251–260 (2012)
10. Li, T., Yu Wu, Y.Z.: Twitter hash tag prediction algorithm (2011), <http://cerc.wvu.edu/download/WORLDCOMP%2711/2011%20CD%20papers/ICM3338.pdf>
11. Mazzia, A., Juett, J.: Suggesting hashtags on twitter, <http://www-personal.umich.edu/~amazzia/pubs/545-final.pdf>
12. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: *Recommender Systems Handbook*, pp. 1–35. Springer (2011)
13. Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. *Data Mining and Knowledge Discovery* 5(1-2), 115–153 (2001)
14. Yang, L., Sun, T., Zhang, M., Mei, Q.: We know what @you #tag: does the dual role affect hashtag adoption? In: The 21st International Conference on World Wide Web, pp. 261–270 (2012)
15. Zangerle, E., Gassler, W.: Recommending #-tags in twitter. In: *Proceedings of the CEUR Workshop* (2011), <http://ceur-ws.org/Vol-730/paper7.pdf>

A Framework for the Design and Synthesis of Coordinated Social Systems

Wynn Stirling, Christophe Giraud-Carrier, and Teppo Felin

¹ Department of Electrical and Computer Engineering, Brigham Young University

² Department of Computer Science, Brigham Young University

³ Department of Organizational Leadership and Strategy, Brigham Young University

Abstract. This paper describes how a nascent collective of individuals can coalesce into a complex social system. The systematic study of such scenarios requires a mathematical framework within which to model the behavior of the individual members of the collective. As individuals interact, they develop social relationships and exchange resources – that is, they develop social capital that quantifies the value of social influence that individuals exert on each other. Social capital can be expressed via conditional preference orderings for each individual. Conditional preferences reflect the influence relationships of an interacting social collective. Conditional preference orderings can then be aggregated via conditional game theory to form a concordant utility that provides an emergent group-level ordering of the harmony of interests of the members of the collective. We can thus develop a complete social model that takes into consideration all social relationships as they propagate through the system. Solution concepts can then be defined that simultaneously account for both group-level and individual-level interests.

1 Introduction

A complex social system comprises individuals whose motives and behavior are difficult to understand and whose actions lead to emergent group-level behavior that is difficult to predict. Computer-based modeling of such societies provides a powerful tool for predicting and explaining human and social behavior. It also provides a powerful synthesis tool for designing artificial societies that are intended to perform useful tasks as a group. A key component of such models is the need for clearly defined notions of rational behavior on the part of the individuals and for a clearly understood aggregation process that leads to an understanding of the behavior of individuals and groups. Extant formal models of individual and social behavior and aggregation, however, are developed in highly idealized situations – in effect, within a social vacuum. In contrast, most social scientists study such behavior and aggregation without drawing on the repertoire of tools developed by computer scientists and mathematicians. We therefore seek to add a level of realism to mathematical and computational models by looking at the role that various social factors play in the formation of individual preferences and the emergence of social choices.

Game theory provides a powerful and parsimonious mathematical framework to model decisions by multiple agents where the outcome for each depends on the choices of all [8,13,23,24,29,38]. Noncooperative, single-stage, strategic-form games, in particular, are the most well known and also the simplest formulation of a mathematical game. Such a game consists of (i) a set of autonomous decision makers, or *players*, $\mathcal{X} = \{X_1, \dots, X_n\}$ ($n \geq 2$), (ii) an action set \mathcal{A}_i for each X_i , and (iii) a *utility* $u_{X_i}: \mathcal{A} \rightarrow \mathbb{R}$ for each $X_i, i = 1, \dots, n$, where $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_n$ is the *outcome space*. For any *action profile* $\mathbf{a} = (a_1, \dots, a_n) \in \mathcal{A}$, the utility $u_{X_i}(\mathbf{a})$, defines the benefit to X_i as a consequence of the outcome \mathbf{a} . These utilities are *categorical* in the sense that $u_{X_i}(\mathbf{a})$ unconditionally defines the benefit to X_i of the group instantiating the action profile \mathbf{a} , ostensibly without regard for the benefit this offers to other agents. In addition to the categorical structure of the utilities, it is usually assumed that each X_i possesses a logical structure that defines how it should play the game. The most widely used logical structure is the doctrine of *individual rationality*: each X_i should act in a way that maximizes its own utility, regardless of the effect on others. When each player subscribes to this notion and believes that all others do so as well, each solves its corresponding constrained optimization problem, resulting in a Nash equilibrium.

These mathematical and logical structures may provide an appropriate vehicle with which to model behavior in an environment of competition and market-driven expectations since, in that environment, the dominant notion of rational behavior is self-interest. It is less clear, however, that self-interest is the dominant notion in mixed-motive social environments, such as those containing opportunities for cooperation, compromise, and unselfishness, as well as for competition, intractability, and avarice. Several researchers have addressed this limitation. Fehr and Schmidt introduce the concept of inequity aversion by including parameters in the utility to account for the player's self-centered concept of itself as a fair individual [10]. Bergstrom defines a player's interdependent utility as a function of its own private utility and the perceived happiness of other players [2]. Sobel introduces psychological factors into the utilities to account for the player's beliefs about the intentions of others [33]. Another class of approaches invokes repeated-play protocols in an attempt to elicit social behavior. Bicchieri argues that rational choice theory must be augmented with a theory of belief formation, that is, a theory of learning [3]. If players are permitted to play the same game many times, they may gain insight regarding the social dispositions of the other players, and may be able to predict their behavior, establish their own reputations, and gain the trust of others to their mutual advantage. As players interact, they may learn to recognize behavioral patterns and settle on stable ones. Sophisticated social behavior, therefore, is viewed as the end result of social evolution. Yet another approach is offered with evolutionary game theory, where the preferences of the players are modified via replicator dynamics [37] or by genetic algorithms [11]. All of these approaches, however, are attempts to introduce social features into the utilities *without altering the mathematical structure of the utilities*. Thus, while they may introduce much needed psychological and social realism into the results, what is still missing is a notion of group benefit.

Multiagent systems are typically designed so that individuals work in a cooperative manner to accomplish some task, but, unfortunately, relying on individual rationality does not foster group rationality [20]. Consequently, conventional game theory has proceeded by making assumptions about individual preferences only and then using those preferences to deduce information about the choices (but not the values) of the group. It might be expected that cooperative game theory possesses some notion of group rationality. This version of game theory permits a subset of players to enter into a coalition such that each receives a payoff that is greater than it would receive if it acted alone. However, cooperative game theory employs categorical utilities and its solutions concepts are based squarely upon the assumption of individual rationality. Each player enters into a coalition solely on the basis of benefit to itself and, even though each may be better off for having joined, a notion of “group benefit” is not an issue when forming the coalition. Shubik has encapsulated the current state of conventional game theory as a framework within which to model the interrelationships of a complex social system. Following Shubik [31], we argue that the classical utility structure is not appropriate when it comes to behavior that falls outside the confines of self-interested behavior where preferences are already given. The objective of this paper is to investigate and provide a viable alternative.

Our approach is to move further upstream and get closer to the headwaters of the way preferences are actually formed. Our starting point is to extend beyond the classical notion of categorical preferences to accommodate *conditional preference orderings*, that is, preference orderings for an individual that can depend on the preferences of other individuals. We then propose to define these preferences in terms of the *social capital* possessed by the members of a society that empowers them to exert influence on other members. These influence linkages constitute the edges of a graph, or network, that models the social relationships that exist among the members of a group. The next step is to *aggregate these conditional utilities to form a social model of the group*. This social model may then be used to extract preference orderings for the group and for its members, leading to sophisticated solution concepts designed to reconcile the interests of the group with those of the individuals.

2 Preference Formulation and Social Capital

Perhaps the main attribute of a member of a collective is its ability to form preferences over the consequences of actions taken by the members of the collective. Whereas much of classical theory focuses on the behavior of the members of a collective whose social structure is assumed to be in place (i.e., each member has well-defined notions of preference and rational choice), we offer ways in which the members of a collective can form their preferences as they interact. Our fundamental assumption is that the members of a collective do not form their preferences in a vacuum; rather, they naturally consider the influence (e.g., political, economic, and social) that others may have on them. If an individual is confined to a categorical preference ordering, then it must effectively

aggregate the influence of others with its own narrow self-interest to arrive at a single compromise preference ordering. But distilling such potentially conflicting information into a single categorical ordering can strip the representation of important, and even essential, social content. Furthermore, categorical preferences do not permit individuals to modify their preferences upon contact with others. Goodin also argues that members of a complex social system possess metapreferences; that is, preferences for preferences [12]. Hence, we propose that, rather than relying on categorical preference orderings, a more sophisticated approach is to represent the preferential attributes of the members of a complex social system with *conditional* preference orderings. Our approach is to express these preferences using the mathematical syntax of probability theory. We thus introduce the notion of a *conditional utility*.

Consider the following scenario: X_1 and X_2 are to purchase an automobile under the following division of labor: X_1 will choose the color, either red (r) or green (g), and X_2 will choose the model, either a convertible (c) or a sedan (s). The corresponding outcome space is $\mathcal{A} = \{(r, c), (r, s), (g, c), (g, s)\}$. Thus, for any action profile $(a_1, a_2) \in \mathcal{A}$, a categorical utility for X_1 is of the form $u_{X_1}(a_1, a_2)$, and the expression $u_{X_1}(r, c) > u_{X_1}(g, c)$ means that X_1 prefers a red convertible to a green convertible, with a corresponding categorical utility for X_2 . Suppose, however, that X_1 does not possess a categorical utility over the joint action space. Instead, let X_1 reason as follows: if X_2 were to most prefer a green sedan, then X_1 would prefer the green sedan to the green convertible. Such a statement is a hypothetical proposition. X_1 does not need to know for certain that X_2 does indeed most prefer a green convertible. That statement is merely the antecedent of a hypothetical proposition whose consequent is X_1 's preference ordering. Symbolically, we may express this relationship as $u_{X_1|X_2}(g, s|g, s) > u_{X_1|X_2}(g, c|g, s)$, where the argument to the left of the conditioning symbol “|” is the profile under consideration by X_1 and the argument on the right is a hypothetical profile for X_2 (this syntax is similar to that of conditional probability). This relationship does not commit X_1 to prefer a green sedan to a green convertible. In fact, X_1 may also possess the following conditional preference ordering: $u_{X_1|X_2}(g, c|g, c) > u_{X_1|X_2}(g, s|g, c)$. These two expressions indicate a willingness for X_1 to conform its preferences to the preferences of X_2 , but this willingness need not apply to all situations. For example, it may also be true that $u_{X_1|X_2}(r, s|r, c) > u_{X_1|X_2}(r, c|r, c)$, that is, X_1 is not willing to conform to X_2 's preferences if X_2 were to most prefer a red convertible.

This utility structure permits the modeling of social relationships that may or may not go beyond self-interest. One explanation for X_1 's conditional preferences is altruism: a willingness to defer to X_2 's wishes, thereby expanding beyond self-interest. Another possible explanation is that X_1 has a well-defined notion of aesthetics, but does not really care what X_2 's preferences are – its only concern is that the car they purchase meets its personal aesthetic criteria. Other explanations are certainly possible. The critical issue is that X_1 's preferences can be influenced by X_2 's preferences. This *social influence* is in contrast with notions of influence employed by classical game theory, where it is assumed that

each player's payoff is influenced by the actions of the other player, but *not* by the other player's preferences (at least ostensibly).

The concept of social capital provides a natural starting point to reason about the formation and evolution of social influence. Whereas most forms of capital are functions of what individuals possess, social capital is grounded in a) the *relationships* that exist among individuals, and b) the *resources* that are available to individuals because of these relationships [7,117,27]. Social capital quantifies the value of social relationships for achieving some individual or group benefit based on the resources present in the underlying network [1,4,18,26]. Here, we adopt the definition of social capital proposed by Bourdieu and Waquant: "the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition" [5]. Hence, an individual X_i 's social capital is created and evolves as a result of social interaction with individuals who possess resources X_i needs. Social interactions may be used to create and strengthen bonds within homogeneous groups, or to leverage diversity by creating bridges across heterogeneous individuals and groups. They may take on various forms, involving different levels of engagement (e.g., emailing, attending a meeting, going out to dinner) and the possible exchange of resources (e.g., information, goods, services). Social interactions are also affected by individual attitudes and dispositions. For example, some individuals are naturally more philanthropic than others, while some always consider "what's in it for them." Similarly, some people are more naturally grateful than others (thus feeling a strong sense of reciprocity), while others may suffer from a sense of entitlement (thus feeling detached). As social interactions take place, relationships among social agents may change in intensity, new relationships may arise, and existing relationships may dissolve. In turn, the strengths of these relationships affect the accessibility and mobilization of resources within the social network.

In general, a *social resource* is a specific asset, material or symbolic, available through social connections within the network. Not all resources are created equal. An individual may attach much value to their material possessions while another might not. An individual may have a naturally altruistic attitude while another may find it more difficult to part with or share resources with others. An individual may value a resource highly today and much less tomorrow, because that resource ages (e.g., it may be easier to let someone borrow your old beat up truck than your brand new sports car), or more of the resource becomes available (e.g., it is generally easier to give away money when there is a surplus than when there is only enough to meet one's own needs), or attitude towards the resource changes (e.g., as children get older they find it easier to pass their once most valued toys on to their younger siblings). In addition, it is generally the case that the flow of resources depends not only on how the owner values them but also on their relationship to the requestor. For example, if r is X_j 's personal vehicle, X_j will likely access to the use request of a family member or a close friend, but not to that of a stranger. On the other hand, if r is one of X_j 's screwdrivers, X_j will likely let almost anyone borrow it. We model these ideas by saying that X_i may

get access to a resource r owned by X_j if and only if $s_{X_j X_i}^{ESN} \geq v_{X_j}^r$, where $s_{X_j X_i}^{ESN}$ is the strength of the explicit relationship from X_j to X_i ¹ and $v_{X_j}^r$ is the value that X_j assigns to resource r . It follows that the social capital of the members of a network is defined as $sc(X_i) = \sum_{X_j} |\{r \in R_{X_j}^p : s_{X_j X_i}^{ESN} \geq v_{X_j}^r \wedge C_{X_i}(r) = Yes\}|$, where $R_{X_j}^p$ is the set of resources possessed by X_j , and $C_{X_i}(r)$ is an indicator function that determines whether r is relevant to X_i .

We argue that the social capital possessed by each individual in a collective is the appropriate starting point for constructing a model for a complex social system. There is indeed an intuitive connection between one’s social capital and one’s ability to influence others, and hence an opportunity to model the formation and evolution of preferences. In contrast to the conventional game-theoretic assumption that the way preferences are formed is irrelevant to the function of the society, our approach constructs the preferences as consequences of the social structure. As the car-buying example illustrates, these preference orderings may be conditional, and thus need to move beyond the assumption that all individuals possess categorical preferences. The challenge invoked by this novel structure is that the existing decision methodologies that are commonly used to characterize multiagent decision problems is no longer applicable. Thus, we need to extend the theory beyond the approaches taken by classical game theory and social choice theory.

3 Conditional Game Theory

Classical game theory, as developed by von Neumann and Morgenstern, requires each individual to possess categorical utilities [36]. As commonly employed in general economic theory, it is assumed that all social complexity can be expressed via categorical utilities, and that individual rationality is the logical structure behind whatever solution concept that is applied. In an attempt to move beyond these mathematical and logical limitations, Stirling has developed an extension of game theory, termed *conditional game theory*, that is designed to accommodate complex social relationships that cannot easily be expressed via categorical utilities, and for which individual rationality is inadequate to characterize the behavior of the members of a complex society [34, 35].

Let $\mathcal{X} = \{X_1, \dots, X_n\}$ be an agent-based system. We denote by $pa(X_i) = \{X_{i_1}, \dots, X_{i_{p_i}}\}$ the subset of \mathcal{X} that socially influences X_i . The subset $pa(X_i)$ is termed the *parent set* of X_i . For each $X_{i_j} \in pa(X_i)$, $j = 1, \dots, p_i$, let \mathbf{a}_{i_j} denote an action profile that X_i hypothesizes as the outcome most preferred by X_{i_j} . The profile \mathbf{a}_{i_j} is called a *conjecture* for X_{i_j} . The collection $\{\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}\}$ of conjectures is a *joint conjecture* for $pa(X_i)$. For each joint conjecture $\{\mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}\}$ for $pa(X_i)$, the *conditional utility* for X_i is a mapping

¹ Explicit social relationships are clearly directed. The amount of social capital X_i may realize from a relationship with X_j is not predicated upon the value that X_i places in the relationship, but rather upon the value that X_j places in it. For example, if X_i is seeking a job reference from X_j , the reference will only be as strong as X_j thinks of X_i , and not the other way.

$u_{X_i | \text{pa}(X_i)}(\cdot | \mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}}) : \mathcal{A} \rightarrow \mathbb{R}$. A conditional game then comprises a) a set of players $\mathcal{X} = \{X_1, \dots, X_n\}$ where $n \geq 2$, b) a set of outcomes $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_n$, where \mathcal{A}_i is the action set for X_i , $i = 1, \dots, n$, and c) a set of conditional utilities $\{u_{X_i | \text{pa}(X_i)} : \mathcal{A} \rightarrow \mathbb{R}\}$, where $\text{pa}(X_i) = \{X_{i_1}, \dots, X_{i_{p_i}}\}$ is the parent set for X_i . The function $u_{X_i | \text{pa}(X_i)}(\mathbf{a}_i | \mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}})$ is the utility that X_i ascribes to action profile \mathbf{a}_i , given the conjectures \mathbf{a}_{i_k} , $k = 1, \dots, p_i$. If $p_i = 0$, then X_i possesses a categorical utility, denoted u_{X_i} . A conditional game reverts to a classical game if all utilities are categorical.

As conditional preferences propagate through the system via influence linkages, social bonds are created among agents. The end result of this propagation is an aggregation function that combines all of the individual conditional utilities to form a group-level utility termed a *concordant utility*, which provides a complete social model of the group. A new theory of preference aggregation has also been developed, which establishes conditions under which the aggregation of conditional utilities possesses the same mathematical syntax as probability mass functions (albeit with different semantics² and aggregating them is achieved by applying the chain rule of probability [34, 35]. In other words, the aggregation of conditional utilities is mathematically equivalent to computing the joint probability of a family of discrete random vectors as the product of conditional probability mass functions. Thus, the concordant utility is of the form $U_{X_1 \dots X_n}(\mathbf{a}_1, \dots, \mathbf{a}_n) = \prod_{i=1}^n u_{X_i | \text{pa}(X_i)}(\mathbf{a}_i | \mathbf{a}_{i_1}, \dots, \mathbf{a}_{i_{p_i}})$.

Since it is a function of n action profiles, the concordant utility cannot be used directly to define a group-level ordering. Rather, it provides a representation of the social consistency of the group, in that it gives a measure of the degree of severity of controversy. To illustrate, consider a two-stakeholder group $\{X_1, X_2\}$. Let \mathbf{a} and \mathbf{a}' be such that \mathbf{a} is best for X_1 and next-best for X_2 , and \mathbf{a}' is worst for X_1 and best for X_2 . It is reasonable to argue that if both were to conjecture \mathbf{a} , the degree of controversy would be fairly small, since both agents receive a reasonable reward. If both were to conjecture \mathbf{a}' , however, the outcome would be worst for one and best for the other; hence the degree of controversy would be quite large, and the condition $U_{X_1 X_2}(\mathbf{a}, \mathbf{a}) > U_{X_1 X_2}(\mathbf{a}', \mathbf{a}')$ would obtain.

In general, $U_{X_1 \dots X_n}(\mathbf{a}_1, \dots, \mathbf{a}_n) \geq U_{X_1 \dots X_n}(\mathbf{a}'_1, \dots, \mathbf{a}'_n)$ means that if the group were jointly to conjecture $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$, the level of controversy for \mathcal{X} would be less than or equal to what it would be if the group were jointly to conjecture $\{\mathbf{a}'_1, \dots, \mathbf{a}'_n\}$. The concordant utility characterizes all of the social relationships that exist in the group and permits the definition of an emergent notion of group preference, namely, an aversion to controversy. Thus, the concordant utility induces an ordering for the group; namely, that $\{\mathbf{a}_1, \dots, \mathbf{a}_n\}$ is preferred to $\{\mathbf{a}'_1, \dots, \mathbf{a}'_n\}$ if the severity of controversy is less with the former than with the latter. It should be noted that a lack of controversy does not necessarily imply a condition of harmony or agreement. For example, with an athletic contest, the

² The power of probability theory is succinctly expressed by Shafer (cited in [25], p. 15]): “probability is not really about numbers; it is about the structure of reasoning.” This reasoning structure is not only applicable to the stochastic domain; it applies equally well to reasoning about the utility of alternative outcomes.

opposing players do not cooperate in the sense of pursuing a common objective; rather, success in playing the game depends on their ability to perform in opposition to each other. In other words, there is a preference, from the group perspective, for the players to have disputes regarding their desired behavior. In such a case, diametrically opposed conjectures would, from a group perspective, have a low degree of controversy and, accordingly, a high concordant utility.

Although the concordant utility provides a complete description of all of the interconnections among the members of the collective, in its present form it does not enable the members of the collective to make decisions, since it is a function of multiple conjectures, not just one action profile.

4 Solution Concepts

Given a conditional game with influence relationships that admit the DAG structure, the issue now becomes one of defining solution concepts that are compatible with this more sophisticated structure. To do this, we must extract information from the concordant utility. We propose a two-step procedure comprising a meso-to-macro (intermediate to global) step followed by a macro-to-micro step (global to local) step. The meso-to-macro step aggregates the conditional preference orderings to form a global representation of the group in the form of a concordant utility. Once formed, we may use this representation to perform a macro-to-micro reduction to individual preferences, which then may be used to define solution concepts that simultaneously take into account both individual and group preferences.

4.1 Conditioned Nash Equilibria

Once a conditional game is defined, it is natural to consider how to apply classical solution techniques. We may easily extend the Nash equilibrium to the conditional case as follows. Suppose $\mathbf{a} = (a_1, \dots, a_i, \dots, a_n)$ is a fixed action profile. For each X_i , let $\mathbf{a}'_i = (a_1, \dots, a'_i, \dots, a_n)$; that is, \mathbf{a}'_i differs from \mathbf{a} in the i th position. The action profile \mathbf{a} is a *conditioned Nash equilibrium* if $u_{x_i | pa(x_i)}(\mathbf{a} | \underbrace{\mathbf{a}, \dots, \mathbf{a}}_{p_i \text{ times}}) \geq u_{x_i | pa(x_i)}(\mathbf{a}'_i | \mathbf{a}, \dots, \mathbf{a})$ for all $a'_i \neq a_i, i = 1, \dots, n$. If

all utilities are categorical, the conditioned Nash equilibrium becomes a classical Nash equilibrium. A conditioned Nash equilibrium permits each agent to define a rational choice for itself that also takes into consideration the interests of others, and therefore allows the agent to extend its sphere of interest beyond narrow self-interest. This approach may be a manifestation of enlightened self-interest, but it does not lead to a notion of group preference.

4.2 Concurrence

Since the concordant utility is a function of the conjectured action profiles of all players and these profiles may all be different from each other, this utility does

not provide a basis for making decisions unless we constrain all of the profiles to be the same. In that case, the concordant utility provides a measure of benefit (the severity of controversy) for the group if all participants were to agree on the same profile. Given a concordant utility $U_{\mathbf{x}}$, a *concurrent utility* is defined as the concordant utility evaluated at a common profile, namely $c_{\mathbf{x}}(\mathbf{a}) = U_{\mathbf{x}}(\mathbf{a}, \dots, \mathbf{a})$. A *concurrency*, denoted \mathbf{a}_c , is then a profile that maximizes the concurrent utility, i.e., $\mathbf{a}_g = \arg \max_{\mathbf{a}} c_{\mathbf{x}}(\mathbf{a})$. A concurrency maximizes the functionality of the group in the sense that the severity of controversy is minimized, regardless of the benefit to its members. If all members of the group were intent on, and only on, maximizing concordance, they would unanimously choose a concurrency.

4.3 Marginalization

During the process of aggregation, the local social relationships, as characterized by the *ex ante* conditional utilities, propagate throughout the group to form the concordant utility $U_{\mathbf{x}_n}$. The *ex post* unconditional utilities u_{x_i} , $i = 1, \dots, n$ are then extracted from the concordant utility by marginalization, yielding $u_{x_i}(\mathbf{a}_i) = \sum_{\sim \mathbf{a}_i} U_{\mathbf{x}_n}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. The *ex post* utilities take into consideration all of the social relationships that are expressed *ex ante* via conditional utilities. They define each agent's personal preferences after having taken into account the interests of all agents who influence it. Once the marginal utilities are defined, the history of their creation ceases to be relevant to the application of classical techniques. They are unconditional and indistinguishable in structure from *ex ante* categorical utilities. Consequently, they may be used according to any classical solution concept (e.g., we may compute the *ex post* Nash equilibrium action in the usual way that Nash equilibria are defined). Such an approach, however, while perhaps providing more psychologically realistic individual utilities, does not get us any closer to a notion of group preference or group rationality.

4.4 Group and Individual Welfare

Classical game theory drives a wedge between the concept of what is good for individuals and what is good for the group. Shubik warns against the anthropomorphic trap of ascribing judgment and choices to a group [30]. This sentiment may be true in the context of individual rationality and categorical utilities, but when the preferences of agents are influenced by the preferences of other agents, it is premature to argue that a notion of group wants or preferences cannot be defined. Again drawing on the probabilistic analogy, if two random variables are independent, then knowing the value of one of them conveys no information about the value the other takes. However, if they are not independent, then knowing the value of one of them does indeed say something about the value the other takes. In other words, some notion of group association exists between the two random variables. By the same reasoning, if two agents are not praxeologically independent, then some notion of group association (sociality) exists between them. If so, then it may indeed be meaningful to define a notion of group preferences that does not obviate notions of individual preferences.

In Section 4.2 we introduced a notion of group preference, but at the expense of considering individual preferences, and in Section 4.3 we were able to extract *ex post* individual utilities from the concordant utility, but no notion of group preference was considered. Here we show how both preference notions can be handled simultaneously. Let us first consider preference orderings for the group. To identify such an ordering, we must focus on the concordant utility. The concordant utility, however, does not directly serve as the basis for taking action, since it is a function of multiple profiles (conjectures), and only one profile can actually be implemented. Nevertheless, just as we may extract marginals from the concordant utility for each individual, we may also extract a marginal for the group. To proceed, we observe that since each agent can control only its own actions, what is of interest is the utility of all agents *making conjectures over their own action spaces*. Consider the concordant utility $U_{\mathbf{x}_n}(\mathbf{a}_1, \dots, \mathbf{a}_n)$. Let a_{ij} denote the j th element of \mathbf{a}_i ; that is, $\mathbf{a}_i = (a_{i1}, \dots, a_{in})$ is X_i 's conjecture profile. Next, form the action profile (a_{11}, \dots, a_{nn}) by taking the i th element of each X_i 's conjecture profile, $i = 1, \dots, n$. Now let us sum the concordant utility over all elements of each \mathbf{a}_i except the ii -th elements to form the *group welfare function*³ $v_{\mathbf{x}}$ for \mathbf{X} , yielding $v_{\mathbf{x}}(a_{11}, \dots, a_{nn}) = \sum_{\sim a_{11}} \dots \sum_{\sim a_{nn}} U_{\mathbf{x}}(\mathbf{a}_1, \dots, \mathbf{a}_n)$, where $\sum_{\sim a_{ii}}$ means the sum is taken over all a_{ij} except a_{ii} . The *individual welfare function* v_{X_i} of X_i is the i -th marginal of $w_{\mathbf{x}}$, that is, $v_{X_i}(a_{ii}) = \sum_{\sim a_{ii}} w_{\mathbf{x}}(a_{11}, \dots, a_{nn})$.

The group welfare function provides a complete *ex post* description of the relationships between the members of a multi-agent group. Unless its members are praxeologically independent, this utility is not simply an aggregation of individual utilities. It represents a true meso-to-macro transformation of individual conditional utilities to a group-level utility (the global level) as the individual conditional preferences propagate through the group, resulting in an emergent notion of group preference. Such a preference ordering is strictly a mathematical notion, and corresponds to the degrees of concordance that the outcomes provide. We define the *maximum group welfare* solution as $\mathbf{a}^* = \arg \max_{\mathbf{a} \in \mathcal{A}} v_{\mathbf{x}}(\mathbf{a})$. Once the group-level utility is defined, we may extract individual utilities (the local level) for each player as a function of its own actions. We may then define the *maximum individual welfare* solution as $a_i^\dagger = \arg \max_{a_i \in \mathcal{A}_i} v_{X_i}(a_i)$. If $a_i^\dagger = a_i^*$ for all $i \in \{1, \dots, n\}$, the action profile is a *consensus* choice. In general, however, a consensus will not obtain, and negotiation may be required to reach a compromise solution.

4.5 Negotiation

The existence of group and individual welfare functions provides a rational basis for meaningful negotiations; namely, that any compromise solution must at least provide each agent with its security level, i.e., the maximum amount of benefit it could receive regardless of the decisions of others. The security level for X_i is the maximin profile, defined as $s_{X_i} = \max_{a_i} \min_{\sim a_i} u_{X_i}(a_1, \dots, a_i, \dots, a_n)$.

³ The term “social welfare” is heavily used in welfare economics as a measure of the benefit to a group in a social choice context. In our game-theoretic context, we use “group welfare” in lieu of “social welfare” when discussing the benefit to the group.

In addition to individual benefit, we must also consider benefit to the group. Although a security level, per se, for the group cannot be defined in terms of a minimum guaranteed benefit (after all, the group itself does not actually make a choice), a possible rationale for minimum acceptable group benefit is that it should never be less than the smallest benefit to the individuals. This approach is consistent with the egalitarian principles of justice espoused by Rawls, who argues, essentially, that a society as a whole cannot be better off than its least advantaged member [28]. Accordingly, let us define a security level for the group as $s_{\mathbf{x}} = \min_i \{s_{x_i}\} / n$, where we divide n since the utility for the group involves n agents. Then, let $N_{\mathbf{x}} = \{\mathbf{a} \in \mathcal{A}: v_{\mathbf{x}}(\mathbf{a}) \geq s_{\mathbf{x}}\}$ be the *group negotiation set*, $N_{x_i} = \{a_i \in \mathcal{A}_i: v_{x_i}(a_i) \geq s_{x_i}\}$, $i = 1, \dots, n$ be the *individual negotiation sets*, and $R_{\mathbf{x}} = N_{x_1} \times \dots \times N_{x_n}$ be the *negotiation rectangle* (i.e., the set of profiles such that each member's element provides it with at least its security level). We can now define the *compromise set*, $C_{\mathbf{x}} = N_{\mathbf{x}} \cap R_{\mathbf{x}}$, which simultaneously provides each member of the group at least its security level, as well as meeting the group's security level. If $C_{\mathbf{x}} = \emptyset$, then no rational compromise is possible at the stated security levels. One way to overcome this impasse is to decrement the security level of the group iteratively by a small amount, thereby enlarging $N_{\mathbf{x}}$ until $C_{\mathbf{x}} \neq \emptyset$. If $C_{\mathbf{x}} = \emptyset$ after the maximum reduction in group security has been reached, then no rational compromise is possible, and the group may be considered dysfunctional. Another way to negotiate is for individual members to iteratively decrement their security levels. Once $C_{\mathbf{x}} \neq \emptyset$, any element of this set provides each member, as well as the group, with at least its security level. If $C_{\mathbf{x}}$ contains multiple elements, then a tie must be broken. One possible tie-breaker is $\mathbf{a}_c = \arg \max_{\mathbf{a} \in C_{\mathbf{x}}} v_{\mathbf{x}}(\mathbf{a})$, which provides the maximum benefit to the group such that each of its members achieves at least its security level.

5 Discussion and Conclusion

We have introduced a computational framework to deal with the role that social relations play in preference formation and aggregation. Extant models focus solely on self-interested actors and categorical preferences, or they postulate other-regarding behaviors (e.g., fairness), but these models have not systematically incorporated the social relations and ensuing social capital, which bind and enable agents, and influence both their preferences as well as emergent social choices. We have placed particular emphasis on conditional preferences, which provide a general framework for considering many factors that naturally influence social interaction and aggregation. For example, conditioning might be a function of expertise, where the established hierarchy of relations and influence is conditioned by who knows what, or the expertise needed for the particular choice at hand. In other words, if a particular member of the group is seen as more knowledgeable, then their preferences might be weighted more heavily and they may influence the joint outcome. This conditioning, of course, may also apply to multi-agent systems where some agents are seen as less knowledgeable, and thus their preferences and information are discarded by others. Our computational framework thus opens the door to a more realistic

and accurate study of important real-world problems. We mention two here as an illustration.

Organizations and Markets. The question of preference aggregation and collective behavior is central in the context of organizations and markets. If organizations indeed are composed of “individuals and groups whose preferences, information, interests, or knowledge differ” [21], and if “administrative activity [indeed] is group activity” [32], then the aggregation of heterogeneous preferences and information is also central for understanding the choices and behavior of firms in markets [14]. Another, more upstream question also relates to the emergence of organization in markets [6]. That is, how do individuals coalesce in markets to initiate joint, collective action? Extant theory focuses largely on the role of an “entrepreneur” in choosing which activities and transactions to engage with (for an overview, see [40]). However, the collective aspects and social choices behind entrepreneurship have received little, if any, analytic attention. That is, the theory of the firm, for example, is relatively atomistic in focusing on singular entrepreneurs rather than specifying the social processes behind collective decision-making about nascent organizational activity and strategies. With our framework as a basis, we can study (via simulations) the emergence of organizations and the evolution of firm strategies as individuals and firms interact with each other over time and resolve differences in their beliefs and expectations about possible organizational strategies.

Extremism and Terrorist Violence. Most explanations of political behavior are firmly rooted at the individual-level [16]. Thus, compelling explanations for some of the most common acts of dissidence remain elusive. For example: Why do some individuals engage in suicide terrorism when arguably they likely will not be alive to reap any of the benefits? Why do extremists join violent movements when the probability of detection (and therefore punishment) is so high? These questions have been considered primarily by economists, political scientists, and sociologists, and most explanations rely almost exclusively on individual notions of rationality, even when analyzing groups [9,15,19]. Rationalist explanations are in some ways compelling, but they can be incomplete and the logic somewhat tortured. Suicide terrorists are argued to be self-interested because they seek extra-world rewards (think 72 virgins) rather than material benefits in this life [22]. Extremists are thought to be self-interested and primarily engaged in ethnic or political violence because of promised (or hoped for) selective incentives [16], such as killing in order to avoid being killed. And yet numerous empirical accounts of terrorism and violent conflict clash with such notions of rationality, and more generally indicate that social bonds within larger groups matter [39]. Because individual-level explanations face so many “anomalies,” much conflict scholarship jumps directly to group-level explanations. But such group-level explanations, which treat groups as unitary actors, ignore completely the fact that groups are social constructs that depend on the correct aggregation of individual preferences and behaviors. Social capital and networks clearly matter in the context of terrorism, civil war, and extremism. The concepts of

conditional preferences and utilities as well as social capital of our framework allow a rigorous examination of how rebel, extremist, and terrorist groups emerge and shape each others' individual preferences, and how those preferences aggregate to the group level in ways that imply particular individual and group-level equilibrium behavior as they compete against an established state.

While we do not claim that ours is a definitive solution, we argue that the notion of socially conditioned preferences is essential to explaining the formation of both individual and social preferences. Similarly, we feel strongly about the dynamic aspects of how social relations and interactions evolve over time. Here, we have pointed out how social capital might dynamically capture the extant set of social relations and their evolution. As individuals interact, they learn more about each other and their respective abilities or similarities and differences. Thus, social capital might develop, where information about who knows (or prefers) what might then in turn influence subsequent interaction. Thus we think there is power in looking at the dynamic aspects of social interaction, where the conditioning relationships change over time as agents, and the system itself, adapts. In that sense, we feel that our proposed framework can be applied to a wide set of multi-agent settings and contexts.

References

1. Adler, P., Kwon, S.W.: Social capital: Prospects for a new concept. *The Academy of Management Review* 27(1), 17–40 (2002)
2. Bergstrom, T.C.: Systems of benevolent utility functions. *Journal of Public Economic Theory* 1, 71–100 (1999)
3. Bicchieri, C.: *Rationality and Coordination*. Cambridge University Press, Cambridge (1993)
4. Borgatti, S., Jones, C., Everett, M.: Network measures of social capital. *Connections* 21(2), 27–36 (1998)
5. Bourdieu, P., Wacquant, L.: *An invitation to reflexive sociology*. University of Chicago Press (1992)
6. Coase, R.: The nature of the firm. *Economica* 4(16), 386–405 (1937)
7. Coleman, J.: Social capital in the creation of human capital. *American Journal of Sociology* 94, S95–S120 (1988)
8. Easley, D., Kleinberg, J.: *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, Cambridge (2010)
9. Fearon, J., Laitin, D.: Explaining interethnic cooperation. *American Political Science Review*, 715–735 (1996)
10. Fehr, E., Schmidt, K.: A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817–868 (1999)
11. Golberg, D.E.: *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley (1989)
12. Goodin, R.: Laundering preferences. In: Elster, J., Hylland, A. (eds.) *Foundations of Social Choice Theory*, pp. 75–101. Cambridge Univ. Press, Cambridge (1986)
13. Jackson, M.O.: *Social and Economic Networks*. Princeton University Press, Princeton (2008)
14. Knudsen, T., Levinthal, D.A.: Two faces of search: Alternative generation and alternative evaluation. *Organization Science* 18(1), 39 (2007)

15. Kuran, T.: Ethnic norms and their transformation through reputational cascades. *Journal of Legal Studies* 27(2), 623–659 (1998)
16. Lichbach, M.: *The Rebel's Dilemma*. University of Michigan Press, Ann Arbor (1998)
17. Lin, N.: *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press, NY (2001)
18. Lin, N.: A network theory of social capital. In: Castiglione, D., van Deth, J.W., Wolleb, G. (eds.) *The Handbook of Social Capital*, pp. 50–69. Oxford University Press (2008)
19. Lohmann, S.: The dynamics of informational cascades: The monday demonstrations in leipzig, east germany, 1989-1991. *World Politics* 47, 42–101 (1994)
20. Luce, R.D., Raiffa, H.: *Games and Decisions*. John Wiley, New York (1957)
21. March, J., Simon, H.: *Organizations*. Blackwell, Oxford (1993)
22. Moghadam, A.: Palestinian suicide terrorism in the second intifada: Motivations and organizational aspects. *Studies in Conflict and Terrorism* 26(2), 65–92 (2003)
23. Nisan, N., Roughgarden, T., Tardos, E., Vazirani, V.V.: *Algorithmic Game Theory*. Cambridge University Press, Cambridge (2007)
24. Parsons, S., Wooldridge, M.: Game theory and decision theory in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 5, 243–254 (2002)
25. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo (1988)
26. Portes, A.: Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology* 24, 1–24 (1998)
27. Putnam, R.: *Bowling Alone: the Collapse and Revival of American Community*. Simon & Schuster, NY (2000)
28. Rawls, J.: *A Theory of Justice*. Harvard University Press, Cambridge (1971)
29. Shoham, Y., Leyton-Brown, K.: *Multiagent Systems*. Cambridge University Press, Cambridge (2009)
30. Shubik, M.: *Game Theory in the Social Sciences*. MIT Press, Cambridge (1982)
31. Shubik, M.: Game theory and operations research: Some musings 50 years later. Yale School of Management Working Paper No. ES-14 (May 2001)
32. Simon, H.A.: *Administrative behavior*. Free Press, New York (1947)
33. Sobel, J.: Interdependent preferences and reciprocity. *Journal of Economic Literature* XLIII, 392–436 (2005)
34. Stirling, W.: Conditional game theory: A generalization of game theory for cooperative multiagent systems. In: *Proceedings of The Third International Conference on Agents and Artificial Intelligence, Rome, Italy, Rome, Italy*, pp. 345–352 (2011)
35. Stirling, W.: *Theory of Conditional Games*. Cambridge University Press, Cambridge (in press, 2012)
36. von Neumann, J., Morgenstern, O.: *The Theory of Games and Economic Behavior*. Princeton Univ. Press, Princeton (1944); 2nd ed. (1947)
37. Weibull, J.W.: *Evolutionary Game Theory*. MIT Press, Cambridge (1995)
38. Weiss, G. (ed.): *Multiagent Systems*. MIT Press, Cambridge (1999)
39. Wood, E.: *Insurgent Collective Action and Civil War in El Salvador*. Cambridge University Press, New York (2003)
40. Zenger, T., Felin, T., Bigelow, L.: Theories of the firm–market boundary. *The Academy of Management Annals* 5(1), 89–133 (2011)

Swayed by Friends or by the Crowd?

Zeinab Abbassi¹, Christina Aperjis², and Bernardo A. Huberman²

¹ Department of Computer Science, Columbia University, New York, NY, USA

² Social Computing Group, HP Labs Palo Alto, CA

zeinab@cs.columbia.edu,

{christina.aperjis,bernardo.huberman}@hp.com

Abstract. We have conducted three empirical studies of the effects of friend recommendations and general ratings on how online users make choices. We model and quantify how a user deciding between two choices trades off an additional rating star with an additional friend's recommendation when selecting an item. We find that negative opinions from friends are more influential than positive opinions, and people exhibit "more random" behavior in their choices when the decision involves less cost and risk. Our results are quite general in the sense that people across different demographics trade off recommendations from friends and ratings from the general public in a similar fashion.

1 Introduction

When making choices, people use information from a number of sources including friends, family, experts, media, and the general public. Two sources that are particularly relevant in an online setting are the opinions of friends and ratings from the general public. Friends are believed to influence choices of their friends. In many cases, however, recommendations from one's friends are in contrast to opinions of individuals in the general public who are not one's friends.

In this paper we study how an online user's decision is influenced by recommendations from friends and ratings from the general public, particularly when these two sources of information are in conflict with each other. This question is interesting for two reasons. First, understanding how people trade off friends' opinions with ratings from the general public helps to determine the weight assigned by consumers to these two sources when they are uncertain about choosing one of two possible options. Second, this information can be used when designing algorithms that display these two sources of information in order to increase the probability of a user selecting one of the options. For example, an online social network platform that has information about how a user's friends and the general public have rated two different items can display to the user the item that she is more likely to select. On the other hand, if users tend to disregard some source of information, this source need not be shown to the user. Finally, an advertiser that wishes to make the user choose a certain item may strategically choose which pieces of information to show.

Specifically, focusing on friends and the general public as two components of social influence is important because these sources of social information are already used in a variety of algorithms and applications online. Social recommender systems take

into account the actions of a user's friends and make recommendations accordingly. Social search is also gaining more attention. Google recently launched its +1 button for search results and ads in order to improve its search algorithm. If a user thinks that a search result or an ad is useful she can click on the +1 button. The +1 will be displayed along with the user's name in the search results to all her friends who subsequently search a similar query. For users who are not friends, only the number of +1's will be displayed. Facebook uses a similar approach for business pages with the intention of getting higher click-through rates. The model that we suggest can be leveraged to design better algorithms for these and other similar applications.

In particular, in this paper we ask the following questions:

- How much are one's choices influenced by the opinions of her friends compared to ratings from the general public? What mathematical model predicts this?
- Do friends' negative opinions have a stronger or weaker effect than friends' positive opinions about an item?
- Do friends' opinions have the same effect on one's decision in higher risk situations versus lower risk situations?

To answer the above questions, we performed user studies on Mechanical Turk involving around 350 participants using positive and negative opinions from friends, as well as ratings from the general public; the latter was represented by the average number of stars. We find that the choice between two options fits a logit model. Our major contributions are (1) Our model is able to predict the probability of selection of an item by a user given two choices when recommendations from friends and star ratings from the general public is displayed, (2) We find that negative opinions from friends are more influential than positive opinions, and (3) We observe that people exhibit more random behavior in their choices when the decision involves less cost and risk. Our results are quite general in the sense that people across different demographics trade off recommendations from friends and ratings from the general public in a similar fashion.

2 Related Work

A number of empirical studies have considered the effect of social influence in various contexts, including prescription drug adoption and use [1], viral and word of mouth marketing [2,3,4], health plans [5], crime rates [6], online graphical perception tasks [7] and investment in the stock market [8,9]. Tucker et al. focus on how quality and popularity influence decisions on a wedding website [10]. Salganik et al. study the effects of social influence over time on the popularity of songs in an artificial online music market [11]. Guo et al. study the effect of messages from friends in online shopping [12].

Some previous works have compared recommendation strategies that are based on friends with recommendations based on similar users who are not necessarily friends (collaborative filtering) [13,14,15]. These papers found that recommendations based on friends behavior or direct suggestions were more useful to the users. Since the users did not know which recommendations came from friends and which from collaborative filtering, social influence could not be measured in these studies.

The existing literature does not consider the effect of multiple sources of social influence at the same time. Each of the aforementioned papers studies a specific source of social influence (e.g., friends or the general public). However, different groups of people have different levels of influence on one's decision. With the availability of social network data, in many online settings, recommendations from friends are available in addition to ratings and reviews from other people. In this work we focus on these two sources of social influence.

3 Method

In this section we describe the experimental design and report some statistics about the data we collected. Our goal is to study how people trade off information from friends and the general public when choosing between two items. Moreover, our experiments allow us to compare a setting where the information from friends consists of positive recommendations to a setting where the information from friends consists of negative opinions.

Furthermore, we compare people's choices with respect to two types of decisions: one that involves a monetary cost (booking a hotel) and a low risk decision that involves no monetary cost (watching a movie trailer). We chose booking hotels because the user cannot go and check it out before deciding and should rely on the information she gets from others. Similarly, a user may not have any information about a movie trailer before she watches it. We can think of the setting with the movie trailers as a less serious decision, since it involves less cost (just a couple of minutes of one's time) and risk. Users often make choices of this type online, e.g., when watching Youtube videos, clicking on a link or ad, etc.

In total, we conducted three user studies: booking a hotel with positive recommendations from friends (Study 1), booking a hotel with negative opinions from friends (Study 2) and watching a movie trailer with positive recommendations from friends (Study 3).

To collect the data we conducted the three studies with 350 participants each in the form of surveys on Amazon's Mechanical Turk (MTurk) during July and August 2011. MTurk is a crowdsourcing online marketplace where *requesters* use human intelligence of *workers* to perform certain tasks, also known as HITs (Human Intelligence Tasks). Workers browse among existing tasks and complete them for a monetary payment set by the requester [16]. Once a worker completes the task, the requester can decide whether to approve it. In particular, if the requester believes that the worker did not complete the task correctly, he can reject her work. In that case, the worker does not get paid for the particular task and her approval rate is decreased.

A number of papers discuss how to use MTurk to conduct behavioral research in a variety of disciplines [16][17]. Horton et al. use MTurk to replicate three classic economics experiments and confirm their results [18]. Heer and Bostock demonstrate that MTurk perception experiments are viable and contribute new insights for visualization design [19].

For our studies, we only hired workers that had approval rates of over 95%, that is, workers who had performed well in the past. We asked each worker to put herself in the following hypothetical situation: she is about to book a hotel (resp. watch a movie

trailer) an on e-commerce site (resp. online), and among the options, she has come down to two between which she is indifferent. The website has an underlying social network of friends (or it runs on top of an online social network). For each of the two options, we provide the following information:

- (i) the overall rating (in terms of stars on the scale of 1 to 5) based on ratings from a large number of previous customers (resp. users) in the case of selecting which hotel to book (resp. which movie trailer to watch)
- (ii) the number of friends who recommend (resp. have negative opinions about) the option in the case of positive (resp. negative) recommendations

For each question, the option that has more stars is the one that is less recommended by friends; that is, we did not use a pair of options where one clearly dominated the other.

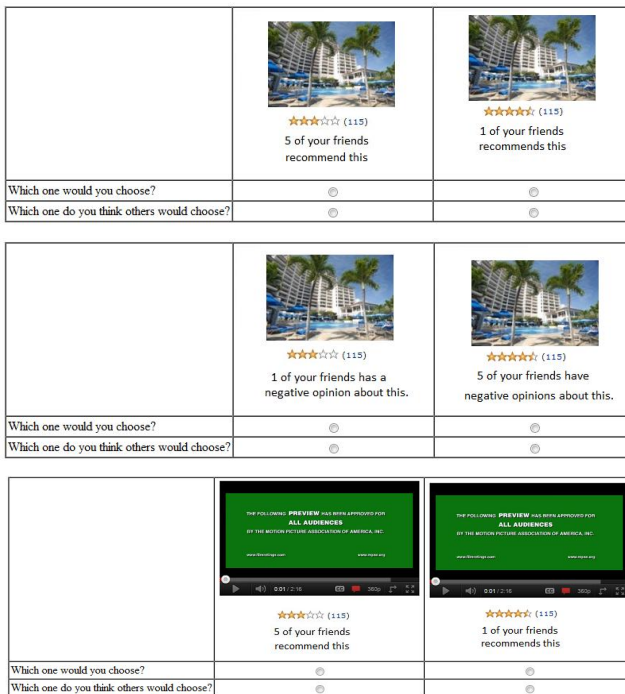


Fig. 1. Sample questions for Studies 1, 2, and 3 (top to bottom)

A sample question from each survey is shown in Figure 1. Each question consists of two parts: 1) “Which one would you choose?” and 2) “Which one do you think others would choose?” The answer to the first question provides information on how a particular worker trades off information from friends and the general public when making her choice. We only use the data from this question for our analysis.

¹ The data is available upon request from the authors.

Even though we do not use the answers to the second question for our analysis, we included it in the survey for two reasons. First, by asking the second question the subjects felt that the survey's goal was to study how people think their decisions are similar to the decisions of others, weakening the reactivity effect [20]. Second, we offered three bonus payments (\$5 or \$10 each, which is 50 or 100 times the amount we paid for each HIT) to the three workers whose answers to the second question was closest to the others' answers to the first question in order to incentive workers to answer the survey carefully.

Apart from using the bonus to incentivize workers to put some thought when answering the survey, we incorporated two "validity check" questions in the survey. If a worker did not answer these two questions correctly, we did not include any of her responses in our analysis (and rejected her work). In the first such question one option clearly dominated the other in terms of both the number of stars and friends' recommendations (one option had only one star and 10 negative recommendations from friends, whereas the other had more stars and 10 positive recommendations from friends). The second test question was a repeated question with the order of choices reversed and without graphics. Overall, we rejected 33% of the responses across all 3 studies because they were invalid. The average completion time for each valid HIT was 174.8 seconds while the average completion time for the invalid HITS was 153.3; this suggests that the workers that were rejected had not taken the task as seriously as the rest of the workers.

In addition to the "validity check" questions, each study consisted of 8 questions (with the format of Figure 1) which we use in our analysis and 3 demographics questions asking about the gender, the age and the education level of the respondent. Overall, 36% of the approved respondents, were female. Other demographic information for the approved respondents according to the self-reporting of the workers are shown in Tables 1 and 2.

Table 1. Age Distribution

| Age | Percentage |
|---------------|------------|
| age < 20 | 6 |
| 20 < age < 30 | 59 |
| 30 < age < 40 | 24 |
| 40 < age < 50 | 8 |
| 50 < age < 60 | 3 |
| age > 60 | 0 |

Table 2. Education Level Distribution

| Education Level | Percentage |
|---------------------|------------|
| Highschool graduate | 19 |
| Associates Degree | 4 |
| Bachelors Degree | 53 |
| Graduate Degree | 24 |

4 Results

For each question, there are two options that the worker can select from, which we refer to as option 1 and option 2. We denote the number of stars S_i and the number of friends' recommendations by F_i , for $i = 1, 2$. To predict the probability that option 1 is selected, we conduct a logistic regression² on our dataset of choices with the difference in the number of friends (i.e., $F_1 - F_2$) and the number of stars (i.e., $S_1 - S_2$) for each question as the predictor variables. We denote the corresponding coefficients by α_f and α_s respectively.

We also ran the logistic regression with an intercept, but the intercept was not statistically significant. This is good news, since a statistically significant intercept in this case would imply bias (that is, the position in which an option is presented would affect the probability that it is selected). We next report the results from each survey separately.

4.1 Study 1: Positive Opinions for Hotels

Model 1. We first only considered the difference in the number of stars and friends as predictor variables. The estimated coefficients along with other parameters are shown in Table 3, and as can be seen both are statistically significant. Observe that both coefficients are positive; this is intuitive, since more stars (resp. more positive recommendations) indicate that the option is better and thus the worker is more likely to select it. Finally, the pseudo $-R^2$ for this model³ is 0.95, indicating that the fit is very good.

Interpretation of the Coefficients. We first interpret the coefficients for our model in terms of marginal effects on the odds ratio. The odds ratio measures the probability that the dependent variable is equal to 1 relative to the probability that it is equal to zero. For the logit model, the log odds of the outcome is modeled as a linear combination of the predictor variables; therefore, the odds ratio of a coefficient is equal to $\exp(\text{coefficient})$. Since $\alpha_s = 0.735$, we conclude that a unit increase in $S_1 - S_2$, multiplies the initial odds ratio by $\exp(0.735) = 2.07$. For the friends predictor variable, the odds ratio is equal to $\exp(0.204) = 1.22$. Another way to interpret the coefficients is in terms of relative change in the probability when there is one unit of change in one of the predictor variables while other parameters remain the same. In this case the relative probability increases by at most 10% with a unit change in $F_1 - F_2$ and by at most 35% with a unit change in $S_1 - S_2$.

² We note that a number of other empirical studies also use the logit choice function to model social influence [21,65].

³ We computed Efron's pseudo $-R^2$ which is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

where, N is the number of observations in the model, y is the dependent variable, \bar{y} is the mean of the y values, and $\hat{\pi}$ is the probabilities predicted by the logit model. The numerator of the ratio is the sum of the squared differences between the actual y values and the predicted π probabilities. The denominator of the ratio is the sum of squared differences between the actual y values and their mean [22].

Table 3. Study 1: Positive Opinions for Hotels

| Predictor | Estimated Coefficients | z-value |
|------------|------------------------|---------|
| α_f | 0.20471*** (0.027) | 7.597 |
| α_s | 0.73549*** (0.050) | 14.307 |

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo- $R^2 = 0.95$

To further assess the predictive power of the model, we performed **cross validation**. We left out one question at a time and estimated the coefficients using the remaining questions. Then, we predicted the probabilities for the question that was left out. The predicted values were very close in all cases with absolute mean difference of 0.021. The actual values and their differences can be found in Table 4.

Table 4. Cross validation for study 1

| Left out question | Actual | Predicted | Difference |
|-------------------|--------|-----------|------------|
| Q1 | 0.54 | 0.53 | 0.01 |
| Q2 | 0.58 | 0.55 | 0.03 |
| Q3 | 0.71 | 0.62 | 0.09 |
| Q4 | 0.74 | 0.74 | 0.00 |
| Q5 | 0.77 | 0.77 | 0.00 |
| Q6 | 0.74 | 0.72 | 0.02 |
| Q7 | 0.82 | 0.83 | 0.01 |
| Q8 | 0.54 | 0.53 | 0.01 |

Finally, we used one of the questions of this survey twice in Amazon's Mechanical Turk (in two separate HITS) in order to see whether workers would react to the question in similar ways. We found that the percentage of workers that chose the first option of the question was similar in both cases (26% versus 24%), further validating our approach.

Model 1'. In Model 1', we included all self reported demographic information as predictor variables in addition to the stars and friends' recommendation variables. This information includes: gender, age (in five 10-year brackets from 20 to 60 years old: called age1-5), and education level (high school, associates degree, bachelors degree, and graduate degree: called edu1-3). More specifically, we coded the following variables as dummy variables. The estimated coefficients and other information is shown in the second column of Table 5. As can be seen in Table 5, these extra coefficients are not statistically significant. This suggests that people in different demographics trade off ratings from the public and friends' recommendations similarly.

Table 5. Studies 1 and 2 with demographic information

| Predictor | Study 1 | Study 2 |
|---------------------|-----------------|----------------|
| α_f | 0.263*** (0.19) | 0.35*** (0.07) |
| α_s | 0.793*** (0.08) | 0.66*** (0.12) |
| gender _f | -0.006 (0.07) | -0.02 (0.06) |
| gender _s | 0.31 (0.16) | -0.13 (0.11) |
| edu1 _s | 0.11 (0.16) | -0.01 (0.11) |
| edu1 _f | 0.04 (0.08) | -0.02 (0.07) |
| edu2 _s | -0.11 (0.46) | 0.16 (0.40) |
| edu2 _f | -0.16 (0.23) | 0.18 (0.63) |
| edu3 _s | -0.05 (0.16) | -0.03 (-3.1) |
| edu3 _f | -0.02 (0.08) | -0.24 (-0.31) |
| age1 _s | 0.061 (0.39) | -0.12 (0.15) |
| age1 _f | 0.02 (0.21) | -0.14 (0.11) |
| age2 _s | 0.06 (0.16) | 0.17 (0.12) |
| age2 _f | 0.11 (0.08) | 0.16 (0.08) |
| age3 _s | -0.08 (0.34) | -0.05 (0.16) |
| age3 _f | -0.02 (0.18) | -0.13 (0.10) |
| age4 _s | -0.05 (0.16) | 0.10 (0.11) |
| age4 _f | -0.17 (0.08) | 0.04 (0.07) |
| age5 _s | 0.16 (0.40) | 0.05 (0.24) |
| age5 _f | 0.03 (0.22) | 0.15 (0.17) |

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.2 Study 2: Negative Opinions for Hotels

Model 2. In this section we look at negative opinions from friends — instead of positive recommendations. In particular, each option is characterized by the number of stars (based on information from the general public) as well as the number of friends who have negative opinions about it. We run a logistic regression and report the results in Table 6. As can be seen in the table both variables are statistically significant and the pseudo- R^2 measure for this model is 0.95 which implies that the model is a good fit. Moreover, as we would expect, the friends coefficient is negative in this case, as more negative opinions from friends decrease the probability that the worker selects an option.

Table 6. Study 2: Negative Opinions for Hotels

| Predictor | Estimated Coefficients | z-value |
|------------|------------------------|---------|
| α_f | -0.281*** (0.030) | 9.378 |
| α_s | 0.503*** (0.050) | 10.018 |

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo- $R^2 = 0.95$

Interpretation of the Coefficients. Similarly to Study 1, we interpret the coefficients for our model in terms of marginal effects on the odds ratio. For the present model (negative recommendations), the fact that $\alpha_s = 0.503$ means that one unit increase in $S_1 - S_2$, multiplies the initial odds ratio by $\exp(0.503) = 1.65$. In other words, the odds of choosing option 1 increases by 65%. For the friends predictor variable, the odds ratio is equal to $\exp(-0.281) = 0.75$, which means that the odds of selecting option 1 decreasing by 25%. Equivalently, the relative odds of selecting option 1 when $F_1 - F_2$ decreases by one unit is $(\exp(0.281) - 1) \approx 32\%$. Another way to interpret the coefficients is by looking at the relative changes in the probability of choosing each option. In this case the relative probability decreases by at most 14% with a unit change in $F_1 - F_2$ and by at most 26% with a unit change in $S_1 - S_2$.

For this study we did **cross validation** as well to test the predictive power of our model. The results are shown in Table 7. The predicted and actual values are very close (mean absolute difference = 0.231).

Table 7. Cross validation for study 2

| Left out question | Actual | Predicted | Difference |
|-------------------|--------|-----------|------------|
| Q1 | 0.30 | 0.25 | 0.05 |
| Q2 | 0.39 | 0.41 | 0.02 |
| Q3 | 0.43 | 0.44 | 0.01 |
| Q4 | 0.54 | 0.53 | 0.01 |
| Q5 | 0.58 | 0.60 | 0.02 |
| Q6 | 0.38 | 0.39 | 0.01 |
| Q7 | 0.40 | 0.42 | 0.02 |
| Q8 | 0.45 | 0.50 | 0.05 |

Model 2'. Similarly to Model 1', in this model we include all variables: stars, friends' opinions, and demographics information in the model. The results are shown in the last column of Table 5. As for Model 1', the estimated demographic coefficients are not statistically significant, meaning that the addition of demographic information does not improve the predictive power (compared to Model 2). In other words, individuals choose between options in these situations similarly across all demographics.

4.3 Study 3: Positive Opinions for Movie Trailers

Model 3. Our third study considers the effect of positive recommendations from friends in a low risk decision: choosing which movie trailer to watch. We perform a logistic regression and report the estimated coefficients in Table 8. The estimated coefficients are statistically significant; however, in this case pseudo- R^2 is 0.61 which is lower than the pseudo- R^2 's for previous models Models 1 and 2 (0.95). The coefficients for stars and friends are $\alpha_s = 0.349$ and $\alpha_f = 0.167$. As for Model 1, both coefficients are positive, since people are more likely to select an option if it has more stars and/or more positive recommendations from friends. Therefore, the odds ratio for the number of stars is

1.41 and for the number of friends' recommendations is 1.18. By computing the relative probability changes, we conclude that an additional star increases the probability of selecting that option by 18%, whereas an additional recommendation from a friend increases the probability by 8%.

Table 8. Study 3: Positive Opinions for Movie Trailers

| Predictor | Estimated Coefficients | z-value |
|------------|------------------------|---------|
| α_f | 0.167*** (0.049) | 7.101 |
| α_s | 0.349*** (0.027) | 6.014 |

Note: Standard errors are shown in parentheses.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo- $R^2 = 0.61$

5 Discussion

This paper studies how positive and negative opinions from friends affect our decisions compared to ratings from the crowd for different types of decisions. Our three user studies result in some interesting conceptual findings about the tradeoff between these two types of social influence.

First, negative opinions from friends are more influential on one's decision than positive opinions. We can see this by comparing the odds ratios of Study 1 and Study 2, in which the number of positive and negative friends' opinions are shown respectively: the odds ratio for the friends variable is higher in Study 2 (1.32 versus 1.22 where the difference is statistically significant with $p = 0.046$), whereas the odds ratio for the stars variable is higher in Study 1 (2.07 versus 1.65 where the difference is statistically significant with $p = 0.001$). In other words, one less negative opinion from a friend has a larger effect than one more positive opinion, whereas one more star increases the odds of an option being chosen less in the case that negative opinions from friends are present. Such an asymmetry between the effect of negative and positive actions and opinions have been studied in the social psychology literature [23,24,25,26,27]. The *positive-negative asymmetry effect* has been observed in many domains such as impression formation [28], information-integration paradigm [29] and prospect theory for decision making under risk [30]. The finding in all the above cited work is that negativity has stronger effects than equally intense positivity. Our results confirm this finding in online settings.

Second, people exhibit more random behavior when the decision involves less cost and less risk. We can see this by comparing the results from Study 1 and Study 3, where the decisions are "which hotel to book" and "which movie trailer to watch" respectively. Booking a hotel clearly involves a monetary cost and some risk, whereas the worse thing that can happen with a movie trailer is to waste a couple of minutes of one's time. The odds ratios are lower in Study 3 than Study 1 (1.18 versus 1.22 for friends with $p = 0.345$ which is not statistically significant, and 1.41 versus 2.07 for stars where the difference is statistically significant with $p < 0.0001$). This implies that one added star has a smaller influence on one's decision in the case of movie trailers. However, one added friend has basically the same influence as in the case of hotels.

Moreover, the fraction of respondents choosing either option is closer to half compared to the hotel booking surveys. This implies that the choices were more random in this case, which may be explained by the fact that choosing which movie trailer to watch is a less important/serious decision than booking a hotel.

Third, we observe that in all three studies one more star increases the probability of selecting that option more than one more (resp. less) friend in the case of positive (resp. negative) recommendations. Equivalently, the odds ratio of the stars' coefficient is larger than the odds ratio of the friends' coefficient (2.07 versus 1.22, 1.65 versus 1.32 and 1.41 versus 1.18 for studies 1, 2 and 3 respectively where for all three studies the differences are very statistically significant.) This does not mean that the number of friends' positive or negative recommendations does not influence decisions; on the contrary, an additional recommendation (resp. one less negative opinion) from friends changes the probability by at least 18% across all three studies. The fact that an additional star has a larger effect than an additional friend opinion is reasonable if we consider that the number of stars is bounded between 1 and 5, whereas the number of friends' recommendations may take values from a larger range.

Forth, for all of our user studies, we find out that the demographic variables (gender, age, and education level) do not significantly impact the choice that is made, implying that people across different demographics trade off recommendations from friends and ratings from the crowd in a similar way. It also implies that our predictive model and results are generalizable across different demographics.

5.1 Practical Implications

Our studies offer insights that can be useful in various online domains such as recommender systems, social search results ranking, online advertisement placement, online social network newsfeed rankings, and social shopping websites. In these applications when both friends' recommendations and ratings from the general public are available, our estimated model can help the platform determine which option to display or what ranking to display the options for a given user.

As an example, consider a specific user that is searching for a hotel on a booking website. There are two hotels that match the user's search criteria, hotel *A* and hotel *B*. Assume that hotel *A* has 3 stars from customer ratings and 4 (positive) recommendations from the user's friends, and hotel *B* has 4 stars but only 2 (positive) recommendations from friends. According to the results of Study 1, the user is more likely to prefer hotel *B* (if everything else is equal). Thus, if the booking website does not have any additional information about the user's preferences, it should recommend hotel *B* to the user, or equivalently rank hotel *B* higher than hotel *A* if it provides personalized ranking of hotels to the user. Such personalization benefits both the user and the booking website by improving user experience and increasing the chances that the user books a hotel through the website..

The same ideas can be applied to recommender systems based on collaborative filtering and in particular social recommender systems. Social recommender systems leverage the actions of one's friends to determine which items to recommend. The choice and ranking of the items can be obtained using our model. The same is true for social

search. Finally, a marketer that wishes to maximize the probability that a user selects a given item may be able to strategically select what information to show to the user.

5.2 Limitations

In our studies, users could only see the *number* of friends that had positive or negative opinions about an item — and not the names of the corresponding friends. We focused on the number of friends, because in this way we can get more general qualitative results. Moreover, given that people tend to have a large number of friends in online social networks, showing the number of friends (instead of specific names) may be a good way to avoid privacy concerns. Nevertheless, we note that opinions from specific friends could have a different effect than the number of friends, e.g., [31].

Another limitation of our work is that we only consider the difference of star ratings and friends' recommendations between the two options in our analysis. This was sufficient for the purposes of the current paper — given we got our simple logistic regression was a very good fit. Of course, the *absolute* numbers of friends' recommendations and stars could also play a role and there could also be interesting non-linear effects.

Finally, we on purpose set the number of people on which the number of stars is based on equal to a large value (115), which we kept constant across all questions and studies, in order to focus on the effect of the other parameters. The dependence of one's choice on the number of people that the stars are based on is an interesting future direction.

6 Conclusion

Our study of how online users make choices based on information from friend recommendations and ratings from the general public is important for a range of online applications in particular social search results ranking, recommender systems, online advertisement placement, online social network newsfeed rankings, and social shopping websites. When both friends' recommendations and ratings from the general public are available, our estimated model can help the platform determine which option to display or in what ranking to display the options for a given user.

Our results offer insights that can be useful in various online domains. Specifically we found that negative opinions from friends are more influential than positive opinions, and people show more random behavior in their choices when lower cost or risk is incurred.

While this paper focuses on two sources of information, namely friends' opinions and ratings from the general public, our approach can also be applied to the study of how individuals trade off information from other sources, such as experts, celebrities, and the media.

References

1. Iyengar, R., Van den Bulte, C., Valente, T.W.: Opinion leadership and social contagion in new product diffusion. *Marketing Science* 30, 195–212 (2011)
2. Aral, S., Walker, D.: Creating social contagion through viral product design: A randomized trial of peer influence in networks. In: *Proceedings of the 31th Annual International Conference on Information Systems* (2010)

3. Lelis, S., Howes, A.: Informing decisions: how people use online rating information to make choices. In: Proceedings of the, Annual Conference on Human Factors in Computing systems, pp. 2285–2294. ACM (2011)
4. Luca, M.: Reviews, reputation, and revenue: The case of yelp. com. Harvard Business School Working Papers (2011)
5. Sorensen, A.: Social learning and health plan choice. *The RAND Journal of Economics* 37(4), 929–945 (2006)
6. Glaeser, E., Sacerdote, B., Scheinkman, J.: Crime and social interactions. National Bureau of Economic Research, Cambridge (1995)
7. Hullman, J., Adar, E., Shah, P.: The impact of social information on visual judgments. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, pp. 1461–1470. ACM (2011)
8. Al-Hasan, A., Viswanathan, S.: The new roi return on influentials. Working Paper (2010)
9. Hong, H., Kubik, J., Stein, J.: Social interaction and stock-market participation. *The Journal of Finance* 59(1), 137–163 (2004)
10. Tucker, C., Zhang, J.: How does popularity information affect choices? a field experiment. *Management Science* (forthcoming 2011)
11. Salganik, M., Dodds, P., Watts, D.: Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762), 854 (2006)
12. Guo, S., Wang, M., Leskovec, J.: The role of social networks in online shopping: information passing, price of trust, and consumer choice. In: ACM Conference on Electronic Commerce (2011)
13. Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S.: Personalized recommendation of social software items based on social relations. In: *RecSys*, pp. 53–60 (2009)
14. Groh, G., Ehmig, C.: Recommendations in taste related domains: collaborative filtering vs. social filtering. In: *GROUP*, pp. 127–136 (2007)
15. Sinha, R.R., Swearingen, K.: Comparing recommendations made by online systems and friends. In: *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries* (2001)
16. Mason, W., Suri, S.: Conducting behavioral research on amazon’s mechanical turk. *Behavior Research Methods*, 1–23 (2010)
17. Kittur, A., Chi, E., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 453–456. ACM (2008)
18. Horton, J.J., Rand, D.G., Zeckhauser, R.: The online laboratory: Conducting experiments in a real labor market. *Experimental Economics* (2011)
19. Heer, J., Bostock, M.: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 203–212. ACM, New York (2010)
20. Heppner, P., Wampold, B., Kivlighan, D.: *Research design in counseling*. Brooks/Cole Pub. Co. (2008)
21. Paez, A., Scott, D., Volz, E.: A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design* 35(6), 1055–1069 (2008)
22. Hardin, J., Hilbe, J., Hilbe, J.: Generalized linear models and extensions. Stata Corp. (2007)
23. Baumeister, R., Bratslavsky, E., Finkenauer, C., Vohs, K.: Bad is stronger than good. *Review of General Psychology* 5(4), 323 (2001)
24. Peeters, G., Czapinski, J.: Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology* 1, 33–60 (1990)

25. Taylor, S.: Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin* 110(1), 67 (1991)
26. Cheung, C., Lee, M.: Online consumer reviews: Does negative electronic word-of-mouth hurt more? In: *AMCIS 2008 Proceedings*, p. 143 (2008)
27. Hao, Y., Ye, Q., Li, Y., Cheng, Z.: How does the valence of online consumer reviews matter in consumer decision making? differences between search goods and experience goods. In: *2010 43rd Hawaii International Conference on System Sciences (HICSS)*, pp. 1–10. IEEE (2010)
28. Anderson, N.: Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology* 70(4), 394 (1965)
29. Anderson: *Foundations of Information Integration Theory*. Academic Press, New York (1981)
30. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291 (1979)
31. Golbeck, J., Hendler, J.: Filmtrust: Movie recommendations using trust in web-based social networks. In: *Proceedings of the IEEE Consumer Communications and Networking Conference*, vol. 96. Citeseer (2006)

Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 U.S. Republican Presidential Primaries

Lu Chen, Wenbo Wang, and Amit P. Sheth

Kno.e.sis Center, Wright State University, Dayton, OH 45435 USA
{chen,wenbo,amit}@knoesis.org

Abstract. Existing studies on predicting election results are under the assumption that all the users should be treated equally. However, recent work [14] shows that social media users from different groups (e.g., “silent majority” vs. “vocal minority”) have significant differences in the generated content and tweeting behavior. The effect of these differences on predicting election results has not been exploited yet. In this paper, we study the spectrum of Twitter users who participate in the on-line discussion of 2012 U.S. Republican Presidential Primaries, and examine the predictive power of different user groups (e.g., highly engaged users vs. lowly engaged users, right-leaning users vs. left-leaning users) against Super Tuesday primaries in 10 states. The insights gained in this study can shed light on improving the social media based prediction from the user sampling perspective and more.

Keywords: Electoral Prediction, Twitter Analytics, Social Intelligence, User Categorization, Engagement Degree, Tweet Mode, Content Type, Political Preference.

1 Introduction

Over 80% of Americans use at least one social network, and people spend nearly 23% of their online time on social networks¹. Among those popular social network sites, Twitter has over 140 million active users, generating over 340 millions tweets per day². The topics being discussed in social networks cover almost every aspect of our lives. On one hand, researchers are making every effort to make sense of the social data to understand what is going on in the world. On the other hand, there is a surge of interest in building systems that harness the power of social data to predict what is about to happen. It has been reported that social data is used to predict box-office revenues [1, 13], stock market [3, 9, 18], and election outcomes [2, 11, 15–17], etc.

Existing studies using social data to predict election results have focused on obtaining the measures/indicators (e.g., mention counts or sentiment of a party

¹ <http://www.socialmediaexaminer.com/26-promising-social-media-stats-for-small-businesses/>

² <http://blog.twitter.com/2012/03/twitter-turns-six.html>

or candidate) from social data to perform the prediction. They treat all the users equally, and ignore the fact that social media users engage in the elections in different ways and with different levels of involvement. A recent study [14] has shown that significant differences exist between silent majority (users who tweeted once) and vocal minority (users who tweet very often) in the generated content and tweeting behavior in the context of political elections. However, whether and how such differences will affect the prediction results still remains unexplored. For example, in our study, 56.07% of Twitter users who participate in the discussion of 2012 U.S. Republican Primaries post only one tweet. The identification of the voting intent of these users could be more challenging than that of the users who post more tweets. *Will such differences lead to different prediction performance?* Furthermore, the users participating in the discussion may have different political preference. *Is it the case that the prediction based on the right-leaning users will be more accurate than that based on the left-leaning users, since it is the Republican Primaries?* Exploring these questions can expand our understanding of social media based prediction, and shed light on using user sampling to further improve the prediction performance.

In this paper, we investigate above questions by studying different groups of social media users who engage in the discussions of elections, and comparing the predictive power among these user groups. Specifically, we chose the 2012 U.S. Republican Presidential Primaries on Super Tuesday³ among four candidates: Newt Gingrich, Ron Paul, Mitt Romney and Rick Santorum. We collected 6,008,062 tweets from 933,343 users talking about these four candidates in an eight week period before the elections. All the users are characterized across four dimensions: engagement degree, tweet mode, content type, and political preference. We first investigated the user categorization on each dimension, and then compared different groups of users with the task of predicting the results of Super Tuesday races in 10 states. Instead of using tweet volume or the overall sentiment of tweet corpus as the predictor, we estimated the “vote” of each user by analyzing his/her tweets, and predicted the results based on “vote-counting”. The results were evaluated in two ways: (1) the accuracy of predicting winners, and (2) the error rate between the predicted votes and the actual votes for each candidate.

The main contributions of this paper are as follows. (1) We group social media users based on their participation (engagement degree, tweet mode, and content type) as well as political preference, and study the participation behaviors of different user groups, (2) we present a method to predict the “vote” of a user based on the analysis of his/her tweets, and count the votes of users to predict the election result, and (3) we examine the predictive power of different user groups in predicting the results of Super Tuesday races in 10 states.

2 Related Work

Using social media data for electoral prediction has attracted increasing interest in recent years. Gayo-Avello [7] provided a comprehensive summary of literature

³http://en.wikipedia.org/wiki/Super_Tuesday

on election prediction with Twitter data. Here, we focus on the literature which is most relevant to our task.

O'Connor et al. [15] discovered correlations between public opinion derived from presidential job approval polls and sentiment based on analysis of Twitter messages. Tumasjan et al. [17] used the number of tweets mentioning a party or candidate to accurately predict the 2009 German federal elections. Sang et al. [16] showed that merely counting the tweets is not sufficient for electoral predictions, and the prediction could be improved by improving the quality of data collection and performing sentiment analysis. Bermingham and Smeaton [2] used both sentiment-based and volume-based measures to predict results of the 2011 Irish General Election. They found that social analytics using both measures were predictive, and volume was a stronger indicator than sentiment.

Meanwhile, some researchers argue that the predictive power of social media might be exaggerated, and the challenges of building the predictive models based on social data have been underestimated, especially for the electoral predictions. Gayo-Avello [8] showed that simple approaches based on mention counts and polarity lexicons failed in predicting the result of 2008 U.S. Presidential Elections. In another study [12], the authors found that the social data did only slightly better than chance in predicting the 2010 U.S. Congressional elections. They pointed out the need of obtaining a random sample of likely voters in order to achieve accurate electoral predictions.

To summarize, existing studies on electoral prediction have focused on exploring the measures and indicators (e.g., tweet volume or sentiment) to predict the election results, and left the problem that whether all the users and their tweets should be treated equally unexplored. Previous research [14] has shown that different groups of users could be very different in tweeting behavior and generated content. Should a user who posts one tweet be handled in the same way as another user who posts 100 tweets in predicating the election? Should a democrat be treated equally as a republican in predicting the republican primaries? We focus on exploring such questions in this paper.

3 User Categorization

Using Twitter Streaming API, we collected tweets that contain the words “gin-grich”, “romney”, “ron paul”, or “santorom” from January 10th 2012 to March 5th 2012 (Super Tuesday was March 6th). Totally, the dataset comprises 6,008,062 tweets from 933,343 users. The data used for this study is collected as part of a social web application – Twitris⁴, which provides real-time monitoring and multi-faceted analysis of social signals surrounding an event (e.g., the 2012 U.S. Presidential Election). In this section, we discuss user categorization on four dimensions, and study the participation behaviors of different user groups.

⁴ <http://twitris.knoesis.org/>

Table 1. User Groups with Different Engagement Degrees

| Engagement Degree | Very Low | Low | Medium | High | Very High | Total |
|-------------------|----------|---------|----------|-----------|-----------|-------|
| Tweets per User | 1 | [2, 10] | [11, 50] | [51, 300] | >300 | |
| User Volume | 56.07% | 35.93% | 6.19% | 1.58% | 0.23% | 100% |
| Tweet Volume | 8.71% | 20.31% | 20.42% | 26.83% | 23.73% | 100% |

3.1 Categorizing Users by Engagement Degree

We use the number of tweets posted by a user to measure his/her engagement degree. The less tweets a user posts, the more challenging the user’s voting intent can be predicted. An extreme example is to predict the voting intent of a user who posted only one tweet. Thus, we want to examine the predictive power of different user groups with various engagement degrees.

Specifically, we divided users into the following five groups: the users who post only one tweet (*very low*), 2-10 tweets (*low*), 11-50 tweets (*medium*), 51-300 tweets (*high*), and more than 300 tweets (*very high*). Table 1 shows the distribution of users and tweets over five engagement categories. We found that more than half of the users in the dataset belong to the *very low* group, which contributes only 8.71% of the tweet volume, while the very highly engaged group contributes 23.73% of the tweet volume with only 0.23% of all the users. It raises the question of whether the tweet volume is a proper predictor, given that a small group of users can produce a large amount of tweets.

To further study the behaviors of the users on different engagement levels, we examined the usage of hashtags and URLs in different user groups (see Table 2). We found that the users who are more engaged in the discussion use more hashtags and URLs in their tweets. Since hashtags and URLs are frequently used in Twitter as ways of promotion, e.g, hashtags can be used to create trending topics, the usage of hashtags and URLs reflects the users’ intent to attract people’s attention on the topic they discuss. **The more engaged users show stronger such intent and are more involved in the election event.** Specifically, only 22.95% of all tweets created by very lowly engaged users contain hashtags, this proportion increases to 39.45% in the *very high* engagement group. In addition, the average number of hashtags per tweet (among the tweets that contain hashtags) is 1.43 in the *very low* engagement group, while this number is 2.68 for the very highly engaged users. The users who are more engaged also use more URLs, and generate less tweets that are only text (not containing any hashtag or URL). We will see whether and how such differences among user engagement groups will lead to varied results in predicting the elections later.

3.2 Categorizing Users by Tweet Mode

There are two main ways of producing a tweet, i.e., creating the tweet by the user himself/herself (original tweet) or forwarding another user’s tweet (retweet). Original tweets are considered to reflect the users’ attitude, however, the reason

Table 2. Usage of Hashtags and URLs by Different User Groups

| Engagement Degree | Very Low | Low | Medium | High | Very High |
|----------------------|----------|--------|--------|--------|-----------|
| Tweets with Hashtags | 22.95% | 26.98% | 30.58% | 32.85% | 39.45% |
| Hashtags per tweet | 1.43 | 1.58 | 1.95 | 2.14 | 2.68 |
| Tweets with URLs | 33.44% | 40.16% | 49.02% | 53.88% | 59.89% |
| Only Text | 50.93% | 43.11% | 34.19% | 29.35% | 25.31% |

for retweeting can be varied, e.g., to inform or entertain the users’ followers, to be friendly to the one who created the tweet, etc., thus retweets do not necessarily reflect the users’ thoughts. It may lead to different prediction performance between the users who post more original tweets and the users who have more retweets, since the voting intent of the latter is more difficult to recognize.

According to users’ preference on generating their tweets, i.e., tweet mode, we classified the users as *original tweet-dominant*, *original tweet-prone*, *balanced*, *retweet-prone* and *retweet-dominant*. A user is classified as *original tweet-dominant* if less than 20% of all his/her tweets are retweets. Each user from *retweet-dominant* group has more than 80% of all his/her tweets that are retweets. In Table 3, we illustrate the categorization, the user distribution over the five categories, and the tweet mode of users in different engagement groups.

Table 3. User Distribution over Categorization of Tweet Mode

| Tweet Mode | Orig. Tweet-Dom. | Orig. Tweet-Prone | Balanced | RT-Prone | RT-Dom. | Total |
|------------|------------------|-------------------|------------|------------|---------|-------|
| Retweet | <20% | [20%, 40%) | [40%, 60%) | [60%, 80%) | >=80% | |
| All Users | 49.04% | 4.76% | 7.22% | 4.27% | 34.71% | 100% |
| Very Low | 55.32% | 0.00% | 0.00% | 0.00% | 44.68% | 100% |
| Low | 41.04% | 9.83% | 16.70% | 8.81% | 23.62% | 100% |
| Medium | 42.01% | 15.41% | 14.78% | 13.21% | 14.59% | 100% |
| High | 38.44% | 15.21% | 16.62% | 15.39% | 14.35% | 100% |
| Very High | 31.89% | 13.88% | 17.03% | 17.73% | 19.47% | 100% |

It is interesting to find that the *original tweet-dominant* group accounts for the biggest proportion of users in every user engagement group, and this proportion declines with the increasing degree of user engagement (55.32% of very lowly engaged users are original tweet-dominant, while only 31.89% of very highly engaged users are original tweet-dominant). It is also worth noting that **a significant number of users (34.71% of all the users) belong to the *retweet-dominant* group, whose voting intent might be difficult to detect.**

3.3 Categorizing Users by Content Type

Based on content, tweets can be classified into two classes – opinion and information (i.e., subjective and objective). Studying the difference between the users who post more information and the users who are keen to express their opinions

could provide us with another perspective in understanding the effect of using these two types of content in electoral prediction.

We first identified whether a tweet represents positive or negative opinion about an election candidate. We used the approach proposed in [4] to learn a candidate-specific sentiment lexicon from the tweet collection. This lexicon contained sentiment words and phrases which were used to express positive or negative opinions about the candidates. Totally, this lexicon comprised 1674 positive words/phrases and 1842 negative words/phrases, which was applied to recognize the opinions about each candidate in tweets. If a tweet contained more positive (negative) words than negative (positive) words about a candidate, e.g., Mitt Romney, it was annotated as “positive(negative)_Mitt_Romney”. If there were no sentiment words found in a tweet about a candidate, e.g., Mitt Romney, it was annotated as “neutral_Mitt_Romney”. Thus, every tweet has four sentiment labels (one for each candidate). “I want Romney to win over Santorum but you must be careful in your negative ads.” was labeled as “neutral_Newt_Gingrich”, “neutral_Ron_Paul”, “positive_Mitt_Romney”, and “negative_Rick_Santorum”.

The tweets that are positive or negative about any candidate are considered *opinion* tweets, and the tweets that are neutral about all the candidates are considered *information* tweets. We also used a five-point scale to classify the users based on whether they post more opinion or information with their tweets: *opinion-dominant*, *opinion-prone*, *balanced*, *information-prone* and *information-dominant*. Table 4 shows the user distribution among all the users, and the users in different engagement groups categorized by content type.

The users from *very low* engagement group have only one tweet, so they either belong to *opinion-dominant* (39%) or *information dominant* (61%). With users’ engagement increasing from low to very high, the proportions of *opinion-dominant*, *opinion-prone* and *information-dominant* users dramatically decrease from 11.09% to 0.05%, 11.75% to 0.42%, and 27.40% to 0.66%, respectively. In contrast, the proportions of *balanced* and *information-prone* users grow. In *high* and *very high* engagement groups, the *balanced* and *information-prone* users together accounted for more than 95% of all users. It shows the tendency that **more engaged users post a mixture of content, with similar proportion of opinion and information, or larger proportion of information.**

3.4 Identifying Users’ Political Preference

Since we focused on the Republican Presidential Primaries, it should be interesting to compare two groups of users with different political preferences – left-leaning and right-leaning. Some efforts [6, 10, 5] have been made to address the problem of predicting the political preference/orientation of Twitter users in recent years. In our study, we use a simple but effective method to identify the left-leaning and right-leaning users.

We collected a set of Twitter users with known political preference from Twel-
low⁵. Specifically, we acquired 10,324 users who are labeled as Republican, conservative, Libertarian or Tea Party as right-leaning users, and 9,545 users who are

⁵ <http://www.twellow.com/>

Table 4. User Distribution over Categorization of Content Type

| Content Type | Opinion-Dom. | Opinion-Prone | Balanced | Info.-Prone | Info.-Dom. | Total |
|--------------|--------------|---------------|------------|-------------|------------|-------|
| Opinion | >=80% | [60%, 80%) | [40%, 60%) | [20%, 40%) | <20% | |
| All Users | 25.89% | 4.74% | 14.75% | 9.92% | 44.70% | 100% |
| Very Low | 39.00% | 0.00% | 0.00% | 0.00% | 61.00% | 100% |
| Low | 11.09% | 11.75% | 30.92% | 18.84% | 27.40% | 100% |
| Medium | 0.59% | 8.02% | 42.85% | 38.60% | 9.94% | 100% |
| High | 0.22% | 1.43% | 53.84% | 42.06% | 2.45% | 100% |
| Very High | 0.05% | 0.42% | 58.98% | 39.89% | 0.66% | 100% |

labeled as Democrat, liberal or progressive as left-leaning users. We denote the top 1000 left-leaning users and top 1000 right-leaning users who have the most followers as L_I and R_I , respectively. Among the remaining users that are not contained in L_I or R_I , there are 1,169 left-leaning users and 2,172 right-leaning users included in our dataset, and these 3,341 users are denoted as T .

The intuitive idea is that a user tends to follow others who share the same political preference as his/hers. The more right-leaning users one follows, the more likely that he/she belongs to the right-leaning group. Among all the users that a user is following, let N_l be the number of left-leaning users from L_I and N_r be the number of right-leaning users from R_I . We estimated the probability that the user is left-leaning as $\frac{N_l}{N_l+N_r}$, and the probability that the user is right-leaning as $\frac{N_r}{N_l+N_r}$. The user is labeled as left-leaning (right-leaning) if the probability that he/she is left-leaning (right-leaning) is more than a threshold τ . Empirically, we set $\tau = 0.6$ in our study. We tested this method on the labeled dataset T and the result shows that this method correctly identified the political preferences of 3,088 users out of all 3,341 users (with an accuracy of 0.9243).

Totally, this method identified the political preferences of 83,934 users from all of the 933,343 users in our dataset. Other users may not follow any of the users in L_I or R_I , or follow similar numbers of left-leaning and right-leaning users, thus their political preferences could not be identified. Table 5 shows the comparison of left-leaning and right-leaning users in our dataset. We found that **right-leaning users were more involved in this election event in several ways**. Specifically, the number of right-leaning users was two times more than that of left-leaning users, and the right-leaning users generated 2.65 times the number of tweets as the left-leaning users. Compared with the left-leaning users, the right-leaning users tended to create more original tweets and used more hashtags and URLs in their tweets. This result is quite reasonable since it was the Republican election, with which the right-leaning users are supposed to be more concerned than the left-leaning users.

4 Electoral Prediction with Different User Groups

In this section, we examine the predictive power of different user groups in predicting the Super Tuesday election results in 10 states. We first recognized the users from each state. There are two types of location information from

Table 5. Comparison between Left-leaning and Right-leaning Users

| Political Preference | Left-Leaning | Right-Leaning |
|----------------------|--------------|---------------|
| # of Tweets | 702,178 | 1,863,186 |
| # of Users | 27,586 | 56,348 |
| Tweets per User | 25.5 | 33.1 |
| Original Tweets | 48.46% | 56.09% |
| Retweets | 51.54% | 43.91% |
| Tweets with Hashtags | 33.02% | 37.99% |
| Hashtags per Tweet | 1.68 | 1.93 |
| Tweets with URLs | 45.95% | 52.75% |
| Only Text | 34.57% | 30.19% |
| Opinion | 41.31% | 41.47% |

Twitter – the geographic location of a tweet, and the user location in the profile. We utilized the background knowledge from LinkedGeoData⁶ to identify the states from user location information⁷. If the user’s state could not be inferred from his/her location information, we utilized the geographic locations of his/her tweets. A user was recognized as from a state if his/her tweets were from that state. Table 6 illustrates the distribution of users and tweets among the 10 Super Tuesday states. We also compared the number of users and tweets in each state to its population. The Pearson’s r for the correlation between the number of users/tweets and the population is 0.9459/0.9667 ($p < .0001$). In the following of this section, we first describe how we estimated a user’s vote, and next report the prediction results, followed by a discussion of the results.

4.1 Estimating a User’s Vote

To answer the question that for whom a user will vote, we need to find for which candidate the user shows the most support. We think there are two indicators that can be extracted from a user’s tweets of one candidate – mention and sentiment. Intuitively, people show their support for celebrities through frequently talking about them and expressing positive sentiments about them.

As described in Section 3.3, we have analyzed each user’s tweets, identified which candidate is mentioned, and whether a positive or negative opinion is expressed towards a candidate in a tweet. For each user, let N be the number of all his/her tweets, $N_m(c)$ be the number of tweets in which he/she mentioned a candidate c , $N_{pos}(c)$ be the number of positive tweets about c from the user, $N_{neg}(c)$ be the number of negative tweets about c from the user. We define the user’s support score for c as:

$$\begin{cases} (1 - \frac{N_{neg}(c)}{N_{pos}(c)+\beta}) \times \frac{N_m(c)}{N} & \text{if } N_{pos}(c) + N_{neg}(c) > 0 \\ \gamma \times \frac{N_m(c)}{N} & \text{otherwise} \end{cases}$$

⁶ <http://linkedgedata.org/About>

⁷ Since geographical analysis is not the focus of this paper, we did not verify if the users are actually from the locations specified in their profiles.

Table 6. Distribution of Tweets and Users over 10 Super Tuesday States

| | | | | | |
|-------------|------------|-----------|-----------|---------------|--------------|
| U.S. State | Alaska | Georgia | Idaho | Massachusetts | North Dakota |
| # of Tweets | 7,633 | 88,555 | 17,331 | 89,842 | 3,763 |
| # of Users | 736 | 13,210 | 1,830 | 15,009 | 661 |
| Population | 722,718 | 9,815,210 | 1,584,985 | 6,587,536 | 683,932 |
| | Ohio | Okalahoma | Tennessee | Vermont | Virginia |
| # of Tweets | 102,880 | 27,747 | 58,384 | 5,525 | 73,172 |
| # of Users | 18,066 | 3,965 | 7,980 | 1,183 | 9,796 |
| Population | 11,544,951 | 3,791,508 | 6,403,353 | 626,431 | 8,096,604 |

where β ($0 < \beta < 1$) is a smoothing parameter, and γ ($0 < \gamma < 1$) is used to discount the score when the user does not express any opinion towards c ($N_{pos}(c) = N_{neg}(c) = 0$). We used $\beta = \gamma = 0.5$ in our study. According to this definition, the more positive tweets (less negative tweets) are posted about c , and the more c is mentioned, the higher the user’s support score for c is. After calculating a user’s support score for every candidate, we selected the candidate who received the highest score as the one that the user will vote for.

4.2 Prediction Results

In this section, we report the comparison of different user groups in predicting Super Tuesday races, and discuss our findings.

To predict the election results in a state, we used only the collection of users who are identified from that state. Then we further divided each user collection of one state over four dimensions – engagement degree, tweet mode, content type, and political preference. In order to get enough users in one group, we used a more coarse-grained classification instead of the five-point scales described in the section of User Categorization. To be specific, we classified users as three different groups according to their engagement degree: *very low*, *low*, and *high**. The *very low* and *low* engagement groups are the same as what we have defined in Section 3.1. The *high** engagement group comprises the users who post more than 10 tweets (i.e., the aggregation of the medium, high and very high groups defined previously). Based on the tweet mode, the users were divided into two groups: *original tweet-prone** and *retweet-prone**, depending on whether they post more original tweets or more retweets. Similarly, the users were classified as *opinion-prone** or *information-prone** according to whether they post more opinions or more information. The *right-leaning* users and *left-leaning* users were also identified from the user collection of each state. In all, for each state, there were nine user groups over four different dimensions.

We also considered users in different time windows. Our dataset contains the users and their tweets discussing the election in 8 weeks prior to the election day. We wanted to see whether it will make any difference to use the data in different time windows. Here we examined four time windows – *7 days*, *14 days*, *28 days* or *56 days* prior to the election day. For example, the *7 days* window

is from February 28th to March 5th. In a specific time window, we assessed a user’s vote using only the set of tweets he/she creates during this time⁸.

With each group of users in a specific state and a specific time window, we counted the users’ votes for each candidate, and the one who received the most votes was predicted as the winner of the election in that state. The performance of a prediction was evaluated in two ways: (1) whether the predicted winner is the actual winner, and (2) comparing the predicted percentage of votes for each candidate with his actual percentage of votes, and getting the mean absolute error (MAE) of the four candidates.

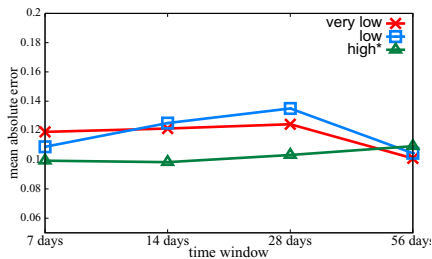
Table 7 shows the accuracy of winner prediction by different user groups in different time windows. The accuracy was calculated as $\frac{N_{state}^{true}}{N_{state}^{pred}}$, in which N_{state}^{true} was the number of states where the winner was correctly predicted, and N_{state}^{pred} (= 10) was the number of all Super Tuesday states. Figure 1 illustrates the average MAE of the predictions in 10 states by different user groups in different time windows. From Table 7 and Figure 1, we do see that different user groups on each dimension show varied prediction performance.

Table 7. The Accuracy of Winner Prediction by Different User Groups

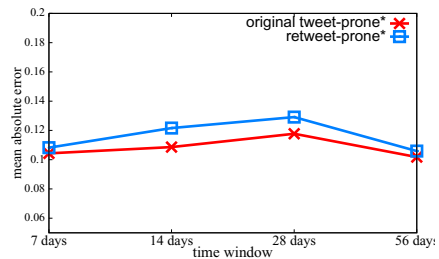
| | 7 Days | 14 Days | 28 Days | 56 Days |
|-----------------------------|--------|---------|---------|---------|
| Engagement Degree | | | | |
| Very Low | 0.5 | 0.4 | 0.3 | 0.6 |
| Low | 0.7 | 0.3 | 0.3 | 0.6 |
| High* | 0.5 | 0.8 | 0.5 | 0.6 |
| Tweet Mode | | | | |
| Original Tweet-Prone* | 0.7 | 0.4 | 0.4 | 0.6 |
| Retweet-Prone* | 0.6 | 0.3 | 0.3 | 0.6 |
| Content Type | | | | |
| Opinion-Prone* | 0.5 | 0.5 | 0.4 | 0.6 |
| Information-Prone* | 0.6 | 0.4 | 0.3 | 0.7 |
| Political Preference | | | | |
| Left-Leaning | 0.5 | 0.2 | 0.3 | 0.6 |
| Right-Leaning | 0.5 | 0.7 | 0.7 | 0.8 |

As shown in Table 7, the *high** engagement group correctly predicted the winners of 5 states in 7 day, 8 in 14, 5 in 28 and 6 in 56 day time windows, respectively, which is slightly better than the average performance of *very low* and *low* engagement groups. In addition, the average prediction error of *high** engagement group is smaller than that of *very low* and *low* engagement groups in three out of the four time windows (see Figure 1a). Comparing two user groups over the tweet mode dimension, *original tweet-prone** group beat the *retweet-prone** group by achieving better accuracy on winner prediction and smaller prediction error in almost all the time windows (see Figure 1b). The two user groups categorized by content type also show differences in predicting the

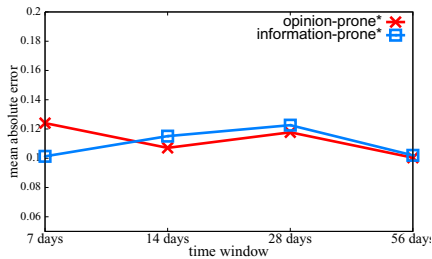
⁸ A user’s vote might be varied in different time windows, since we used different sets of tweets for the assessment.



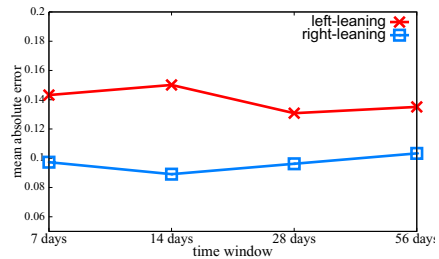
(a) User Groups Categorized by Engagement Degree



(b) User Groups Categorized by Tweet Mode



(c) User Groups Categorized by Content Type



(d) User Groups Categorized by Political Preference

Fig. 1. The Mean Absolute Error (MAE) by Different User Groups in 10 States

elections, but the difference is not as clear as that of user groups on other dimensions. On winner prediction, the *opinion-prone** group achieved better accuracy in 14 day and 28 day time windows, and *information-prone** group achieved better accuracy in 7 day and 56 day time windows. Although the prediction error of *opinion-prone** group was smaller than that of *information-prone** group in three time windows, the gap was quite small (see Figure 1(c)).

It is interesting to find that, among all the user groups, the *right-leaning* group achieved the best prediction results. In Table 7 and Figure 1(d), *right-leaning* group correctly predicted the winners of 5, 7, 7 and 8 states (out of 10 states) in 7 day, 14 day, 28 day and 56 day time windows, respectively. Furthermore, it also showed the smallest prediction error (<0.1) in three out of four time windows among all the user groups. In contrast, the prediction by the *left-leaning* group was the least accurate. In the worst case, it correctly predicted the winners in only 2 states (in the 14 day time window), and its prediction error was over 0.15.

To further verify our observation, we looked at the average prediction error of four time windows for each state, and applied paired t-test to find whether the difference of the average prediction errors in 10 states between a pair of user groups was statistically significant. The test showed that the difference between *right-leaning* and *left-leaning* user groups is statistically highly significant ($p < .001$). The difference between *low* and *high** engagement user groups was also found statistically significant ($p < .01$). However, the difference between *original tweet-prone** and *retweet-prone**, or between *opinion-prone** and *information-prone** was not significant.

In addition, we also compared our results with random predictions. From the winner prediction perspective, all the user groups except the *left-leaning* one beat the random baseline (25% accuracy) in all the time windows. The random baseline showed a mean prediction error (of vote percentage) over 0.13, which is higher than that of all the user groups except the *left-leaning* one.

4.3 Discussion

There are at least two factors that could affect the accuracy of electoral prediction. Firstly, whether the prediction of users' votes is accurate. Secondly, whether the users' opinion is representative of the actual voters' opinion. We interpret the varied prediction results with different user groups based on these two factors.

In our study, the *high** engagement user group achieved better prediction results than *very low* and *low* engagement groups. It may be due to two reasons. Firstly, high engagement users posted more tweets. Since our prediction of a user's vote is based on the analysis of his/her tweets, it should be more reliable to make the prediction using more tweets. Secondly, according to our analysis, more engaged users showed stronger intent and were more involved in the election event. It might suggest that users in the *high** engagement group were more likely to vote, compared with the users in the *very low* and *low* engagement groups.

However, the *low* engagement group did not show better performance compared with the *very low* engagement group. One possible explanation might be that the users from these two groups are not that different. A more fine-grained classification of users with different engagement degrees might provide more insight. Since the prediction is state-based, we could not get enough users in each group (especially the groups of highly engaged users) if we divided users into more groups. It is worth noting that more than 90% of all the users in our dataset belonged to *very low* and *low* engagement groups. Accurately predicting the votes of these users is one of the biggest challenges in electoral prediction.

The results also show that the prediction based on users who post more original tweets is slightly more accurate than that based on users who retweet more, although the difference is not significant. It may be due to the difficulty of identifying users' voting intent from retweets. In most of the current prediction studies, original tweets and retweets are treated equally with the same method. Further studies are needed to compare these two types of tweets in prediction, and a different method might be needed for identifying users' intent from retweets. In addition, a more fine-grained classification of users according to their tweet mode could provide more insight.

No significant difference is found between the *opinion-prone** and the *information-prone** user groups in prediction. It suggests that the likely voters cannot be identified based on whether users post more opinions or more information. It also reveals that the prediction of users' votes based on more opinion tweets is not necessarily more accurate than the prediction using more information tweets.

The *right-leaning* user group provides the most accurate prediction result, which is significantly better than that of the *left-leaning* group. In the best case (56 day time window), the right-leaning user group correctly predict the winners

in 8 out of 10 states (Alaska, Georgia, Idaho, Massachusetts, Ohio, Oklahoma, Vermont, and Virginia). It is worth noting that this result is significantly better than the prediction result of the same elections based on Twitter analysis reported in the news article⁹, in which the winners are correctly predicted in only 5 out of 10 states (Georgia, Idaho, Massachusetts, Ohio, Virginia). Since the elections being predicted were Republican primaries, the attitude of *right-leaning* users could be more representative of the voters' attitude. To some extent, it demonstrates the importance of identifying likely voters in electoral prediction.

This study can be further improved from several aspects. First, more effort could be made to investigate the possible data biases (e.g., spam tweets and political campaign tweets) and how they might affect the results. Second, we estimated the vote intent of each Twitter user in our dataset and aggregated them to predict the election results. However, these users are not necessarily the actual voters. Identification of the actual voters from social media is also an interesting problem to explore. In addition, our work examined the predictive power of different user groups in republican primaries, thus some of our findings may not apply to other elections of different natures, e.g., general elections. However, we believe the general principle that Twitter users are not equal in predictions is common for all elections.

5 Conclusion

In this paper, we studied the spectrum of Twitter users in the context of the 2012 U.S. Republican Presidential Primaries, and examined the predictive power of different user groups in predicting the results from the 10 states that held Republican primaries on Super Tuesday. We divided users into different groups on four dimensions – engagement degree, tweet mode, content type, and political preference. To predict the election results, we first predicted each user's vote based on analyzing the mentions and sentiments of the candidates in the user's tweets, and then counted the votes received by each candidate from every user group. Comparing the prediction results obtained by different user groups, we found the result achieved by right-leaning users was significantly better than that achieved by left-leaning users. The prediction based on highly engaged users was better than that based on lowly engaged users. The users who posted more original tweets provided slightly higher accuracy in the prediction than the users who retweeted more did. To some extent, these findings demonstrate the importance of identifying likely voters and user sampling in electoral predictions.

Acknowledgment. We are grateful to Ashutosh Jadhav, Hemant Purohit, Pavan Kapanipathi and Pramod Anantharam for helpful discussions, Sarasi Lalith-sena and Sujana Udayanga for insightful comments. We also thank the anonymous reviewers for their useful comments. This research was supported by US NSF grant IIS-1111182: SoCS: Social Media Enhanced Organizational Sensemaking in Emergency Response.

⁹ <http://www.usatoday.com/tech/news/story/2012-03-07/election-social-media/53402838/1>

References

1. Asur, S., Huberman, B.A.: Predicting the future with social media. Arxiv preprint arXiv:1003.5699 (2010), <http://arxiv.org/abs/1003.5699>
2. Birmingham, A., Smeaton, A.F.: On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the Sentiment Analysis where AI meets Psychology Workshop at IJCNLP (2011)
3. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. *Journal of Computational Science* (2011)
4. Chen, L., Wang, W., Nagarajan, M., Wang, S., Sheth, A.P.: Extracting Diverse Sentiment Expressions with Target-dependent Polarity from Twitter. In: Proceedings of ICWSM (2012)
5. Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Proceedings of the IEEE 3rd International Conference on Social Computing, pp. 192–199 (2011)
6. Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., Menczer, F.: Political polarization on twitter. In: Proceedings of ICWSM, pp. 89–96 (2011)
7. Gayo-Avello, D.: I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper. Arxiv preprint arXiv:1204.6441 (2012), <http://arxiv.org/abs/1204.6441>
8. Gayo-Avello, D.: Don't turn social media into another 'Literary Digest' poll. *Communications of the ACM* 54(10), 121–128 (2011)
9. Gilbert, E., Karahalios, K.: Widespread worry and the stock market. In: Proceedings of ICWSM, pp. 229–247 (2010)
10. Golbeck, J., Hansen, D.: Computing political preference among twitter followers. In: Proceedings of the Annual Conference on Human Factors in Computing Systems, pp. 1105–1108 (2011)
11. Livne, A., Simmons, M.P., Adar, E., Adamic, L.A.: The party is over here: Structure and content in the 2010 election. In: Proceedings of ICWSM (2011)
12. Metaxas, P.T., Mustafaraj, E., Gayo-Avello, D.: How (Not) to predict elections. In: Proceedings of the IEEE 3rd International Conference on Social Computing, pp. 165–171 (2011)
13. Mishne, G., Galance, N.: Predicting movie sales from blogger sentiment. In: AAAI Symposium on Computational Approaches to Analysing Weblogs (2006)
14. Mustafaraj, E., Finn, S., Whitlock, C., Metaxas, P.T.: Vocal minority versus silent majority: Discovering the opinions of the long tail. In: Proceedings of the IEEE 3rd International Conference on Social Computing, pp. 103–110 (2011)
15. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of ICWSM, pp. 122–129 (2011)
16. Sang, E.T.K., Bos, J.: Predicting the 2011 Dutch Senate Election Results with Twitter. In: Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, pp. 53–60 (2012)
17. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of ICWSM, pp. 178–185 (2010)
18. Zhang, X., Fuehres, H., Gloor, P.A.: Predicting Stock Market Indicators Through Twitter 'I hope it is not as bad as I fear'. *Procedia-Social and Behavioral Sciences* 26, 55–62 (2011)

Web Page Recommendation Based on Semantic Web Usage Mining

Soheila Abrishami¹, Mahmoud Naghibzadeh², and Mehrdad Jalali¹

¹Department of Computer Engineering, Azad University of Mashhad, Mashhad, Iran
{soheilaabrishami, jalali}@mshdiau.ac.ir

²Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
naghibzadeh@um.ac.ir

Abstract. The growth of the web has created a big challenge for directing the user to the Web pages in their areas of interest. Meanwhile, web usage mining plays an important role in finding these areas of interest based on user's previous actions. The extracted patterns in web usage mining are useful in various applications such as recommendation. Classical web usage mining does not take semantic knowledge and content into pattern generations. Recent researches show that ontology, as background knowledge, can improve pattern's quality. This work aims to design a hybrid recommendation system based on integrating semantic information with Web usage mining and page clustering based on semantic similarity. Since the Web pages are seen as ontology individuals, frequent navigational patterns are in the form of ontology instances instead of Web page addresses, and page clustering is done using semantic similarity. The result is used for generating web page recommendations to users. The recommender engine presented in this paper which is based on semantic patterns and page clustering, creates a list of appropriate recommendations. The results of the implementation of this hybrid recommendation system indicate that integrating semantic information and page access sequence into the patterns yields more accurate recommendations.

Keywords: Web usage mining, semantic web, ontology, web page recommendation, page clustering.

1 Introduction

In semantic web content and information is interpretable and understandable not only by humans, but also by computers. In order to support the user in his task, the Web should be enriched by the machines' ability to process the information [1]. Therefore, the Web content and objects should be semantically introduced into the machine world by using ontologies. Ontologies have been created to facilitate knowledge reuse and sharing in the decentralized and distributed context of the Web [5].

The rapidly growing amount of information on the Web causes difficulties for the Internet users to find desired information and due to the huge volume of data and lack of structure in many Web sites finding relevant information on the Web has become a

real challenge. Therefore, the research field of Web usage mining has gained notable consideration for finding user behavioral patterns. Web usage mining is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web logs. One of the most important disadvantages of the current approaches in Web usage mining is that the result is produced in terms of Web pages (i.e. web page addresses); hence, there is no semantic meaning of the common navigation profile. Another disadvantage of classical web usage mining is called the new-item problem, which is the failure to recommend newly added pages or products to the visitors since these products or pages are not in the current common navigation profiles. To overcome this, the common navigation profile can be extracted in terms of semantic information so that the common navigation profile will be in ontological terms and concepts. Therefore, newly added items can be recommended to the user as long as this item's concept is in the common navigation profiles.

From that point, Web site content has started to play a more important role in the Web usage mining process, relying only on Web usage data for user modeling can be inefficient. In recent years, there have been studies where semantic knowledge systems are used in order to provide further improvement in accuracy.

The extracted patterns in web usage mining are useful in several different areas, including recommendation, Web site restructuring, prefetching, etc. Recommendation is a heavily studied research subject, where web usage patterns can improve the accuracy of the task.

In this work, we present a hybrid web recommender system, which incorporates semantic knowledge and sequence information into pattern generation and clusters Web pages by using a constructed ontology for Web site. The information of clustering is used for identifying of irrelevant pages in recommendation set. In [3] a hybrid recommender system is proposed, which integrates document clustering and semantic knowledge with web usage mining. In another hybrid system [4] user sessions are clustered and the navigational patterns are generated without using semantic information and then semantic features, which extracted from domain specific ontologies combined with the navigational patterns. However, in our work, semantic information is used within pattern generation, rather than in postprocessing, and we cluster the web pages based on semantic similarity among individuals in ontology which is created for the Web site. The results of the implementation of our hybrid recommender system show a significant improvement on the quality of recommendation engine output.

The rest of the paper is organized as follows. In Section 2, related work is presented. In Section 3, the details of the proposed approach are described. Section 4 includes the experimental work. Finally, Section 5 presents the concluding remarks.

2 Related Work

In the area of web usage mining, there are various approaches such as clustering, association rule mining, sequence mining, and sequence alignment to mine the data in the server logs. However, those works which combine web usage mining and

semantic web are limited. Bettina Berendt, Andreas Hotho and Gerd Stumme [1, 17] are the authors of one of the first studies of web usage mining on the Semantic Web. In [16] they show how the Semantic Web can improve web usage mining, and how usage mining can help to build up the Semantic Web. In their work, they assumed that the server log contains terms of ontology concepts and individuals, so the mining algorithms like clustering and association rule mining can be applied on it, but there are no details about generation of pattern by using semantic information.

Mobasher et al., [9] have provided a system for incorporating domain ontologies with Web usage mining and in this work the emphasis is on the personalization process and the frequent pattern generation is done by clustering.

The work presented by Adda et al., [14] proposes using metadata about the content that they assume is stored in domain ontology to enhance the quality of the discovered patterns. Authors use two-level taxonomy, while the first taxonomy branches between concepts, the second one branch between relations among concepts, and this increases the time complexity of the algorithm considerably.

Recent studies presented in [13, 18] aim to enhance association rules with domain ontologies. In [18] each web server log file entry converts into a single ontology concept. After applying SPADE algorithm on the converted log file, sequential association rules are generated. Unlike standard sequential rule miners, this approach yields rules that have ontological objects as antecedent and consequent and the result is used for generating web page recommendations to the visitor. On the other hand, in the work presented in [13], the emphasis is on improving the time efficiency of online next page prediction, rather than pattern quality. In [19] they integrate semantic web into web usage mining and clustering of sessions is used for extracting navigational patterns.

3 Proposed Approach

The general architecture of the system has the basics of classical web usage mining systems. The initial steps include the data acquisition and data cleaning. These steps are followed by offline mining part and lastly, the recommendation phase which is the online part of the system.

As illustrated in Fig. 1, the proposed system consists of 4 phases: preprocessing, rule extraction, page clustering, and recommendation. In the rest of this section, the steps of the process are described in more detail.

3.1 Preprocessing

For Web Log Preprocessing, the entries of the Web server log is cleaned, transactions are extracted, and ontology class individuals are mapped to the Web page addresses. The preprocessing method which is used in [2] is chosen for the preprocessing phase our system.

The first phase of the log file parsing is the pruning of the non-responded web requests and eliminating requests made by software agents such as Web crawlers.

After pruning, the navigation history of each session from the log file is extracted. For extraction of navigation history User Identification and Session Identification must be done.

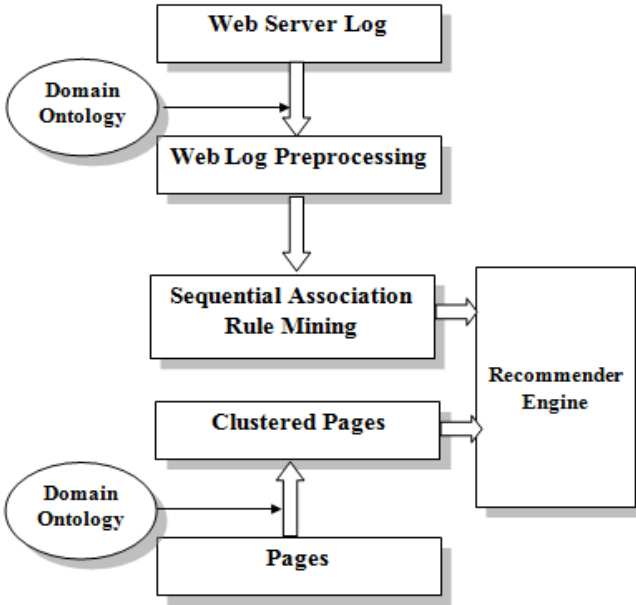


Fig. 1. General Framework of Overall System Design

We identify users with their IP address registered in each record in log file. For Session Identification, we should define session timeout. We assumed 20 minutes session timeout for the experimental procedure.

In order to prevent any postprocessing, in the last step, each page access in every transaction is mapped to the ontology instances defined in the ontology for the web site. Since the automatic extraction of semantic information is a hard task, and it is a research problem on its own, we accomplished this task manually. At first, we construct ontology in OWL associated with our web site by helping Web page structures. The construction of this basic ontology can be achieved through knowledge extraction from Web sites, based on content and structure mining. For mapping the suitable concepts and instances is chosen by considering the content of the Web pages.

Construction and creation of ontology are done using Protégé [11]. This ontology contains the concept of InfoResource, Science and FreeResource in first level. The Science concept includes a hierarchy relation in which higher lever categories include more details than lower level categories (for example Agriculture concept has a child named Forestry). After construction of ontology, the visited Web pages' URLs in navigation history are mapped to one or more individuals of the ontology. This mapping is done according to the semantic annotation of the web pages and web objects on the pages. At the end of this step, transactions consist of ontology individuals instead of pages' URLs.

3.2 Navigational Patterns Creation

After the preprocessing step, the next step is the extraction of frequent navigation patterns. Association rule mining¹ is used to discover interesting relations among items in a large database. Sequential association rule mining² is a limited version of association rule mining. In sequential association rule mining, items in both input (the dataset) and output (association rules) are presented in a sequential form (mostly ordered by a timestamp). In this research, we preferred to use sequential association rule mining since the sequence information in the navigation is retained in the generated patterns. Through sequential rule mining, the results will be in a sequential form that permits us to know which pages are visited and in what order. Since taxonomies are encountered in many ontology definitions and classes may have *is – a* relations, it is necessary to find generalized sequence patterns under a given taxonomy. In order to find generalized association rules, the dataset is extended in such a way that each item is also associated with the item’s taxonomical parent. For efficient and scalable mining of sequential patterns, PrefixSpan [7] is used. The PrefixSpan’s input includes transactions which each transaction is a sequence of events. A sequence is an ordered set of events, and event is a set of one or more items. In Fig. 2, a sequence database is shown.

| <i>Sequence_id</i> | <i>Sequence</i> |
|--------------------|-----------------------------------|
| 10 | $\langle a(abc)(ac)d(cf) \rangle$ |
| 20 | $\langle (ad)c(bc)(ae) \rangle$ |
| 30 | $\langle (ef)(ab)(df)cb \rangle$ |
| 40 | $\langle eg(af)abc \rangle$ |

Fig. 2. A sequence database [7]

SPMF [12] is an open-source data mining framework written in Java. For extracting sequential association rule mining, we used RuleGen [6] algorithm in this framework. So the input of this algorithm is an extended dataset, such taxonomies are introduced into the dataset, and each transaction is a sequence of events.

Some examples for the generated frequent sequence rules are presented in Fig. 3. The rule $\text{NaturalResource} \Rightarrow \text{Irrigation}, \text{LiveStock}, \text{NaturalResource}$ tells that visiting concept “NaturalResource” will be followed by visits to the concepts of Irrigation, LiveStock and NaturalResource.

¹ The support of rule $X \Rightarrow Y$ is the percentage of the transactions that contain both X and Y to all transactions. The confidence is the percentage of the transaction that contains Y to the transaction that contains X.

² For a sequential association rule, $a \rightarrow b \Rightarrow c$, $a \rightarrow b$ is the head of the rule and c is the body of the rule. When body is concatenated to the head, we end up with the sequence $a \rightarrow b \rightarrow c$. Therefore, support for this rule is ((number sessions including $a \rightarrow b \rightarrow c$) / (all transactions)) and confidence for this rule is ((number of sessions including $a \rightarrow b \rightarrow c$) / (number of sessions including $a \rightarrow b$)).

| <i>Sample Rules</i> |
|---|
| <i>NaturalResource => Irrigation , LiveStock , NaturalResource</i> |
| <i>Gardening => Gardening , Irrigation , LiveStock , NaturalResource</i> |
| <i>LiveStock , AgricultureEngineering => AgricultureEngineering , Foresty , Irrigation , LiveStock , NaturalResource</i> |
| <i>Irrigation => AgricultureEngineering , Foresty , Irrigation</i> |

Fig. 3. Sample rules for the library Web site

3.3 Page Clustering By Using Semantic Similarity

In this step, we clustered URLs which were visited by users. We used K-means [20] clustering algorithm for grouping the pages. As described before, in this work, we mapped pages into a set of ontological item sets, so we clustered pages by using semantic similarities between ontology's individuals. For example, we have pages P_1 and P_2 which are mapped into certain individuals.

$P1.html$ is mapped into $obj2, obj3, obj4$

$P2.html$ is mapped into $obj8, obj15, obj17$

Where each P_i is a unique page and each Obj_k is ontology individual.

In [19], an approach, which clusters user sessions that have been mapped into sequence of objects, is presented. In our research, we employed and modified this proposed approach to cluster the pages and used the result of this clustering in order to increase accuracy of our recommendation. In our method, although clustering is used, instead of sessions, pages are clustered. Since our clustering is based on semantic similarity between ontology's individuals, the result of page clustering reflects the similarity of pages. The outcome of clustering is used for identification of irrelevant pages in recommendation phase.

In order to cluster these pages we need two instruments:

- A distance metric to compute the distance between two ontology concepts.
- A distance metric to compute the distance between two ontology instances.

Since instances which are used in mapping are string types, we use "LevenshteinDistance" string comparison algorithm.

The distance between two concepts can be defined as a function of the distance between their attributes and their location in the ontological tree. We assume a concept to be a tuple $C = (A, L)$, where A is set of attributes $(a1, a2, a3, \dots, an)$ and L is a location representation in the taxonomy of the ontology. The distance between two concepts $C1$ and $C2$, $DIST(C1, C2)$, can be defined as the weighted sum of distances of object attributes and tree locations.

$$Dist(C1, C2) = DistA(att1, att2) * w1 + DistL(loc1, loc2) * w2 \quad (1)$$

where $DistA$ is a function that returns the distance between two sets of attributes and $DistL$ is a function that returns the distance between two locations in a tree; $w1$ and $w2$ are weights of the distance functions. For $DistL$ we use the approach by [8] that is based on the positions of the concepts in the concept taxonomy. For $DistA$ function, we employ the string distance definition “LevenshteinDistance”.

By using these metrics for computing distances in the K-means algorithm, the pages are clustered. At the end of this phase, we have clustered URLs according to their semantic information.

3.4 Recommender Engine

This is the final component of the system. It combines the analysis of the usage mining and Web page clustering and produces recommendations for current users. The current user’s navigation path is compared to all sequential association rules to produce recommended pages. For each recommended page, the page is checked to determine in which cluster it is located. After defining maximum clusters, which are clusters with a larger number of pages, the pages of these clusters are added to the final recommendation set.

Since the association rules are composed of ontology individuals, the user navigation history is converted into the sequence of ontology instances. In the recommendation phase, firstly, according to the `window_count` (`window_count` is a parameter which defines the maximum number of previously visited pages which should be used in order to recommend a new page) navigated items are taken as the search pattern. The association rules and user navigation history are joined and the consequent part association rules, whose antecedent part is equal to the search pattern, are extracted and added to the recommendation set. The ontology instances in the rule consequents are mapped back into Web page addresses. As a result, the number of pages in recommendation set increases. In previous work [18], the number of recommended pages is large and it is possible that some irrelevant pages are recommended as well.

In this research, by using information related to page clustering we identify irrelevant pages and only the pages which are unrelated will be deleted. Before these pages are recommended to a user with respect to page clustering information and how many of these pages are located in each cluster, the maximum clusters are selected and the pages which belong to these clusters are recommended to the user. This process is expressed in Algorithm 1. As an example, consider the active user navigation $Web\ page:IEEE \rightarrow Web\ page:ACM$ where these two pages are concrete Web pages with URLs. In the first step, user navigation history is mapped into ontology instances. As the result of this step, user navigation is turned to $ElectronicEngineering \rightarrow ComputerEngineering$. Assume that the only rule we have is $EngineeringMath \Rightarrow CivilEngineering$. Since $EngineeringMath$ is the parent of $ComputerEngineering$, it can be used for recommendation. At last, $CivilEngineering$ is mapped back into pages. If we assume these pages are $p3, p5, p7, p9, p2, p4$ and we have a page clustering like this:

$$C1: p1, p6, p8, p4 \quad C2: p2, p10, p5 \quad C3: p9, p3, p7$$

Pages $p3, p7, p9$ are located in $C3$, $p2, p5$ are located in $C2$ and $p4$ is located in $C1$. Since the number of clusters is three and based on the proposed algorithm, two

clusters which are maximum, are selected and their pages are recommended to the user. Since two clusters C3 and C2 are maximum, Web pages p9, p3, p7, p2, p5 are added to the final recommendation set. By doing so, the number of recommended pages is reduced according to semantic similarity.

Algorithm 1. construct recommendation set(R, T, K, window_count, Clustered Pages)

Comment: Constructing recommendation set algorithm

Comment: R is set of association rules T is active user navigation path K is number of cluster

```

for each rule  $R = \langle a_1, a_2, \dots, a_n \rangle \Rightarrow \langle c_1, c_2, \dots, c_j \rangle$ 
{
  If ( $a_n = t_m$  and  $a_{n-1} = t_{m-1}$  and  $\dots a_{n-\text{window\_count}} = t_{m-\text{window\_count}}$ )
  then  $L = L \cup R$ ;
}
If ( $K > 2$ )
{
  for each page in  $L = p_1, p_2, \dots, p_l$ 
    Identify cluster  $p_i$ 
  Sort the clusters based the number of pages exist in them
  If ( $3 \leq K \leq 7$ )
    then select 2 clusters are maximum and add pages of them to final recommendation set
  If ( $K == 9$ )
    then select 3 clusters are maximum and add pages of them to final recommendation set
  If ( $K == 11$ )
    then select 4 clusters are maximum and add pages of them to final recommendation set
}

```

4 Evaluation Measures and Experimental Result

4.1 Evaluation Measures

Precision and coverage [15] are the two popular metrics that are used to evaluate and compare the performance of the proposed system.

Each testing session (ts) of testing set is divided into two parts. The first n web pages of session ts is used as the input of the recommendation engine which is denote as `Recommend_List` and the second part is simulated as the future requests (page visits) which are compared with the output of the recommendation system and denote as `Real_List`.

The precision of a transaction is given as the number of web pages correctly predicted divided by the total number of web pages predicted.

$$Precision_t = \frac{|Recommend_List_t \cap Real_List_t|}{|Recommend_list_t|} \quad (2)$$

The coverage of a transaction is given as the number of web pages correctly predicted divided by the total number of web pages visited by the user.

$$Coverage_t = \frac{|Recommend_List_t \cap Real_List_t|}{|Real_List_t|} \quad (3)$$

The precision and coverage were evaluated for all the transactions in the testing dataset and their averages were calculated. The average precision and average coverage values helped to evaluate the system.

In order to get a single evaluation measure, the M-metric, define in [10] is used.

$$M = \frac{2 * coverage * precision}{coverage + precision} \tag{4}$$

4.2 Experimental Results

In this work, the experiments were conducted on the navigation logs that belong to library Web site of Ferdowsi University of Mashhad (<http://c-library.um.ac.ir>). We selected that Web pages related to Information Resources and Open Access parts. After the preprocessing, the log file includes 1300 sessions and contains 5,200 Web page views. The average number of Web pages in a session is 4. As described before, the ontology model of these parts consists of three concepts InfoResource, Science and FreeResource in first level which Science concept has a taxonomy levels.

The dataset is divided into two parts. The first part is the training part, 75% of the dataset, and the other part is reserved for testing, included the remaining 25% of the dataset. By applying the association rule mining algorithm on training data set, association rules are generated. In addition, by using K-means algorithm on mapped pages, clusters are determined. In our experiments, we used K=9 for K-means clustering since the best results are obtained by this value for K.

The first set of experiments shows that using semantic information improves the rule and recommendation quality. As it is seen in the table 1, in part (a) the resulting precision, coverage and matching rate values are higher than the result shown in part (b). As it was expected, using semantic information improves pattern and recommendation quality. In part (b) since semantic information does not contribute to the pattern structure in this experiment, the generated patterns and recommendations have low precision and coverage values.

Table 1. Comparison of the proposed system with recommender system without semantic web, window_count=1, min_confidence=1, k=9

| | <i>Support</i> | <i>Precision</i> | <i>Coverage</i> | <i>M</i> |
|---|--|------------------|-----------------|-----------|
| Results by Applying Semantic and page clustering K=9 (a) | 0.43023 | 0.496032 | 0.161199 | 0.243323 |
| | 0.33720 | 0.603175 | 0.214802 | 0.316789 |
| | 0.23255 | 0.603175 | 0.214802 | 0.316789 |
| | 0.19767 | 0.603175 | 0.214802 | 0.316789 |
| | 0.13953 | 0.515873 | 0.169488 | 0.255148 |
| | Results without Semantic (Just URL) (b) | 0.0697 | 0.0238095 | 0.0010131 |
| 0.0581 | | 0.08333 | 0.028255 | 0.0422019 |
| 0.046511 | | 0.0809524 | 0.0285255 | 0.0418899 |
| 0.03488 | | 0.166667 | 0.037723 | 0.0608728 |
| 0.02325 | | 0.153886 | 0.0734597 | 0.0994469 |

In second set of experiments, we compared our work with approach presented in [18] by applying on our log file. In the first experiments, we have evaluated the effect of change in minimum support on precision values by using `window_count=1` and comparison between our approach and method [18]. In our experiment and as Fig. 4 shows initially, precision value increases when the minimum support increases. This is due to the facts that, when the minimum support increases, weaker association rules are eliminated, and more accurate recommendations can be generated. However, after a breaking point, precision starts to decrease, since the number of association rules and recommendations decrease. Fig. 4 illustrates better result for precision recommendation in our system over [18]. As it was mentioned before, in [18] the number of pages, which are recommended, is large. However, in our method by using page clustering unrelated pages in recommendation sets are identified and those pages which are unrelated are omitted. Consequently, a more precise recommendation set can be presented to users.

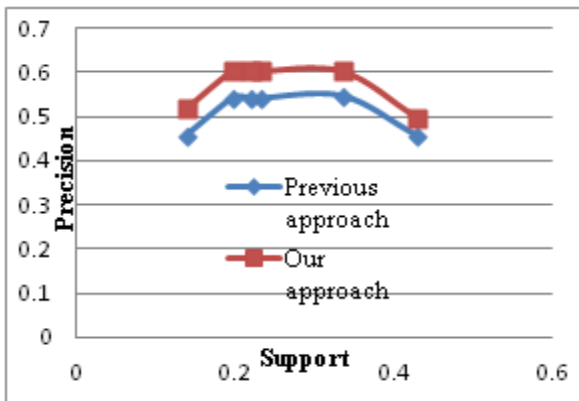


Fig. 4. Precision comparison of the proposed system with the system in [18], `window_count=1`, `min_confidence=1`, `k=9`

Fig. 5 shows coverage comparison between our work and the work in [18]. As Fig. 5 depicts, the coverage of our system is less than the previous system. The reason for this is that in the system [18] when an individual is mapped back to URL's page, the number of pages which are obtained is large, the coverage metric is increased accordingly. However, in our system, using page clustering, we eliminate irrelevant pages and in some parts the number of recommended pages is reduced noticeably consequently, coverage metric is reduced as well. This reducing of extra pages exerts positive effect, and as it's obvious in Fig. 4, the precision metric is also improved.

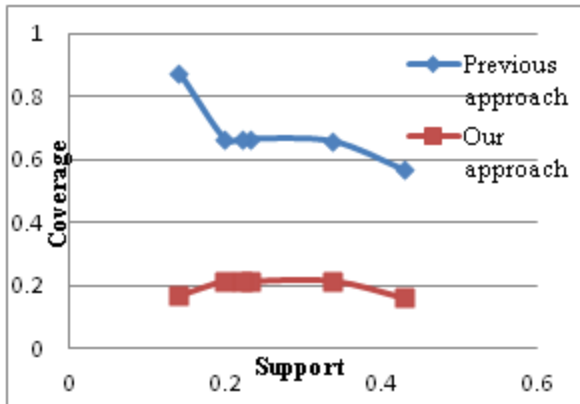


Fig. 5. Coverage comparison of the proposed system with the system in [18], window_count=1, min_confidence=1, k=9

In the last group of experiments, the recommendations are generated under window_count=2 to display the effect of window count in results. Fig. 6 demonstrates that the precision asset value in both method is less than precision values in window_count=1, but as we expected our method for precision metric works better than [18] in window_count=2 as well.

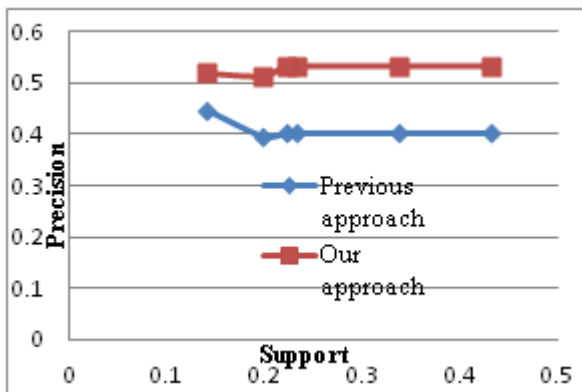


Fig. 6. Precision comparison of the proposed system with system in [18], window_count=2, min_confidence=1, k=9

Fig. 7, shows coverage values as same as precision are decreased too and our approach has lower values than the [18] like window_count=1. In last set of experiments important observation is that the increase in window count has an insignificant and negative effect on the precision and the coverage, hence most recent visit appears to be the most effective one on the recommendation.

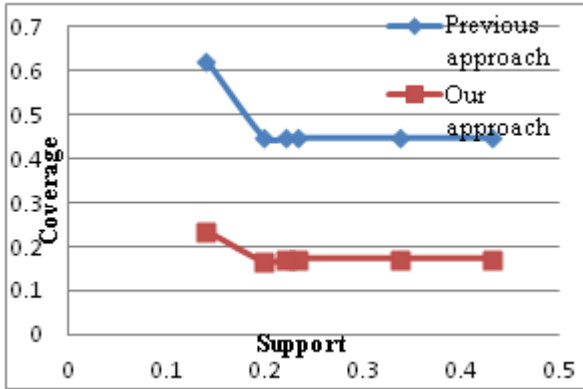


Fig. 7. Coverage comparison of the proposed system with system in [18], window_count=2, min_confidence=1, k=9

5 Conclusion

Extracted pattern in Web usage mining has an important role in many areas, including recommendation, Web personalization, Web construction, Website organization and Web user profiling. In this paper, we have proposed an approach to extract user navigation behavior by using semantic web usage mining. To achieve this aim, we incorporate semantic web into generated patterns. With this combination the created rules contain ontology individuals instead of web pages' URLs.

The success of the system is measured by evaluation of the recommendations. In the presented work, improved pattern quality is used to recommend pages more accurately. Since the number of recommended pages is large, unrelated pages must be determined and omitted. Information from page clustering based on semantic similarity is used for identifying irrelevant pages which are obtained by rules and added to the recommendation set. In previous works, the number of pages which were recommended was large but in this research by using page clustering, that pages which are unrelated will be omitted so, the number of pages is narrowed down. The reducing number of pages affect positively. Therefore, the precision of our proposed system is enhanced. In this way, recommender engine, by using the created rules, selects some pages to recommend but before that irrelevant pages are eliminated according to page clustering. Our experiments suggest that integrating semantic knowledge with web usage mining and in addition to clustering of pages by using of concept similarity in ontology, can indeed be useful in recommender systems.

References

1. Berendt, B., Hotho, A., Stumme, G.: Towards Semantic Web Mining. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 264–278. Springer, Heidelberg (2002)
2. Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. *Knowledge and information systems* 1, 5–32 (1999)

3. Wei, L., Lei, S.: Integrated Recommender Systems Based on Ontology and Usage Mining. *Active Media Technology*, 114–125 (2009)
4. Samizadeh, R., Ghelichkhani, B.: Use of semantic similarity and web usage mining to alleviate the drawbacks of user-based collaborative filtering recommender systems use. *International Journal of Industrial Engineering and Production Research (IJIE)*, English (2010)
5. Etmnani, K., Delui, A.R., Naghibzadeh, M.: Overlapped ontology partitioning based on semantic similarity measures. In: 2010 5th International Symposium on Telecommunications (IST), pp. 1013–1018. IEEE (2010)
6. Zaki, M.J.: SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning* 42, 31–60 (2001)
7. Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on Knowledge and Data Engineering* 16, 1424–1440 (2004)
8. Maedche, A., Zacharias, V.: Clustering Ontology-Based Metadata in the Semantic Web. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) *PKDD 2002. LNCS (LNAI)*, vol. 2431, pp. 383–408. Springer, Heidelberg (2002)
9. Dai, H., Mobasher, B.: Integrating semantic knowledge with web usage mining for personalization. *WebMining: Applications and Techniques.[20082 06211]* (2009), <http://maya.cs.depaul.edu/~mobasher/papers/DM042WM2Book.pdf>
10. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Improving the effectiveness of collaborative filtering on anonymous web usage data. In: *IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization ITWP01 2001* (2001)
11. <http://protege.stanford.edu/>
12. <http://www.philippe-fournier-viger.com/spmf/index.php>
13. Mabroukeh, N.R., Ezeife, C.I.: Using domain ontology for semantic web usage mining and next page prediction. In: *Information and Knowledge Management*, pp. 1677–1680. ACM (2009)
14. Adda, M., Valtchev, P., Missaoui, R., Djeraba, C.: Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing* 45–52 (2007)
15. Nakagawa, M., Mobasher, B.: Impact of site characteristics on recommendation models based on association rules and sequential patterns. In: *IJCAI 2003 Workshop on Intelligent Techniques for Web Personalization* (2003)
16. Stumme, G., Hotho, A., Berendt, B.: Usage Mining for and on the Semantic Web: next generation data mining. In: *NSF Workshop* (2002)
17. Stumme, G., Hotho, A., Berendt, B.: Semantic web mining: State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web* 4, 124–143 (2006)
18. Senkul, P., Salin, S.: Improving pattern quality in web usage mining by using semantic information. *Knowledge and information systems* 30, 527–541 (2012)
19. Yilmaz, H., Senkul, P.: Using Ontology and Sequence Information for Extracting Behavior Patterns from Web Navigation Logs. In: 2010 IEEE International Conference on Data Mining Workshops, pp. 549–556. IEEE (2010)
20. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: 5-th Berkeley Symposium on Mathematical Statistics and Probability, California, USA, p. 14 (1967)

Scalable Analysis for Large Social Networks: The Data-Aware Mean-Field Approach

Julie M. Birkholz¹, Rena Bakhshi², Ravindra Harige²,
Maarten van Steen², and Peter Groenewegen¹

¹ Organization Sciences Department,
Network Institute, VU University Amsterdam, The Netherlands
{j.m.birkholz,p.groenewegen}@vu.nl

² Computer Science Department,
Network Institute, VU University Amsterdam, The Netherlands
rbakhshi@few.vu.nl, ravindra.harige@student.vu.nl, steen@cs.vu.nl

Abstract. Studies on social networks have proved that endogenous and exogenous factors influence dynamics. Two streams of modeling exist on explaining the dynamics of social networks: 1) models predicting links through network properties, and 2) models considering the effects of social attributes. In this interdisciplinary study we work to overcome a number of computational limitations within these current models. We employ a *mean-field model* which allows for the construction of a population-specific model informed from empirical research for predicting links from both network and social properties in large social networks.. The model is tested on a population of conference coauthorship behavior, considering a number of parameters from available Web data. We address how large social networks can be modeled preserving both network and social parameters. We prove that the mean-field model, using a data-aware approach, allows us to overcome computational burdens and thus scalability issues in modeling large social networks in terms of both network and social parameters. Additionally, we confirm that large social networks evolve through both network and social-selection decisions; asserting that the dynamics of networks cannot singly be studied from a single perspective but must consider effects of social parameters.

1 Introduction

Dynamics of social networks are receiving increasing attention in multiple research domains [1-3]. Theoretical developments posit that dynamics are influenced by network [4] and social processes [2]; with recent theory suggesting that the two co-evolve [1]. Methods to explore dynamics of networks traditionally implement evolving graph models, using inferential statistics to assert the likelihoods of the creation, maintenance or dissolution of edges. Two distinct classes of modeling exist: 1) exclusively modeling the effect of network structures on dynamics [5, 6], and 2) modeling effects of social parameters and network effects for small networks (~ 1000 nodes) [2]. Both types of models prove that network processes affect the dynamics of networks. Network models have been able to accurately predict a small percentage of edges, suggesting that dynamics may also be fed by other processes. Social-parameter models have proved social attributes, in combination with network structures, play a role in network dynamics.

Despite this growing knowledge from both model classes, these models have limitations. The main limitation relates to using an evolving graph model which calculates statistical probabilities of individual nodes. This approach generally leads to a super-linear growth in computational load as the network size increases, partly caused by the quadratic growth in the number of links that need to be considered. Both models attempt to overcome this through different means. One is limited to either testing the effect of a few parameters on a large network, or a number of parameters on small networks. Consequently, neither provide a terrain to empirically confirm the effect of both network and social parameters in large social networks.

In order to better understand the dynamics of large social networks, a different computational approach must be taken to overcome the issue of scalability in present models. In this paper we review the two existing model classes used to investigate dynamic social networks, and present a model for overcoming a number of acknowledged limitations. Using a mean-field model approach we are able to overcome scalability issues in previous models through aggregation of individual nodes. Parameters are developed using a data-aware approach which combines empirical research from Social Science and standard inferential statistics to develop a population-specific model for exploring the dynamics of collaboration in science.

We consider the question whether mean-field modeling allows us to describe the behavior of a social system, considering a number of network and social parameters. In this first application of the mean-field model to large social networks, we aim to explain the effect of a set of parameters governing networking patterns of collaboration in Dutch Computer Science (CS). Four parameters are considered in this research: institutional affiliation, scientific age, cosmopolitanism of knowledge production, and visibility of the scientists. We prove that mean-field models expand the empirical testing ground of dynamic network models through increased scalability. This allows us to better understand dynamics of large social networks, covering space that has not been investigated in the past using a mean-field approach.

The paper is set up as follows. In Section 2 we review the state of social network models, specifically highlighting the limitations of present models. In Section 3 we explain the mean-field model, discussing in detail the computational advantages of the model as well as the steps taken to implement a data-aware approach for improved specifications. In Section 4 we test the model on the coauthorship networks of papers from the conference proceedings for Dutch computer scientists, collected from the DBLP data set for 2006 – 2010. Finally, we conclude with the results and implications for scalable, data-aware modeling solutions for explaining dynamics of social networks.

2 Network Models

The evolution of a network is driven by the addition, maintenance, and dissolution of interactions (edges) between nodes over time. Evolving graph models are the most commonly implemented models to explain the dynamics of networks [7–9]. These models assume that nodes are added one-by-one to the network, in discrete time. They infer the probability of a link emerging given a node-transition rate using a Markovian model of simulation. Within this model type two distinct approaches exist investigating social

network dynamics: 1) global network-structure link-prediction models, and 2) social-parameter models integrating social factors into link prediction.

Models with pure network-structure prediction assumptions derive from the vast research on global network structures. Studies on network properties confirm that many real-world networks display small-world properties in which high node clustering is combined with short average internode distances [7,10]. Networks have also been found to behave according to a power-law scale-free phenomenon where a relatively small number of nodes have numerous connections [3,11,12]. Additionally, networks have properties of clustering hierarchies [3], and tendencies of transitivity or “triangles of interaction” describing the manner in which ties between node A and B , and between node B and C facilitate a likely tie between A and C .

From this knowledge on network properties a second generation of studies emerged addressing how a social network can be modeled using properties intrinsic to the network. These global network-structure link-prediction models provide insight into not yet identified or observed linkages [13], as well as to infer not directly observed likely links [14–16]. Within these studies two approaches are taken to predict links: (1) computing node-level measures from greater network structures and, (2) meta-level analyses. In this study we consider only node-level measures (which are comparable to the gap we aim to fill in this research), while still maintaining the network structure.

Several approaches for predicting social network linkages have been proposed, for a complete list see [5]. Despite the extensive research of different measures used to model the network dynamics, all of these models suffer from low fitness, with random link prediction performing just as well as Katz’s model of path collection- predicting links by the sum of collected path lengths per individual [17]. This has led informaticians to explore the effects of additional parameters in understanding network dynamics. A second model type works to address the effect(s) of social parameters on the dynamics of social networks. The justification for these models arose from research on social networks which proved that social selection plays a key role in relation formation [18–20]. Models of this type allow us to question how a social network can be modeled using both network and social properties of nodes. These models also infer edges through evolving graph models but consider state spaces with both network and social parameters. Two model types are commonly used to investigate the inference of these dual parameters: stochastic actor models (SIENA) [2] and exponential random graph models (ERGM) [21].

The key distinction in these models, from the network-only models, is the combination of link prediction based on both local effects, as well as on “social circuits” that capture the influence of more distant ties on behavior [22]. This leads to an exponential growth of the state space due to the consideration of more parameters, requiring extensive computing power in prediction. Given the computational complexity of calculating this for every node these models are not easy to develop in a way that convergence emerges in large networks [22]. Consequently, these classes often limit the size of networks through a theoretical boundary of inferring statistics for a bounded network. This reduces the burden of having to perform computations on potentially very large graphs, but also effectively limits application to small networks (~ 1000 nodes).

In summary, these two model classes provide a testing ground to explore dynamics, but are both not without limitations. Both network and social parameters have scalability problems. As we discuss next, in order to empirically explore the effect of both network and social parameters on large social network dynamics a scalable solution is required.

3 Modeling Framework

We propose a *mean-field approach* for studying social networks; (equally behaving) individual nodes are grouped according to their *states*. This approach is used for an optimized analysis of large-scale systems, allowing for a prediction of the average behavior of the system. The mean-field theory has been applied previously, e.g., to large-scale gossip systems in [23]. Concisely, the state of the system is represented by a distribution, or a vector of fractions of nodes $\delta_s(t)$ in each state s at time unit t . The evolution of the stochastic system is governed by a so-called master equation of the form:

$$\delta(t + 1) = M_{\delta(t)} \cdot \delta(t) \quad (1)$$

$M_{\delta(t)}$ is the matrix, each entry of which is a transition probability from a state s at time t to state s' at time $t + 1$. Thus, we are effectively reducing the global state space, thereby increasing the computational efficiency of the model, and in turn, allowing us to consider more parameters as well as more nodes.

Moreover, we use the notion of *classes*, introduced in [23], to distinguish between equally behaving nodes affiliated to different categories. To this end, the mean-field model predicts average behavior of sets of nodes of each class given a number of social and network parameters. We highlight the modeling steps:

Forming a Model. In order to model the network, first we need to define the system in the form of its parameters. This will form a state of the system. Given the type of network under study, the effects of system parameters are considered using either manual classification or statistical classification (e.g., [24]) to identify the set of significant parameters to form states and classes. For example, some parameter u can be a theoretically informed organizational constraint (e.g. an organization, a background, etc).

Applying Abstraction Refinement. The theory underlying the mean-field model requires also the population of each state to be large enough to be approximated by the law of large numbers. The size of the population in a sampled data set may force one to consider further abstraction for the ranges of the parameters, thereby reducing the size of the system state space. For instance, if chosen parameters for the system are the number of papers per author $p \in \mathbb{N}$ and the number of an author's coauthors $c \in \mathbb{N}$, the number of possible states of the system will simply be a product $\mathbb{N} \times \mathbb{N}$. Some parameters can be restricted in their value ranges without loss of the accuracy of the model itself.

Computing the Model Input. To execute the model, input data is needed on the initial state of the system, as well as on distributions for networking behavior, which will be used for the matrix $M_{\delta(t)}$. The input distributions for the mean-field model include three categories: (1) communication, (2) idle, and (3) collision. Communication describes the interaction between nodes, and idle is a state of no interaction. Collision

is the disappearance or decay of an interaction. The distributions of interaction (links, from a graph-theoretical perspective) are estimated for each class, which determines the nonuniform behavior by different classes for the model. We compute these distributions statistically from the sampled data set.

Estimation of Distributions. The aforementioned transition probability distributions are determined using a discrete-time model to identify the optimal time slicing for the studied data set. Such a time slice corresponds to one time unit in the model. The distribution for probability of transition from one class to another one is also used in the master equation (1) (for a more detailed equation, cf. [23] Fig. 10). The method used for estimation of the probability distributions is a Hidden Markov Model (HMM) [25].

Applying Automated Mean-Field Framework. Armed with the knowledge regarding states, classes and transition rates, obtained from the previous steps, we apply an automated mean-field framework to infer average behavior of the system. We repeat the earlier steps until all parameters are included for a time period covered by the data set. We use the resulting mean-field model to make average link predictions on the system given the parameters under consideration. The model provides a number of advantages over models discussed in Section 2, such as greater flexibility in modeling behavior of nodes through a number of processes. The use of HMMs provides an additional round of probability in node interactions, to compensate for the aggregation. Moreover, such a model allows us to consider both social parameters as well as network structures. Unlike simulation or deployed models, the model is flexible given a theoretical knowledge of the interactions under study. In analyzing the system under question we set the formal specifications which provide detailed processes of specification.

Considerations for Extensions of Social Networks. The challenge in applying the mean-field model to social networks is to derive accurate predictions of the local behavior of the nodes within defined classes. Particularly, for social networks, model abstractions need to be done using a data-aware approach. A data-aware approach implies that both classes and parameters are informed through an intense, robust knowledge of the system under study, as well as the content of edges in the network data. It is a requirement that this is approachable through a theoretically or empirically grounded conceptual scheme on both the system under study and the mechanisms that inform the parameters considered in simulation models. Consequently, not all social networks and or systems can be analyzed using such an approach.

Additionally, we argue for an interdisciplinary approach in development of the model as data needs to be intensely explored to inform parameters by both a data engineer and validated by social scientists or informed experts of the system under study. This implies, unlike other models, that the data-aware approach is essential to determining accurate results, which can be compared in model-fit tests. This results in a model that specifically fits the needs of the system under study, and which can be adapted per population given the basic set of rules for abstraction we describe. In the next section we lay out the general steps for the application of a mean-field model.

4 Application

As discussed in the previous section a set of requirements are necessary for implementing a mean-field model to investigate the effect of social and network factors on network dynamics: network data, parameter data, and knowledge from empirical studies of the system under study. We explain the case studied here and detail the abstraction steps undertaken to model the effect of network and social parameters on network dynamics.

4.1 Network Data

A majority of computational analyses of large social networks implement coauthor or similar co-occurrence networks to examine network dynamics [3]. Coauthorship networks, via publication data, provide a representation of a specific social interaction-successful collaboration, in producing an output- dissemination of knowledge through publication. Moreover, publication data is readily accessible on the Web providing large, reliable, and scalable data sets to model network dynamics.

In addition to the use of coauthorship data to study network dynamics, empirical studies on coauthorship provide a framework to develop measures to consider in the model testing. In science studies, coauthorship is a standard measure for collaboration in science. Collaboration is increasingly common in science; from the near disappearance of single-authored papers to the growth in prevalence of an increasing numbers of coauthors on academic publications [26]. A decade of studies on collaboration in science have proved the effect of different social variables on collaborative behavior of scientists [27, 28]. Recent studies have found that task types and a number of external factors influence collaborative behavior of scientific processes [29]. Both institutional and short geographical distances play a key role in the collaborative behavior of scientists [30, 31]. Given these studies we have a basis at which to both test informed parameters and link findings to knowledge on collaborative tendencies of scientists.

In this paper we explore a system of collaborative behavior of scientists in testing the mean-field model for large social networks. We select one nation and discipline – Dutch computer scientists, to investigate dynamics as to limit known exogenous effects of different knowledge production practices between disciplines and nations. Effectively, we comment only on the average behavior of the system of Dutch CS. The field of CS was chosen for three reasons: the traditions of the field with a diversity of subfields within the discipline; the known tendency for collaboration through coauthorship; the validity and reliability of online sources documenting publications. The Dutch context provides a diversity of cases at which to examine different institutional processes.

A source list of 434 tenured Dutch computer scientists in 2010 was acquired from the Nederlands Onderzoekdatabank, an official body that keeps records on research in the Netherlands. To identify a valid and reliable set of coauthorship data for the Dutch computer scientists a snapshot of DBLP DataBase was queried. (DBLP is one of the most comprehensive bibliographic indices for the field of CS.) Within this set the list of Dutch computer scientists was queried for all publications of scientists from 2006 - 2010 (the year of our list of tenured scientists). This list was manually cleaned to disambiguate names. From this list the name of the publication was queried to identify the

unique author IDs of each author per publication. These unique author IDs were queried to pull full publication lists of each author (Dutch scientists and their coauthors).

Conference proceedings were selected for the case study as conferences in CS require at least one author to physically present work at a conference to be published. Conferences provide a good fit for the assumption of interaction in previous computer models as a potential meeting points for coauthors. Additionally, it provides a number of clear timestamps discerning possible transition periods, with most conferences occurring annually, with regular cycles. Conference proceedings are denoted in this data set by the BibTeX entry `@inproceedings`, allowing us to further query for proceedings-only publications. This resulted in 3639 scientists, and 2757 conference-proceeding publications. Nodes represent individual scientists and links represent shared coauthorship of proceedings. From this data set of individual authors we also collect data on the social parameters.

4.2 Parameters

In this study we aim to include parameters that are informed from previous empirical studies in the field of science studies. Four parameters are considered in the model: scientific age, cosmopolitanism of knowledge production, visibility, and institutional affiliation. For the collection of social parameter data in this study the Web is used, providing a reliable method for collecting meta-data on scientists within publication records [32]. The use of Web data as the source of meta data is integral in this first model development as it reduces the burden of data collection of social variables (compared to traditional social science data of surveys or interviews). This allows us to quickly test the effect of social parameters on behavior with a considerable amount of reliability from merging meta-data from additional online databases.

The parameters – scientific age, cosmopolitanism of knowledge production, and visibility are calculated from within the DBLP data set. Scientific age was selected because tenure and rank are both said to play a role in collaborative behavior of scientists, with scientists of a higher tenure more likely to collaborate than mid-range, tenure-seeking colleagues [33]. We first noted publication per author in the DBLP data set for which we compute per year per author as his or her scientific age. A second parameter, cosmopolitanism, relates to the socio-technical acquired capabilities of scientists suggesting that access to potential coauthors in a field plays a key role in collaboration [27]. This parameter was measured through previous coauthorship experience. The number of coauthors per year per author is computed from the DBLP. The third parameter aims to comment on the visibility of the scientist. The visibility of the scientist is the likely popularity through publication magnitude. These three parameters allow us to consider a number of possible social factors that are not network effects but rather social attributes on the scientists' networking behavior.

One additional parameter was collected for consideration in the model – the institution. Previous studies proved that the institution is statistically significant with respect to how scientists collaborate [29-31]. The institution is identified through a query of two databases. These data are considered static in this model, unlike the previously mentioned data, as we assume minimal change of institution in the five-year period under study. The automatic collection of historical data on institutional affiliation is not currently stored in one database, to our knowledge, thus we assume a five-year period

as a valid period of time to accurately measure inference. A query using Microsoft Academic Search – a database which includes the DBLP data set is used to identify institutions. To locate additional missing data another database, ArnetMiner.org was used. The remaining unidentified institutions were queried manually giving us a total of 1358 identified institutions. In order to disambiguate institutional names, to have a reliable and valid set of data, this list was queried in geocoding Web service Yahoo! PlaceFinder [34]. This query provides a proximity measure for each institution and a uniform institutional affiliation based on common GPS coordinates.

These four parameters provide a setting to explore the application of the mean-field model in large social networks. The occupancy measure at time $\delta(t)$ in our model is the fraction of people in state (p, c, h, u) , where p is a number of publications, c is a number of coauthors, h is scientific age, and u is affiliation. We test the following social science hypothesis: institutions effect the patterns of collaborative behavior (by behavior we mean average number of coauthors, and average number of papers). In addition to these social parameters we also include the network parameter of transitivity. As discussed in section 1, social networks have tendencies of transitivity [3,7]. We consider the social parameters in predicting the triadic interactions between nodes.

4.3 Classes Abstraction

In principle, any of our parameters could be considered a class. When studying a social system, however, we need to consider known social and organizational constraints. In order to define a class we investigate the four possible parameters under consideration in this model. We first consider known effects.

Our system is already bounded by the selection of one national science structure and one scientific discipline. The effect of the institution provides a valid and logical boundary at which to explore aggregation. Additionally, we know that geographical location also plays a key role in collaboration, which we aim to consider in the abstraction. Consequently, we employ institutions as classes in our mean-field model, and as one of the parameters u contributing to a state (p, c, h, u) of a collaboration network. Due to limitation of the data-mining techniques to automatically extract full history of scientific employment, we assume that a scientist has one affiliation during the four year period.

The data set for our model consist of 3639 Dutch authors with 749 different institutions. However, the theory underlying our mean-field model requires that the population of each class should be large enough to be approximated by the law of large numbers. To this end, we applied an abstraction on classes (institutions) based on statistical metrics for the given distribution D of computer scientists among institutions.

Since both our data set and results are focused on the system of Dutch computer scientists, we distinguish (1) institutions in the Netherlands, and (2) institutions in other countries. For each of these categories we estimate a statistical threshold of the significance of the institution. This threshold depends on the dispersion of the distribution D' of scientists sampled for each of the categories of institutions. If values are highly dispersed, then we set the threshold to be the average number of affiliated scientists.

To measure the statistical dispersion for the scientists' distribution S , we compute a *sample covariance*, which is the average distance to the mean value between any two values in the distribution S . To allow for some dispersion, we compare the arithmetic

mean for S and its sample covariance: if the sample covariance for a subset $S \in D$ is higher than the mean, then the values of the sampled D' are highly dispersed.

In addition to estimation of the significance threshold, this simple test is applied in two steps: (1) for the continental abstraction, and (2) the country-wide abstraction. In case 1, we sample data for all universities per continent (using the UN list of countries per continent and GPS coordinates). In the case of high dispersion in the number of scientists in institutions in one continent, we proceed to test the dispersion of the number of scientists affiliated with institutions in one country. We merge only those institutions that have a number of scientists below the mean of the entire distribution D . The histogram in Fig. 1 shows the number of scientists in each class, before and after the classes abstraction. The number of classes has been reduced from an initial 749 to 157, effectively reducing also the state-space size.

4.4 Other Parameters Abstraction

Scientific Age. The scientific age h is based on the first publication date of an author according to DBLP. The earliest possible publications in DBLP date back to 1971, which inevitably leads to an increase by a factor 40 of the state-space size of our model. Considering our sampled data set with only 3639 scientists, the distribution of the population in such a state space is very sparse. Thus, we identify five main groups of scientific age, categorizing age into ten-year periods as to generalize about generations of scientists: 70, 80, 90, 2000, 2010. In general, scientific careers require substantial investments to establish tenure. These positional differences, whether it being established tenure, or a starting PhD, all influence the manner in which scientists undertake collaboration [27, 33]. Our abstraction granularity is fine enough to strongly indicate the scientific position of researchers, e.g., senior staff, junior staff.

Visibility. The visibility of the scientists is measured by the annual number of conference publications. We choose only conference publications, as a potential interaction point, assuming that scientists encounter future collaborators during conferences. Without loss of generality, we limit the highest number of conference publications per year to 12 assuming it takes on average one month of preparation per publication. Those scientists that publish 12 and more papers per year we distinguish as fast publishers with a parameter value of 12.

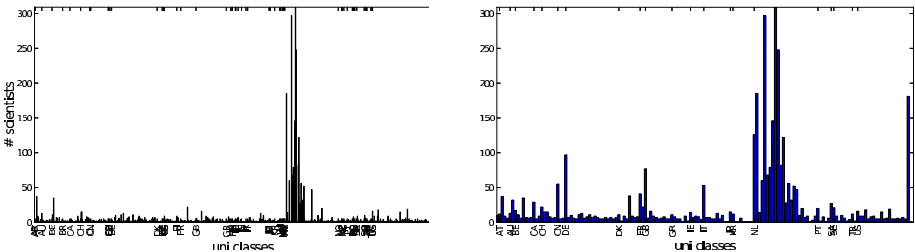


Fig. 1. The distribution of scientists among institutions before (left) and after (right) the abstraction

Cosmopolitanism. The cosmopolitanism of the science is measured by number of coauthors, indicating how well connected a scientist is. We studied the distribution of the number of coauthors on our sampled data set. We observed that there are few publications with a large (more than 12) number of coauthors on a single paper. A high number of coauthors on a paper generally indicates a participation in a large research project. This results in an unnecessary large state-space size of the model, given the sampled authors in this sample. To tackle this, we distinguish five categories of coauthor count per paper: “non cooperative” (0) for the papers with one author, “regular” (1) for the papers with up to 3 coauthors, “high” (2) with up to 6 coauthors on the paper, “team” (3) with up to 10 coauthors, and a “large project” (4) for papers with more than 10 coauthors. Since we consider the unique coauthors of a scientist as possible network contacts within one year, we take the annual number of coauthors relative to the number of the publications per year per person.

4.5 Transitions and Distributions

There are three categories of distributions needed to derive from our data set for our mean-field model: (1) communication κ , (2) idle η , and (3) collision ϕ . *Communication* is defined as collaboration via shared coauthorship between two scientists resulting in a conference paper. Both *idle* and *collision* states signify the decay of communication; in fact, for our application, these probability distributions are both an identity function. Moreover, in terms of the model, selection of the collaboration partner is governed by the distribution function *contact*, which specifies the collaboration network topology.

Computing Transition Probabilities. We first measure from the collected data the evolution of collaboration between scientists (nodes) for each year 2006–2010. That is, we compute the state vector $\delta(t)$, entries of which are the fractions of nodes in every possible state of the system at time t . This state vector $\delta(t)$ is used in the initial configuration for the model: we sum up all fraction of nodes with scientific age h from class u , $\delta_{(p,c,h,u)}(t)$ for all possible p and c and set the result as $\delta_{(0,0,h,u)}(0)$ at the beginning of each year t . In the model, we split the time frame onto a week τ , for finer granularity, with 52 weeks in each year.

Consider states $A = (p_a, c_a, h_a, u_a)$ and $B = (p_b, c_b, h_b, u_b)$. For each pair of classes u_a and u_b , we compute the probability $\text{contact}(u_a, u_b)$ that a node from u_a contacts any node in u_b in year t as follows. Each paper i with c_i -authors by a node from u_a and a node from u_b gives the probability $P_i(c_i, u_a, u_b) = \frac{1}{m(u_a) \cdot c_i}$ that the node from class u_a contacts a node from u_b . Here, $m(u_a)$ is the number of nodes in class u_a . Since we have to take into account that papers jointly written by nodes from u_a and u_b may have other coauthors, divisor c distributes the share of contribution to each coauthor. Then, $\text{contact}(u_a, u_b)(t)$ is obtained as follows: $\text{contact}(u_a, u_b)(t) = \sum_{i(u_a)} \sum_{i(u_b)} P_i(c_i, u_a, u_b)$, where $i(u_a)$ and $i(u_b)$ means “for each author of paper i from class u_a ” (u_b , respectively).

The computation of the collaboration distribution $\kappa_{(A,B)}(t)$ is as follows. For each paper penned by authors in states A and B (within a one-year time frame), we observe all possible state transitions (i.e. before and after collaboration). The result is an expression of the form:

$$\kappa_{(A,B)}(t) = \{(p_1, (A, B), (A_1, B_1)), \dots, (p_n, (A, B), (A_n, B_n))\}$$

where p_i is the probability that the nodes in state A at time t make a transition to state A_i at time $t+1$ (and, those in state B move to state B_i , respectively). All these distributions are normalized to a weekly timescale.

Estimating Distributions. These rates may vary from year to year thereby requiring an average to be determined for every of these distributions to ensure accuracy in the model. To that end, we obtained probabilities, as described earlier, for the years 2006–2008, and use an HMM approach to sample the underlying distribution. Our goal is to approximate the set of pairs that have positive probability of collaborating. Our mean-field model takes these sampled distributions as its input.

5 Results

The mean-field model allows us to predict average behavior. The analytical results to the statistical results for the years 2009 and 2010 are compared to the ones produced by the mean-field model. Institutions are labeled and sorted in lexicographical order; this list is enumerated and corresponds to the number on the x -axis (similar to Fig. 1). Classes 98-116 correspond to Dutch institutions. As we can see from Fig. 2a the mean-field results for the larger institutions corresponds with the statistics from the data set for 2010. Our data set does not list all papers of the coauthors of coauthors, but we divide by all people in the class; so statistics produced are lower than actual.

Institutional Factor. The results produced by the alternative mean-field model with uniform distribution *contact* for collaborations between different institutions show that the sample distribution is non uniform. This *contact* distribution produces the equal probability of collaboration between any two scientists in the whole network, irrespective their affiliations, and thus forms a baseline for comparison to see whether affiliations are statistically significant. The comparison is shown in Fig. 2b. As we can see, the uniform *contact* distribution predicts higher output for foreign institutions but lower for Dutch institutions, since the output is then uniformly “redistributed”.

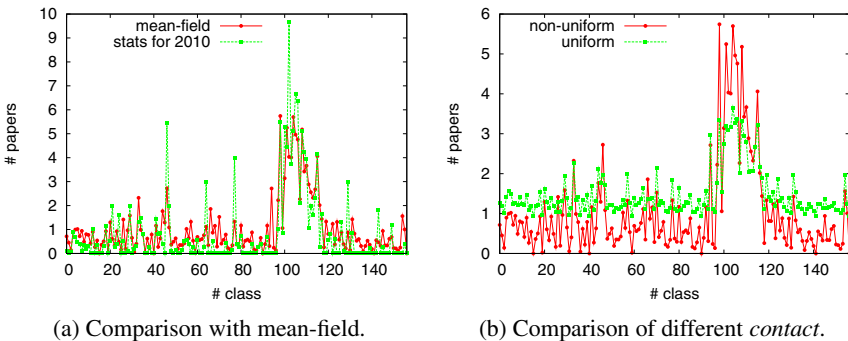


Fig. 2. Average output for different classes

| Sci. age | avg # pubs. | $u_a \leftrightarrow u_b, u_b \leftrightarrow u_c$ | avg. $u_a \leftrightarrow u_c$ |
|----------|-------------|--|--------------------------------|
| 2010s | 1.8 | ≥ 0.0 | 1.0 |
| 2000s | 1.61 | ≥ 0.2 | 1.13 |
| 1990s | 1.76 | ≥ 0.4 | 1.15 |
| 1980s | 1.95 | ≥ 0.6 | 1.20 |
| 1970s | 2.3 | ≥ 0.8 | 1.27 |
| | | ≥ 1.0 | 1.32 |

(a) Average output for different scientific age. (b) Triad relations.

Fig. 3. Results for the age impact and triad relations for Dutch institutions

Impact of Scientific Age. Fig. 3a shows the average number of papers for different scientific age. The results from only Dutch institutions were averaged. The mean-model shows that a principle of preferential attachment [3] is occurring in the network based on age, with higher tenured scientists acquiring more collaborators and papers. The average output per scientific age per institution, was also computed; see results in [35], which displayed differing tendencies in collaboration patterns.

Link Prediction. In accessing the manner in which links are made through transitivity: if class *A* has a paper in common with *B*, and class *B* with *C*, then *A* has stronger connectivity with *C*. Within this system we consider the institution parameter, allowing us to reflect on the initial hypothesis – an institution plays a role in the collaborative patterns of scientists. The connectivity factor based on the distribution *contact*, which in turn, depends on the probability $P_i(c_i, u_a, u_b)$, the number of coauthors from a certain institution implicitly contributes to strength of the connectivity between institutions. Fig. 3b shows the generalized triad relations of Dutch institutions; considering a scientific age in *contact* produces results in [35].

6 Discussion and Conclusion

In investigating the system of Dutch computer scientists’ collaborative behavior through the mean-field model we observed systematic networking behavior associated with a number of social parameters, which aid in describing the networking dynamics of scientists. The past collaborative partners of one’s institution plays a key role in how future collaborations unfold. With every conference proceeding with another institution the chance of collaborating with the institution increases. Age also matters; the age of the scientists plays a role in the visibility of a scientist (number of publications) within the system. The cosmopolitanism of the scientists (number of co-authors) also contributes to the likelihood of future interaction. Consequently the mean-field model allows us to describe the Dutch CS system of conference paper collaboration to be governed by a number of social variables, where ties can be predicted given previous relationships among common institutions, reinforcing clustering tendencies in these networks.

In this first application of the mean-field model in predicting both social and network parameters for large social networks, we also recognize a number of shortcomings. The first is the sensitivity of the data-aware approach and thus the empirically informed

aggregations of nodes into clusters from such an approach. Future work should aim to consider additional social parameters, such as performance, gender, discipline, length of time known in understanding the system. To improve the precise description of states the notion of idle and collisions in the model should be improved for social networks. Additionally, we acknowledge that this explorative study of the mean-field model did not address both the potential for shift classes reflecting the fluidity of actual organization constraints in social life, as well as model checking. These limitations are related to the current state of computing techniques, in first data-mining techniques which does not currently allow us to collect such refined information on social beings, and secondly the lack of methods to appropriate accurate model checking.

The incorporation of the modeling knowledge with population specific dynamics we are able to identify the conditions under which links emerge given a set of both network and social parameters through the mean-field model. This allows us to provide informed predictions to comment on the mechanism(s) under which specific patterns of behavior emerge in large social networks. Mean-field models provide a meta-scopic method, which overcomes limitations of the network only and social parameter models. Meta-scopic models of this sort allow us to incorporate both the micro (considered in evolving graph models) and the mega networking processes to infer links through a data-aware approach. Additionally, it provides an empirical terrain at which to explore the effects of both network and social parameters on large social networks.

Acknowledgements. We thank Paul T. Groth for the initial DBLP data set, and Jörg and Stefan Endrullis for their support in the "refitting" of the automated mean-field framework for the social domain.

References

1. Ahuja, G., Soda, G., Zaheer, A.: The genesis and dynamics of organizational networks. *Organization Science* 23, 434–448 (2012)
2. Snijders, T., van de Bunt, G., Steglich, C.: Introduction to actor-based models for network dynamics. *Soc. Networks* 32, 44–60 (2010)
3. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* (74), 47–97 (2002)
4. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* 286, 509–512 (1999)
5. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. *J. ASIST* 58(7), 1019–1031 (2007)
6. Moore, C., Ghoshal, G., Newman, M.E.J.: Exact solutions for models of evolving networks with addition and deletion of nodes. *Phys. Rev. E* 74, 036121 (2006)
7. Newman, M.: Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci* 101, 5200–5205 (2004)
8. Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications* 311(3–4), 590–614 (2002)
9. Grossman, J.W.: Patterns of collaboration in mathematical research. *Notices of the AMS* 52(1), 35–41 (2005)
10. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* 393(49), 440–442 (1998)

11. de Solla Price, D.: Introduction to the special issue on network dynamics. *Science* 149(3683), 510–515 (1965)
12. Akkermans, H.: Web dynamics as a random walk: How and why power laws occur. In: *Proc. Conf. of Web Science (WebSci)*. ACM (to appear, 2012)
13. Krebs, V.: Mapping networks of terrorist cells. *Connections* 24(2), 43–52 (2002)
14. Goldberg, D., Roth, F.: Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci.*, 4372–4376 (2003)
15. Popescul, A., Ungar, L.: Statistical relational learning for link prediction. In: *Proc. Conf. on Artificial Intelligence*, pp. 81–90. ACM (2003)
16. Taskar, B., Wong, M.F., Abbeel, P., Koller, D.: Link prediction in relational data. In: *Proc. of Neural Information Processing Systems*, pp. 659–666. MIT Press (2003)
17. Katz, L.: A new status index derived from sociometric analysis. *Psychometrika* 18(1), 39–43 (1953)
18. Granovetter, M.: The strength of weak ties. *American Sociological Review* 78, 1360–1380 (1973)
19. Krackhardt, D.: The strength of strong ties: the importance of philos in organizations. *Netw. and Organiz.: Structure, Form, and Action*, 216–239 (1992)
20. Ennett, S., Bauman, K.: The contribution of influence and selection to adolescent peer group homogeneity, the case of adolescent cigarette smoking. *J. of Personality and Social Psychology* 67, 653–663 (1994)
21. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p^*) models for social networks. *Soc. Networks* 29(2), 173–191 (2007)
22. Robins, G.: Exponential random graph models for social networks. In: *Handbook of Social Network Analysis*. Sage (2011)
23. Bakhshi, R., Endrullis, J., Endrullis, S., Fokkink, W., Haverkort, B.: Automating the mean-field method for large dynamic gossip networks. In: *Proc. of QEST*, pp. 241–250. IEEE Computer Society (2010)
24. Bishop, C.M.: *Neural Networks for Pattern Recognition*. OUP (1995)
25. Stratonovich, R.: Conditional markov processes. *Theory of Probability and its Applications* 5, 156–178 (1960)
26. Grenne, M.: The demise of the lone author. *Nature* 450(1165) (2007)
27. Bozeman, B., Crley, E.: Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy* 33(4), 599–616 (2004)
28. Stokols, D., Misra, S., Moser, R., Hall, K., Taylor, B.: The ecology of team science: understanding contextual influences on transdisciplinary collaboration. *American Journal Preventive Med.* 35(2S), S96–S115 (2008)
29. Börner, K., Contractor, N., Falk-Krzesinski, H.J., Fiore, S.M., Hall, K.L., Keyton, J., Spring, B., Stokols, D., Trochim, W., Uzzi, B.: Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Sci. Transl. Med.* 2(49), 49cm24 (2010)
30. Rodriguez, M., Pepe, A.: On the relationship between the structural and socioacademic communities of a coauthorship network. *J. Informetrics* 2(3), 195–201 (2009)
31. Jones, B.F., Wuchty, S., Uzzi, B.: Multi-university research teams: shifting impact, geography, and stratification in science. *Science* (322), 1259–1262 (2008)
32. Mika, P., Elfring, T., Groenewegen, P.L.M.: Application of semantic technology for social network analysis in the sciences. *Scientometrics* 68(1), 3–27 (2006)
33. de B. Beaver, D., Rosen, R.: Studies in scientific collaboration. Part III. Professionalization and natural history of modern scientific coauthorship. *Scientometrics* 1(3), 231–245 (1979)
34. Yahoo! PlaceFinder, <http://developer.yahoo.com/geo/placefinder/>
35. Birkholz, J.M., Bakhshi, R., Harige, R., van Steen, M., Groenewegen, P.: Scalable analysis for large social networks: the data-aware mean-field approach. Technical Report arXiv:1209.6615, CoRR (2012)

A Survey of Recommender Systems in Twitter

Su Mon Kywe, Ee-Peng Lim, and Feida Zhu

Singapore Management University, Singapore
{monkywe.su.2011, eplim, fdzhu}@smu.edu.sg

Abstract. Twitter is a social information network where short messages or tweets are shared among a large number of users through a very simple messaging mechanism. With a population of more than 100M users generating more than 300M tweets each day, Twitter users can be easily overwhelmed by the massive amount of information available and the huge number of people they can interact with. To overcome the above information overload problem, recommender systems can be introduced to help users make the appropriate selection. Researchers have begun to study recommendation problems in Twitter but their works usually address individual recommendation tasks. There is so far no comprehensive survey for the realm of recommendation in Twitter to categorize the existing works as well as to identify areas that need to be further studied. The paper therefore aims to fill this gap by introducing a taxonomy of recommendation tasks in Twitter, and to use the taxonomy to describe the relevant works in recent years. The paper further presents the datasets and techniques used in these works. Finally, it proposes a few research directions for recommendation tasks in Twitter.

Keywords: Twitter, recommender systems, personalization.

1 Introduction

1.1 Motivation

Twitter is an online social information network launched in July 2006. By 2012, the number of Twitter users has grown to over 140 million [1]. Unlike many other online social networks, the user-user relationships in Twitter network can be social or informational, or both. This is because users not only follow other users for maintaining social links, but also for gaining access to interesting information generated by others [13, 15]. For example, Twitter has been often used to share information and sentiments about live events including the 2011 Egypt's revolution [5].

As Twitter users generate more than 300M tweets each day, these users are also overwhelmed by the massive amount of information available and the huge number of people they can interact with. To overcome the above information overload problem, recommender systems can be introduced to help users make

¹ <http://en.wikipedia.org/wiki/Twitter>

the appropriate selection. While some of these are already deployed so far, most of them are still being studied as research projects in universities and industry labs. These research projects usually address individual recommendation tasks. There is currently no comprehensive survey for the realm of recommendation in Twitter to categorize the existing works as well as to identify areas that need to be further studied. The paper therefore aims to fill this gap by introducing a taxonomy of recommendation tasks in Twitter, and to use the taxonomy to describe the relevant works in recent years.

Our taxonomy is designed considering the unique functions users can perform in Twitter. Before we show the taxonomy, we first review these functions as follows.

- **Tweet** - This refers to posting a message of up to 140 characters, known as tweets. The content of tweets may vary from users' daily activities to news [13]. Some messages may also include URLs to web pages or hashtags to relate tweets of similar topics together. Each hashtag is a keyword prefixed by a # symbol. For example, #Egypt and #Jan25 have been used to group tweets related to Egypt's revolution in January 2011.
- **Retweet** - This refers to forwarding a tweet from another user to the followers. Such re-sharing of tweets is a prevailing mechanism in Twitter to diffuse information.
- **Follow** - This refers to linking to another user and receiving the linked user's tweets after that. The user creating such a link is called the *follower* and the linked user is known as the *followee*.
- **Mention** - One may mention one or more users in a tweet by including in the tweet the mentioned user name(s) prefixed by the @ sign. The mentioned user(s) will subsequently receive the tweet. This is a means for users to gain attention from the other users so as to start new conversations.

1.2 A Taxonomy of Recommendation Tasks for Twitter

Our taxonomy represents the information required for the above user functions. We represent the information involved in different functions by different tuples as shown in the Table II. For example, a tweet action performed can be represented by $tweet_i = \langle u_i, t_i, Url_i, Tag_i \rangle$ where u_i , $text_i$, Url_i , Tag_i denote the user who tweets, tweet's text, the set of URLs and set of hashtags that appear in the tweet respectively.

For each of the above functions, one can define one or more recommendation tasks to aid users in deciding the missing field(s) in the corresponding tuples. For example, a user u_0 trying to perform a tweet function may have written a piece of text, e.g., "SocInfo2012 has announced the keynote speakers" but does not know what hashtag(s) to use. In this case, we have a tuple $\langle u_0, \text{"SocInfo2012 has announced the keynote speakers"}, \{\}, Tag? \rangle$ with the hashtag information to be suggested as represented by the $Tag?$ variable. This tuple with a variable therefore corresponds to a recommendation task that suggests hashtags for a given piece of text written by a given user.

Table 1. Tuple Representations

| Function | Tuple | Function | Tuple |
|-------------|---|-------------|--|
| $tweet_i$ | $\langle u_i, text_i, Url_i, Tag_i \rangle$ u_i : user who tweets $text_i$: tweet’s text Url_i : set of URLs in the tweet Tag_i : set of tags in the tweet | $mention_i$ | $\langle u_i, U_i, text_i, Url_i, Tag_i \rangle$ u_i : user who mentions others U_i : users who are mentioned $text_i$: the tweet’s text Url_i : set of URLs in the tweet Tag_i : set of tags in the tweet |
| $retweet_i$ | $\langle u_i, u_j, t_j \rangle$ u_i : user who retweets u_j : user whose tweet is retweeted t_j : the tweet that is retweeted (URLs and tags may already exist in t_j) | $follow_i$ | $\langle u_i, u_j \rangle$ u_i : user who follows u_j : user who is followed |

One can work out a variety of recommendation tasks by assuming that some field(s) in some tuples are not known. In Figure 1, we show our proposed taxonomy of recommendation tasks in Twitter and each task is accompanied by its corresponding tuple representation and recommendation statement. In the remaining parts of this paper, we will survey some of the recommendation tasks which have been studied or are being studied.

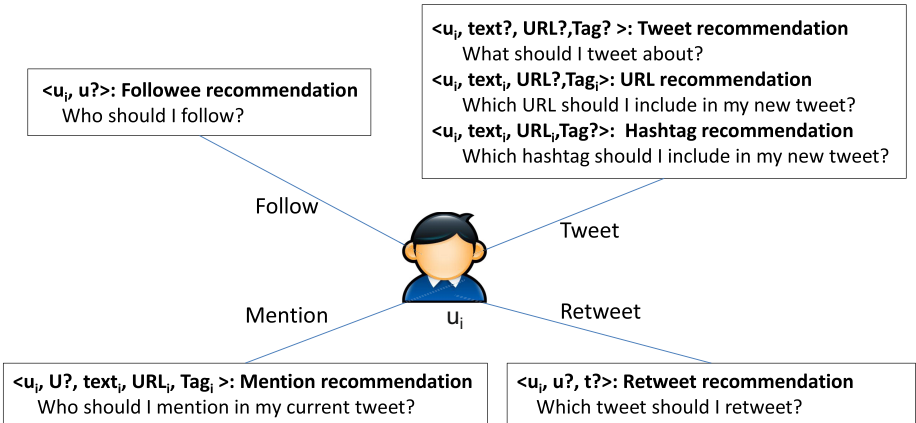


Fig. 1. Taxonomy of Recommendation Tasks in Twitter

1.3 Paper Outline

The remainder of the paper is structured as follows. Section 2 provides a summary of the traditional recommendation methods. Followees, followers, hashtags, tweets, retweets and news recommendation tasks and their methods are summarized in Section 3, 4, 5, 6, 7, and 8 respectively. We finally conclude the paper in Section 9.

2 Traditional Recommender Systems

Recommender systems perform information filtering by suggesting to a user some new items (e.g., songs, books, or movies) to purchase or some new users for building friendships [21]. There are two types of recommender systems – *personalized* and *non-personalized*. The personalized recommender systems consider the preferences of users to be recommended. The non-personalized recommender systems however do not make use of user preferences. An example of non-personalized recommendation method is to return top ten songs of the current month. Most recommendation methods to be surveyed in this paper are personalized. Personalized recommender systems utilize characteristics of items, profiles of users and the interactions or transactions between users and items to predict the users' future item adoptions. Collaborative filtering and content-based approaches are often used in personalized recommendation.

2.1 Collaborative Filtering and Content-Based Recommendation

The underlying assumption of the *user-to-user based collaborative filtering approach* is that if a person X has the same opinion as a person Y on an issue A , X is more likely to adopt Y 's opinion on a different issue B than a randomly chosen person. The recommender system finds people with similar tastes or preferences, according to their past ratings or implicit interactions. Then, the system predicts the preference of a user on an unrated item using the preferences of similar users [23].

Another personalized recommendation approach is *item-to-item collaborative filtering* which is used by Amazon.com's recommender system. Items A and B are highly similar if a relatively large portion of the users who purchase item A also buy item B . Then, the preference of a user over an unrated item B is predicted based on the user's rated item A .

Content-based recommender system finds similar items by comparing their features and characteristics. Then, the recommendation of an item is made to the user who likes or purchases similar items before. In other words, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

2.2 Other Approaches in Social Media

Social recommender systems recommend items based on the preferences of a user's friends or other social media information, such as tags and comments. The recommended items might not necessarily be components of social networks. For example, in the case of Twitter, one can recommend news articles making use of the attention the articles received from Twitter users. Hence, such recommendation may be targeted for users outside of Twitter.

Technique wise, the existing recommendation methods used in social media have to adapt to the unique features in Twitter. For example, the friendship

recommendation methods that work well at social networking sites such as Facebook may not work well in Twitter’s follow link recommendation as the latter is asymmetric (i.e. users do not necessarily follow back those who follow them).

3 Followee Recommendation

In Twitter, users are interested in finding not only their close friends but also new relevant contacts not yet known to them. A user may follow other users whom he or she does not know offline but who share interesting trending topics. These users can be treated as information sources for the user. Depending on the target user needs, different followee recommendation algorithms can be used. For instance, one may use number of common friends to recommend known friends, or use user profile similarity measures to recommend users with similar interests, or popularity scores to find good information sources.

Twitter has a “Who to follow” feature at Twitter home page, the user profile pages, and Connect and Discover pages². It recommends followees who are similar to the existing followees of the target user, and followees of those followees. When the target user visits another user’s profile page, users who are similar to the visited user profile will be recommended. The exact recommendation algorithm behind this feature is however unknown. The recommendation algorithm also includes advertiser accounts which are labeled as “promoted” accounts.

3.1 Topology-Based Methods

Armentano et al. proposed three very similar topology-based approaches for followee recommendation [1–3]. They [3] use both collaborative filtering and content-based recommendation.

Collaborative filtering approach is considered a topology-based method since similar users are found based on follow graph. The authors assume that the target user is similar with the followers of his or her followees. Hence, candidate followees are ranked according to the number of common followees with the target user, page rank and the number of mentions. The top ranked candidates are then recommended as potential followees. The number of common followees represents the similarity between the two users’ preferences. Both page rank and number of mentions determines the popular and reliable information source.

In the content-based approach proposed by the same research group, user interest is represented by the tweet content of his or her followees. The users whose followees’ tweets are similar to the followees’ tweets of the target user will therefore be recommended. The implicit assumption is that a target user is likely to follow those who are similar. This is consistent with the homophily effect where individuals have tendency to bond with similar people [18].

The above two approaches are somewhat different from typical collaborative filtering which recommends followees of similar users, instead of similar users.

² <https://support.twitter.com/>

For instance, a user X follows those who tweet about A , while a user Y follows those who tweet about A and those who tweet about B . In a typical collaborative filtering approach, if user X and user Y are similar, user X should be recommended with those who tweet about B as potential followees. Nonetheless, user Y is recommended as a potential followee in the above approaches.

3.2 Weighted Content-Based Methods

The paper proposed by Garcia [7] identifies features that might be useful for recommending followees. Although five features, namely popularity, activity, location, friends in common and content of the tweets, are predicted to be relevant for recommendation, only popularity and activity have been evaluated. The intuition of the paper is that if a target user has many popular and active followees, other popular and active followees should be recommended to the user. If the target user has only popular followees, only popular followees should be recommended. A similar approach can be applied for target users with active followees.

Popularity is measured by the follower and followee count ratio, while activity is defined by the number of tweets a user has posted since he registered on Twitter. A user is regarded as popular or active when the score is greater than certain threshold. Then, the preference score of a user towards popularity is defined by the fraction of followees who are popular. The preference score of a user towards activity is defined by the fraction of followees who are active. When the preference score of the target user towards popularity or activity is greater than certain threshold, popular or active followees will be recommended.

Moreover, the paper observes that the two features together perform better in prediction than alone. It gives an insight that if more features are considered, the recommendation accuracy can be further improved.

3.3 Structural Methods

A structural approach to contact recommendations in Twitter is introduced by Golder et al. [8]. This work introduces ‘reciprocity’, ‘shared interests’, ‘shared audience’ and ‘filtered people’ methods for recommending followees. The reciprocity method assumes that a user will follow back his or her followers, just to return the attention.

Shared interests and shared audience methods are based on the assumption of homophily, which states that people form ties with like-minded or similar others. A set of users is considered similar or shares the same interest if they are following the same people. Similarly, users who share the same audience or followers are considered similar. A user is then recommended to follow his similar users.

Filtered people of a user are the users whose tweets are retweeted by the followees of this user. The paper states that a user may be interested to follow those filtered people who are the followees of the user’s followees because they may also share the same interest.

3.4 Twittomender

In [9–11], Hannon et al. presents a Twittomender system that recommends followees using both content-based and collaborative-based approaches.

In their content-based approach, users are represented by: (i) their own tweets, (ii) their followees' tweets, (iii) their followers' tweets, or (iv) combination of all of them. In case (i) where a user is represented by his own tweets, users with similar tweets are recommended to the targeted user. In case (ii), a target user is recommended with a list of users whose followees' tweets are similar to those of this user's followees. Cases (iii) and (iv) are treated similarly. In all these cases, each user is represented by TF-IDF weighting scheme [22].

In the collaborative-based approach, the users are represented by IDs of their followees, IDs of their followers or combination of them. IDs are treated as keywords and each user is represented by a set of his follower/followee IDs. Then, TF-IDF weighting scheme is used to find users with similar follower/followee IDs. For example, in the first case where the users are represented by IDs of their followees, a followee is more likely to represent the user's interest if it is not followed by a lot of other people (IDF score). When two users have such common followee, they are more likely to be similar than if they share a common followee who is followed by many users.

Experiments have shown that the above collaborative methods is more precise than the content-based methods. The three most precise methods are, the combination of all the individual methods, followed by the method where users are represented by their followees' IDs, and the method where users are represented by both of their followees' IDs and followers' IDs.

3.5 Recommendation Based on Followers and Lists

In the paper of Krutkam et al. [14], followee recommendations are made based on the number of followers that the user has, the number of lists or groups that the user is listed in and the number of news related group the user is in. The methods are not personalized. In other words, they suggest the most popular users based on the above methods, without considering the individual user's preferences. According to the surveyed results, recommendation based on the number of followers significantly outperforms recommendation based on the number of lists the user is in.

4 Follower Recommendation

While the needs of the general users are targeted by the followee recommender systems, marketers and politicians are interested in finding out new followers who can spread their tweets by retweeting. The following paper emphasizes on identifying followers who can efficiently share information, recommendations and news (such as conference announcements and events) with like-minded users in a community.

4.1 Tadvise

Nasirifard et al. introduced Tadvise to recommend new followers based on their hashtags. The purpose of Tadvise is to help users know their followers better [19]. A set of hashtags is associated with each user's profile as the hashtags appear in the user's tweets. The weight of each hashtag in the user's profile is defined by the total PageRank of the users who mention the profile's owner with the corresponding hashtag. The intuition behind this is that a hashtag is highly relevant to a user if it is frequently used in the user's incoming tweets by highly authoritative users.

Tadvise then recommends well-connected topic-sensitive users as followers. These users may serve as hubs for broadcasting a tweet to a larger relevant audience. The candidate followers are ranked by their hub scores which represent the number of interested users who could potentially receive tweets from the former.

Given a user and a tweet with at least one hashtag, Tadvise determines whether the tweet will likely diffuse from the user. Firstly, Tadvise identifies if the hashtag(s) used in the tweet are relevant to the followers and followers-of-followers. If there are a large number of relevant followers and followers-of-followers who have high weight profiles for the given hashtag, the tweet is expected to attract much attention. Otherwise, the followers and followers-of-followers may choose to ignore the tweet.

5 Hashtag Recommendation

There are multiple purposes of using hashtags. Some people use them to categorize their tweets. Some use them as mass broadcast media for disasters or special events like elections. Hashtags are also used for brand promotion or micro-meme discussions [12]. Since hashtags are neither registered nor controlled by any user or group, it may be hard for some users to find appropriate hashtags for their tweets. Therefore, recommender systems for suggesting appropriate hashtags to the users are proposed.

5.1 Recommending Hashtags in Twitter with TF-IDF Scheme

The paper by Zangerle et al. [25] assumes that the primary purpose of the hashtags is to categorize the tweets and facilitate the search. The paper recommends suitable hashtags to the user, depending on the content that the user enters without considering user's preference for specific hashtags.

When a user writes a tweet, the recommender system retrieves a set of tweets similar to the given tweet. Similarity score is calculated by TF-IDF scheme. Then, the hashtags are extracted from the retrieved similar tweets and are ranked using their number of occurrences in the whole dataset (OverallPopularityRank

score), their number of occurrences in the retrieved dataset (Recommendation-PopularityRank score) or similarity scores of the tweets (SimilarityRank score). The precision and recall measures of these three ranking scores show that SimilarityRank score is the best among them and the performance of the recommender system is the best when only five hashtags are recommended.

5.2 Suggesting Hashtags on Twitter Using Bayes Model

Another paper which recommends hashtags on Twitter is proposed by Mazzia et al. [17]. Similar to the previous paper, this paper recommends hashtags by observing the content that the user generates. Unlike the previous paper, this paper proposes to use Bayes model which calculates the probabilities of using hashtags.

Before processing the data, the paper cleans the data by removing micro-memes and spams. Micro-memes are detected by identifying tweets which use the same hashtags but are very dissimilar. Spams are filtered by limiting the number of tweets with a particular hashtag from a user. The Bayes model used in this paper is represented by the following formula.

$$p(C_i|x_1, \dots, x_n) = p(C_i)p(x_1|C_i)...p(C_i)p(x_n|C_i)/p(x_1...x_n)$$

where C_i represents the i^{th} hashtag and x_1, \dots, x_n represents the words. $p(C_i|x_1, \dots, x_n)$ is the probability of using hashtag C_i given the words that the user provides and the hashtags with the highest probabilities are recommended to the user. $p(C_i)$ is the ratio of the number of times hashtag C_i is used to the total number of tweets with hashtags. $p(x_1|C_i)...p(x_n|C_i)$ is calculated from the existing data of tweets.

The paper also suggests another model which makes use of Inverse Document Frequency (IDF) to calculate the probability.

$$p(x_1, \dots, x_n|C_i) = p(x_1|C_i)^{(1-t_1)} \dots p(x_n|C_i)^{(1-t_n)}$$

where t_j is the IDF weight of the word x_j .

5.3 High Dimensional Euclidean Space Model

The paper proposed by Li et al. [16] also recommends hashtags based on the information provided by the previous similar tweets. It constructs high dimensional Euclidean space with the words of tweets. Hashtags of the tweets which have the minimal distances are recommended. Distance of tweets in this approach is measured as 1) Euclidean Distance, 2) Ontology Based Distance (OBD), or 3) Centralized Ontology Based Distance (COBD). The comparison of error rates for these three methods shows that OBD method performs the best.

6 Tweet Recommendation

All tweets from the followees of a user are displayed in the user's home page. When the user is following many active users, there are chances that the user

might miss out reading some interesting tweets. With the careful information filtering, important tweets can be chosen and emphasized according to the user's preference.

6.1 User Oriented Tweet Ranking: A Filtering Approach to Microblogs

A personalized tweet filtering approach is proposed [24], which introduces two methods – ranking incoming tweets and ranking targeted users. In the first method, for each user, tweets are ranked according to their probabilities of being retweeted by the user. In the second method, for each tweet, users are ranked according to their probabilities of retweeting the tweet. The underlying assumption is that a tweet is considered relevant and recommended to a user if the user is likely to retweet the tweet.

This paper treats the ranking as a classification problem. First, the classifier is trained with four features, namely author-based, tweet-based, content-based and user-based features.

- Author-based features are features that can be inferred from the user profile, such as number of followers, tweet rate, age of the account, etc.
- Tweet-based features are the syntactic features of the tweet, such as hash-tags, URLs, etc.
- Content based features are the ones related to the information contained in the tweet, such as minimum cosine distance to other tweets.
- User-based features are related to the user whose tweet is being ranked, such as “Is the author following me?”, “Is the author my conversation friend? (i.e. did we mention each other before?)”.

The trained classifier will predicts whether a given tweet is likely to be retweeted by a given user, depending on the above features. Tweets with high probabilities of being retweeted by the target user will be recommended.

7 Retweet Recommendation

Currently, there is no paper about personalized retweet recommendation. However, the work [24] introduced in Section 6.1 can be considered as a retweet recommender system because they are suggesting tweets according to the probabilities of being retweeted by the user. Tadvise [19] identifies whether the hash-tags used in the tweets are relevant to the followers of the targeted user. It can also be used to recommend tweets which the user should retweet for his followers.

8 News Recommendation

Since the tweets are actively written or retweeted by the user, they can be assumed to strongly reflect the user's interest. The following two papers recommend news articles to the user based on the tweets generated by that user.

8.1 Recommending URL from Information Streams

The paper by Chen et al. [4] takes URL as a unit of news information in Twitter. They design and implement a URL recommender system called Zerozero88 which recommends URLs that a particular user might find interesting. This paper uses a *choose-and-rank* approach, where a candidate set of URLs is chosen first and then ranked according to two methods summarized as follows.

The candidate set of URLs are chosen by *followees of followees* and *popularity* methods. The first method is based on the intuition of the locality – neighborhood of a user is considered similar and relevant to the user, such that the URLs posted by a user’s neighborhood are likely to produce high quality recommendations. Therefore, this approach selects only the URLs posted by the followees and followers of followees of a user. In the second method, popularity score of URLs are utilized to select the candidate set.

After choosing the candidate URL set, two methods are used to rank the candidate URL set. The first method uses topic relevance and the second uses social process. In the topic-relevance method, two factors are considered, which are the similarity between the tweets containing candidate URLs and the tweets of this user, and the similarity between the tweets containing candidate URLs and the tweets of this user’s followees. In the social process method, candidate URLs are ranked according to the vote powers of the users who tweet the URL. The vote power of a user is proportional to his follower count, and inversely proportional to the frequency of tweeting.

After testing different combinations of choosing and ranking methods, the paper concludes that using the *followees of followees* approach in choosing candidate set gives the highest probability of recommending the most interesting URLs. For the ranking methods, the method which performs best is the one that combines 1) the similarity between the tweets containing candidate URLs and the tweets of this user, and 2) the vote powers of the users who tweet the URLs.

8.2 Personalized News Recommendation by Analyzing Tweet Contents

The personalized news recommender system by Morales [6] uses tweets to build user profiles and recommend interesting Yahoo news articles to users based on the supervised learning method. The recommendation ranking algorithm is given by the following formula.

$$R_T(u, n) = \alpha \cdot \sum_T(u, n) + \beta \cdot \Gamma_T(u, n) + \gamma \cdot \prod_T(n)$$

where

$R_T(u, n)$ = Ranking of news n for user u ;

$\sum_T(u, n)$ = Content-based relatedness between user u and news n at time T ;

$\Gamma_T(u, n)$ = Social-based relatedness between user u and news n at time T ;

$\prod_T(n)$ = Popularity of news n at time T ;

α, β, γ = Coefficients that specifies the relative weights of the components.

The paper uses spectrum entity extraction system [20] and applies the concept of entity to find the relatedness between tweets and news articles. Content-based relatedness ($\sum_T(u, n)$) captures the intuition that if the news articles and the user's tweets are under common entities, then the news is relevant to the user. Social-based relatedness $I_T(u, n)$ computes the relevant scores by taking into account of the tweets authored by the neighboring users. Other features, such as age, hotness and click count of news articles are also applied in the learning algorithm. For the purpose of testing and evaluation, Twitter user IDs and Yahoo toolbar cookie IDs are linked by the simple heuristic that a user visits his own account more often.

9 Conclusions

Several recommender systems have been proposed to help Twitter users perform information sharing and social interactions more easily. Our paper outlines a taxonomy to classify all the recommendation tasks into a few categories defined around the types of user functions in Twitter. Using the taxonomy, we have surveyed several recommendation methods specially developed for Twitter. To the best of our knowledge, this is the first time a taxonomy is used to classify recommendation tasks in Twitter. Our survey shows that while some recommendation tasks have been well studied, there are some tasks that could be included in future social media mining research. For instance, the current hashtag recommendation systems only consider the content of tweets but not user preferences or effectiveness of hashtags in spreading information. There are also very few works on mention or retweet recommendation. When solutions to these recommendation tasks are developed and evaluated with high accuracies, one can envisage a more comprehensive range of recommendations personalizing the use of Twitter.

Acknowledgements. This research/project is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

1. Armentano, M.G., Godoy, D.L., Amandi, A.A.: Recommending Information Sources to Information Seekers in Twitter. In: International Workshop on Social Web Mining
2. Armentano, M.G., Godoy, D.L., Amandi, A.A.: Towards a Followee Recommender System for Information Seeking Users in Twitter. In: The 2nd International Workshop on Semantic Adaptive Social Web
3. Armentano, M.G., Godoy, D.L., Amandi, A.A.: A Topology-Based Approach for Followees Recommendation in Twitter. In: 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, Barcelona, Spain (July 2011)
4. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and Tweet: Experiments on Recommending Content from Information Streams. In: The 28th International Conference on Human Factors in Computing Systems (2010)

5. Choudhary, A., Hendrix, W., Lee, K., Palsetia, D., Liao, W.K.: Social media evolution of the Egyptian revolution. *Communications of ACM* 55(5), 74–80 (2012)
6. De Francisci Morales, G., Gionis, A., Lucchese, C.: From Chatter to Headlines: Harnessing the Real-Time Web for Personalized News Recommendation. In: *The 5th ACM International Conference on Web Search and Data Mining* (2012)
7. Garcia, R., Amatriain, X.: Weighted Content Based Methods for Recommending Connections in Online Social Networks. In: *The 2nd ACM Workshop on Recommendation Systems and the Social Web*, Barcelona, Spain (June 2010)
8. Golder, S.A., Marwick, A., Yardi, S., Boyd, D.: A structural approach to contact recommendations in online social networks. In: *Workshop on Search in Social Media*, In conjunction with ACM SIGIR Conference on Information Retrieval
9. Hannon, J., Bennett, M., Smyth, B.: Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In: *The 4th ACM Conference on Recommender Systems* (2010)
10. Hannon, J., McCarthy, K., Smyth, B.: Finding Useful Users on Twitter: Twitteromender the Follower Recommender. In: *The 33rd European Conference on Advances in Information Retrieval* (2011)
11. Hannon, J., McCarthy, K., Smyth, B.: The Pursuit of Happiness: Searching for Worthy Followeres on Twitter. In: *The 22nd Irish Conference on Artificial Intelligence and Cognitive Science* (August 2011)
12. Huang, J., Thornton, K.M., Efthimiadis, E.N.: Conversational Tagging in Twitter. In: *The 21st ACM Conference on Hypertext and Hypermedia*, pp. 173–178 (2010)
13. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: Understanding Microblogging Usage and Communities. In: *The 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pp. 56–65 (2007)
14. Krutkam, W., Saikaw, K., Chaosakul, A.: Twitter Accounts Recommendation Based on Followers and Lists. In: *3rd Joint International Information and Communication Technology* (2010)
15. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: *The 19th International Conference on World Wide Web* (2010)
16. Li, T., Yu Wu, Y.Z.: Twitter hash tag prediction algorithm. In: *World Congress in Computer Science, Computer Engineering, and Applied Computing* (2011)
17. Mazzia, A., Juett, J.: Suggesting hashtags on twitter. EECS 545 (Machine Learning) Course Project Report, <http://www-personal.umich.edu/~amazzia/pubs/545-final.pdf>
18. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks.
19. Nasirifard, P., Hayes, C.: Tadvise: A Twitter Assistant Based on Twitter Lists. In: Datta, A., Shulman, S., Zheng, B., Lin, S.-D., Sun, A., Lim, E.-P. (eds.) *SocInfo 2011*. LNCS, vol. 6984, pp. 153–160. Springer, Heidelberg (2011)
20. Paranjpe, D.: Learning Document Aboutness from Implicit User Feedback and Document Structure. In: *ACM Conference on Information and Knowledge Management* (2009)
21. Ricci, F., Rokach, L., Shapira, B. (eds.): *Recommender Systems Handbook*. Springer (2011)
22. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* 24(5), 513–523 (1988)
23. Schafer, J.B., Konstan, J.A., Riedl, J.: E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery* 5(1-2), 115–153 (2001)

24. Uysal, I., Croft, B.W.: User Oriented Tweet Ranking: a Filtering Approach to Microblogs. In: The 20th ACM International Conference on Information and Knowledge Management (2011)
25. Zangerle, E., Gassler, W.: Recommending #-Tags in Twitter. In: Workshop on Semantic Adaptive Social Web 2011, in connection with the 19th International Conference on User Modeling, Adaptation and Personalization (2011)

A Multi-view Content-Based User Recommendation Scheme for Following Users in Twitter

Milen Chechev and Petko Georgiev

Faculty of Mathematics and Informatics, Sofia University, Bulgaria
{milen.chechev, petkog}@fmi.uni-sofia.bg

Abstract. This paper describes recommendation techniques that help users to find potentially interesting people to follow at Twitter. The explored techniques are based on a confirmed assumption that the recent activity of users is indicative of their latest friend preferences. Several content-based recommendation strategies are explored, compared and tested. Among them the foundations for a novel hybridization framework are provided and a multi-view approach towards modeling user profiles is considered. The training and test database is crawled with real users and tweets from the Twitter network. A non-standard evaluation scheme is applied in an offline testing context for the various algorithms. Conclusions are drawn as to the viability, relative predictive power and accuracy of the recommendation approaches.

Keywords: recommendation system, social network, Twitter, content-based, multi-view hybridization.

1 Introduction

In recent years social networks like Facebook and Twitter are taking the lead in providing ways for online communication and data sharing among the Internet community. The figures for the number of active users in each network reach easily over several hundred million. In fact the number of Twitter users soared to just over half a billion a few months ago. With the increasing amount of users who publish data on a regular basis via messages and micro-blogging, the volume of data regarding people's social activity becomes incomprehensibly massive and provides numerous opportunities and challenges for data miners.

With the immense availability of user-generated content brought by social networks including Twitter researchers have started investigating the different aspects of modeling user activity. This focus on activity is of prime importance because users become easily overwhelmed by the huge amounts of information resources and data generated by peers. Recognizing relevant or interesting content among the piles of data has turned into quite a challenge for the average user.

One of the players among the large online social networks where users generate excessive amount of content is Twitter. This internet social phenomenon offers its registered users the ability to post short messages (tweets) of up to 140 characters and

allows each user to gain access to a continuous stream of tweets (known as their timeline) posted by other users explicitly identified as friends (followees). The user is said to be a follower of their friends and the relation between two users is not necessarily reciprocal. In response to the call for systematizing and filtering interesting content for users the presented paper is concerned with evaluating to what extent the latest activity of Twitter users predetermines their choice of most recent followees. It suggests a correlation between the most recently published tweets of users and their latest friend preferences. The paper exploits the mentioned correlation assumption and examines the viability of content-based techniques for recommending potential friends to Twitter users.

2 Related Work

In Twitter user activity is represented with the tweets published by users. Researchers have identified two main streams of recommending potentially intriguing content: (1) prioritizing and filtering interesting tweets and (2) identifying sources of information (other users) that tend to generate messages of value for the subject of recommendation. The authors of [1] are targeting the second goal (2) by examining different strategies for modeling the user profile of a Twitter participant. A number of content-based approaches to recommendation have been evaluated which model the user as a document in the vector space of key words mentioned in the tweets of a user's own timeline or those sent by their followers. In addition the authors have managed to implement a recommendation strategy which resembles collaborative filtering by modeling the user with identifiers of their friends or followers.

[4] on the other hand present a topology-based recommender that inspects the network structure of a target user's relations. They build recommendations by looking at the followees' followers preferences with the intuition that other users who follow the same friends as the target user's ones are likely to already have identified relevant sources of information similar to the target user's tastes. In [8] the same authors are extending their study so that they have also built a content-based recommender representing the content of the tweets from the public timeline of the user which consists of the messages sent by friends. Upon evaluating the relative accuracy of both approaches the authors have come to the conclusion that both algorithms perform similarly at different thresholds for the number of recommended users. [9] are also one of the few that focus on building user followee recommendations by determining relevant features when modeling a user profile (profile popularity based on the ratio between friends and followers and activity based on the range of different posts).

The first goal (1) has been the target of much more scientific focus than the second one. [2], [5], [7] and [11] are concentrating on different strategies for recommending interesting content available from the links in the tweets. The cited links to external resources provide a means for extending the volume of topics mentioned in tweets and give the users a chance to express themselves beyond the constraints of 140 characters. [2] look at ways for modeling user's preferences by building semantically-enriched profiles. They have adopted entity resolution by unifying the set of concepts

mentioned in users' tweets and utilize that approach to recommend news items similar to the ones referred to by the user. [5] in contrast focus solely on recognizing potentially interesting URLs for the subject of recommendation. The authors take into consideration 3 dimensions when forming their URL recommendations - the source of the URL (globally popular or mentioned in the user's friend network), the type of the content-based user profile and the level of popularity of the URL in the user's friend network. [11] develop a system for ranking and searching of interesting URLs. The relevance of the URL is defined based on a combination of different link-specific features such as link authority, link popularity and link longevity (the time between the first and last mention of the item). [7] look at ways for evaluating and optimizing the performance of collaborative-filtering approaches in the context of recommending links in Facebook. The authors have the necessary means for implementing collaborative-filtering as there exists an explicit user binary rating (like/dislike) for the links. In contrast, in Twitter the explicit way for classifying some content as interesting is through marking tweets as favorites. However, the amount of favorites compared to the number of generated tweets is quite small.

The authors of [3] heavily research the first goal (1) by considering a large set of features that contribute to the ranking of how interesting tweets are. When summarizing the results of their experiments they conclude that among the best features for classifying tweets as interesting are the presence of a URL in the text and the length of the tweet.

To sum up, it should be mentioned that the goals (1) and (2) that have been identified here are by no means exhaustive with regard to the ways for recommending content of interest to users. One alternative example is the work of [6] who try to predict information spreading in Twitter by building a model that calculates the probability of a tweet to be retweeted. The process of retweeting a tweet is in essence forwarding the tweet so that other users who follow the retweeter can gain access to the same piece of information.

3 Sample Network Data Harvesting and Analysis

Providing recommendations for potentially relevant users to follow on Twitter is an important and challenging task that requires the close observation of the users' social activity. The first step towards understanding the users' preferences inferred from their online social behavior is gathering activity and relation data from a large enough representative subset of the Twitter network. The targets of the harvesting process are users with public profiles in Twitter (unprotected accounts) which means that their tweets are readily available for viewing by other Twitter users. Besides, only users who have identified their main language of communication in the social network as English are considered.

To build a sample network of users the Twitter API is utilized and a breadth-first search is performed by following the friend relations in Twitter of the authors of the paper. Initially, the search is limited at depth 2 or in other words the authors' friends and their friends are included in a network which amounts to 33737 Twitter users in total. As the aim of the research is evaluating ways to recommend users to follow,

what is also needed is information regarding the friend relationships for a subset of the gathered users' general profiles.

So far the harvested data does not form a compact graph structure where the users are mainly following other users in the same built network. In order to understand what the density of the edges in the user graph is, 1054 of the users are selected with a relatively large number of friends (around 5000 for the majority of them). For those users a calculation is made concerning the number of friends they have in our network and the results are that on average 250 friends are already among the 33737 users. To further enrich the graph of known users with relations we gather for the handpicked ones the 7 latest added friends.

It is worth pointing out that the focus of the harvesting process is obtaining as much information as could be extracted for the latest friends of targeted users. The manner in which users are gathered is crucial to the evaluation method employed later for assessing the effectiveness of the social recommendation strategies and as far as the authors are aware it has not been observed in other related work. The main idea behind the discussed approach is tightly connected with the latest activity of the Twitter users. The intuition is that the latest tweets of the users are representative of their latest interests in topics and concepts and it is highly likely that the latest added friends of the users reflect these preferences.

For the purposes of evaluation and testing a representative set of users are picked whose friend relations reflect the observed distribution (table 1). After the latest expansion of the network of known users their total number reaches 40306. The test users are selected so that their median for the friends count is exactly the one found for the current user set (table 1). Furthermore, the public general profiles of their latest 5 friends are harvested. The user's latest followees are easily obtained with the Twitter REST API which returns their identifiers in reverse order of addition by the target user.

Table 1. Statistics for the expanded network of users (40306)

| | <i>Friends</i> | <i>Followers</i> |
|----------------|----------------|------------------|
| <i>Min</i> | 0 | 1 |
| <i>Max</i> | 766119 | 22914722 |
| <i>Average</i> | 1845 | 24989 |
| <i>Median</i> | 227 | 249 |

Now out of our further expanded network of 43562 users a subset (21779) is selected such that the test users and their 5 latest friends are inside. Finally, for these users the latest 200 tweets on average are crawled (if they have as many). The result is a database of 4365222 tweets gathered in a period of one week in April 2012.

4 Recommendation Strategies

In Twitter the social activity of users is represented with the short messages they publish and the network dynamics of their friends and followers. The fluctuations in the

structure of the users' social graph happen because of their deliberate decision to follow and unfollow certain users. The act of adding a user to a friend set is much more common than removing a user and it reflects the conscious decision of a Twitter participant to begin following the tweet feed of a potentially interesting source of information. Understanding the reasons for choosing one friend over another is essentially the core of any viable recommendation strategy that suggests users of interest to follow. It is assumed that what people talk about lately in Twitter is what they would like to hear more about which in turn unambiguously determines their latest friend choices. Therefore, the next part of the paper explores several content-based techniques that model user preferences based on tweet activity and lays the foundations for the verification of the correlation between friends and tweets.

So far the authors have come across several content-based strategies for building a user profile presented by other authors but none of them has considered the predictive power of the tweet text elements in isolation. For example, the authors of [1] have focused on 4 strategies for building a pure content-based user profile but all of them examine the whole content of the tweets. The tweet has a rich structure of meta-data which deserves its own attention. Apart from free text, a tweet may contain links to external resources, mentions of other Twitter users (beginning with the character @) and hashtags (explicitly marked key words represented with a continuous sequence of characters without intervals preceded by the symbol #). A challenging question which needs answering is to what extent the different elements of the tweet text represent users' preferences for following one user over another.

To address the issues raised by the above mentioned question a different approach is taken towards building a content-based user profile which takes into account the separation of concerns between the tweet free text and the entity meta-data. A profile of a user is built by examining a set of views based on the published content of their latest tweets:

(V1) Free-text based - considering only the free text without additional elements in the tweets

(V2) Hashtag based - considering only the hashtags of the tweets

(V3) Link based - considering only the domains referred to by links in users' tweets

(V4) A weighted hybrid approach that combines (V1) and (V2)

The first view (V1) of the user content-based profile which is brought in the spotlight is concerned with the text of the tweet freed from all meta-data. For each user u in the network of 21779 Twitter participants the text of the latest 200 tweets is merged into one document of keywords. Thus all things that user u has said in their latest 200 tweets apart from links, hashtags and user mentions is subjected to indexing. For the indexing step Apache Lucene¹ has been used. It is a Java library for information retrieval that offers functionality for indexing large volumes of text data and more importantly gives the ability to search through the corpus of indexed documents for relevant matches that satisfy an information need presented with a query. As a result of the indexing, each user u is modeled in the term-vector space of the words they

¹ <http://lucene.apache.org/>

have written in tweets. The extracted textual content is processed with enhanced tools for text analysis - a lower-case filter is applied, tokens which belong to a list of smart stop words are removed and a Porter Stemmer is finally used to process the remaining of the text stream. Consequently, the user profile in this view is a vector of key terms each of which is given a special weight. The used values for weights are given according to the commonly used TF-IDF weighting scheme (term frequency - inverse document frequency). Formula (1) gives the equation for the TF-IDF value where $tf(i, d)$ is the term frequency of term i in user document d , $df(i)$ is the number of documents containing i at least once and N is the total number of indexed users.

$$w(i, d) = tf(i, d) * \log \frac{N}{df(i)} \quad (1)$$

The TF-IDF weighting scheme gives a higher score to terms that are uniquely mentioned in few of the documents but that have high frequencies within the document. Thus words which are frequently used by a user but which are not common among all Twitter users achieve high scores to denote that such words are much more representative of the user's topic preferences.

Next given that user profiles have been built as vectors of keywords, a question arises how a recommendation can be designed to fit in the chosen approach. For a user u all other users in the built network are compared according to the Apache Lucene scoring scheme. The users (presented as documents) who have the highest similarity score to user u become the objects of the recommendation for u . The Apache Lucene scoring scheme² uses a modified version of the widely adopted cosine similarity for measuring how close to one another two documents are. The advantage of the measure is that shorter and longer documents can be compared safely by considering only the angle between them in the high dimensional vector space and not their length. The cosine similarity (formula 2) between 2 document vectors A and B essentially sums the product of the weights of the words that are mentioned by both users (corresponding to vectors A and B) and normalizes the sum by dividing it by the product of the length of the vectors. The range of the resulting similarity values lies in the closed interval $[0, 1]$ with the value 1 being reached only when the two document vectors are identical. In this way users that have mentioned the same set of key words will have higher similarity scores over others with fewer matching words in tweets.

$$\cos(A, B) = \frac{\sum_{i=1}^n A_i * B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} * \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (2)$$

Thus the motivation for our similarity ranking approach is that the latest users' tweets capture their latest interests and the text of these tweets can be used to determine the users' choices of recently added friends. In other words what is subject to verification here is whether the user and their latest friends talk about the same topics as reflected in their latest tweets.

² http://lucene.apache.org/core/3_6_0/api/all/index.html

The next view (V2) that is proposed is concerned with the hashtag entities and their power to represent the users' tastes. The hashtags are explicitly posted by the user in their tweets and intuitively such entities can be used to capture the explicitly defined topic preferences. As seen in table 2, the average number of hashtags a user in our network has posted in their latest 200 tweets is 57 while the median is 32. What is more, of 21779 almost 90% (19338) have used at least one hashtag in their tweets. It is therefore safe to assume that the hashtags are a reliable source for modeling user preferences.

Table 2. Statistics for the number of hashtags in users' tweets

| | <i>Number of tweets with at least one hashtag</i> | <i>Total number of hashtags in the tweets</i> |
|----------------|---|---|
| <i>Min</i> | 0 | 0 |
| <i>Max</i> | 397 | 2157 |
| <i>Average</i> | 40 | 57 |
| <i>Median</i> | 26 | 32 |

Now to model the user preferences in a view of their hashtags we proceed in a manner similar in spirit to the previous text-only-based approach. All the hashtags of a user are combined into one document which is analyzed and indexed with Apache Lucene. The user is again represented as a vector of keywords (hashtags) but the difference here is that the enhanced analysis step is skipped since each hashtag denotes a special meaning regardless of whether it coincides with a stop word in a natural language. Furthermore, often hashtags are phrases and the Porter Stemmer becomes inapplicable here. Only a lower-case filter is used.

The third view (V3) of the user profile under consideration is a novel one in a sense that as far as the authors are concerned none has evaluated user preferences in such a manner. The view is based on the domain of the links mentioned in the tweets. When citing external resources in their tweets the users refer to different URLs which in turn refer to a smaller set of site domains representing the sites that are visited. On average 18 different domains are mentioned in the latest 200 tweets of the users in the network.

Motivated by the fact that users with similar tastes probably visit a common set of sites we proceed to build for each user a document of mentioned link domains. Again the document is indexed with the help of *Apache Lucene* without applying special text analysis. The user profile in this view is a vector of site domains weighted by the standard TF-IDF score.

It is worth pointing out that the Twitter API provides the means for finding the real domain behind links in tweets. Since the maximum number of allowed characters in a tweet is 140 very often a URL shortener is used to shrink the size of the original URL. However, for each URL in a tweet the Twitter API returns the shortened URL, the expanded (original) version of the URL and how the URL is displayed. Yet it is possible that a small proportion of the users prematurely shorten the published links which are later subject to further shortening by the standard Twitter tools.

So far the methods for building a user profile represent various aspects of the users' interests without taking advantage of their joint power. Using the recommendation strategies (V1), (V2) and (V3) in isolation is important for the understanding of the relative predictive power of the tweet elements but is certainly not enough for a comprehensive recommendation. The isolation does not take advantage of the fact that users provide more insight into their preferences by publishing inter-related content. Next a hybridization framework (V4) that can be used to combine a set of content-based approaches is proposed and is motivated by the need to accurately model the user tastes with more of the available information regarding the user's social network activity. The suggested framework for combining strategies utilizes a weighted hybrid according to the classification presented by [10]. Although the actual implemented hybrid combines (V1) and (V2) the approach works for an arbitrary number of content-based ranking strategies. It should be noted that the framework suggests a non-standard approach towards ranking users' similarity which is evaluated at the next chapter.

Let u be a user who is the subject of our recommendation. Let v denote a potential user whose profile is evaluated as to how similar it is to u . Let $x_1(u, v)$ and $x_2(u, v)$ be the cosine similarity values (lying in the closed interval of $[0, 1]$) computed for the users u and v by two content-based recommendation algorithms. Then the aim of the hybrid weighting scheme is to combine these values in a way that gives a better representation of the similarity between users u and v .

Let the actual similarity between these users be a target function $g(u, v)$. The requirement for this function is to give higher scores (preferably close to 1) for the recently added friends of u . In addition, it should be decreasing for users who were added long time ago by u . The idea behind such a modeling is that most of the u 's recently found friends are intentionally added because of their latest activity (published tweets) in Twitter which is similar to the content with which u has been recently engaged. So let $fn(u)$ denote the number of the friends for user u and $reverseOrder(u, v)$ denote the reverse order of addition for user v as a friend of u if user v is indeed a u 's followee. To illustrate this point if v is the most recently added friend of u , then $reverseOrder(u, v)$ should give a value of 1. Then the functions $o(u, v)$ and $t(n)$ are defined as follows:

$$o(u, v) \begin{cases} reverseOrder(u, v), & \text{if } v \text{ is a friend of } u \\ fn(u) + 1, & \text{if } v \text{ is not } u\text{'s followee} \end{cases} \tag{3}$$

$$t(n) = \frac{o(u, v)}{fn(u)} \tag{4}$$

The goal is to find a function $g(u, v) = g'(t)$ such that it is non-increasing in the interval $[0, 2]$ and approximately gives values in the interval $[0, 1]$. One such a function is (5) which is used in the hybridization approach. Experimentally it has been found out that this function gives the best predictive power.

$$1 - \frac{1}{1 + e^{-t}} \tag{5}$$

Now that the values for x_1 , x_2 and g are clearly defined a training set for a sample of the users is built. The training set is simply a list of m triples (x_1, x_2, g) for a set of

relations between users and their friends. The aim of this set is to be used as the entry point for a regression strategy which tries to find weights w_0 , w_1 and w_2 such that the error between f (as defined in (6)) and g is minimal.

$$f(x_1, x_2) = w_0 + w_1 * x_1 + w_2 * x_2 \quad (6)$$

The error between f and g is given by the objective function (7), and it is optimized (minimized) by running a gradient descent classifier. The used method for building the hybrid model uses the *Java* library for machine learning *Weka*³.

$$J(w_0, w_1, w_2) = \frac{1}{2m} * \sum_{i=1}^m (f(x^{(i)}) - g^{(i)})^2 \quad (7)$$

5 Offline Testing and Evaluation

The evaluation of the recommendation strategies described in the previous section is intended to address two issues that have been raised – namely, whether the latest Twitter user activity presented by the recent content of the published tweets determines the choice of latest friends to follow and whether this assumption can actually be used for the successful application of the built recommendation techniques. In fact, the two questions can be answered by testing the accuracy of the considered recommendation schemes against a carefully selected set of rules whose essence is based on how recently the friends for a test user have been added. The test set consists of 1004 users for whom the latest 5 friends are known and have profiles built according to the proposed algorithms (V1) to (V4). Only the latest followees of the users are being considered for measuring the algorithm accuracy in the test setting and only they are deemed as relevant recommendations. The rationale behind this evaluation scheme attempts to capture the intuition that the latest tweets of the users are indicative of who they deliberately choose to follow. We only test against 5 recently added users because the rate at which new friends are added for the majority of users in a social network is not excessive.

To see how well the algorithms perform a ranked list of recommendations for each test user is provided. This ranked list is nothing more but all 21779 users sorted in decreasing order by their similarity value to the current test user as computed by each of the algorithmic approaches. What is expected is that the most relevant recommended users appear among the first in the list. For each test user all of their friends are removed from the ranked list except the latest added 5 followees. Depending on the position of the latest 5 followees in the recommendation list it can be inferred to what extent the tweet content corresponds to the user interests in friends.

The figure 1 shows the number of test users with at least one of their latest friends ranked by an algorithm with similarity value greater than 0. When applying each of the recommendation approaches for some of the users none of the latest added 5 friends were found to be similar by some of the algorithms. This is partly due to the fact that some of the test users as well as some of their friends do not have a representation in the corresponding profile view. The representation is missing when the user

³ <http://www.cs.waikato.ac.nz/ml/weka/>

has not resorted to using the enriched entities such as hashtags and links in their latest published 200 tweets. A recommendation strategy that uses the entities as the source of content-based profiles fails to give a similarity value greater than 0 when either the tested user or their friend is lacking a representation. The other reason for classifying two users as completely dissimilar is that sometimes there is no common term mentioned by both users. In such a case the formula (2) for the cosine similarity naturally produces a 0 output. In contrast, the event of missing representation is reflected by simply assigning a default value of 0 for the similarity and it is not a result of a computation. In either case the dissimilar users are listed at the end of the queue of potential recommendations with a farthest position ranking and are counted towards the final measurement of the algorithmic accuracy.

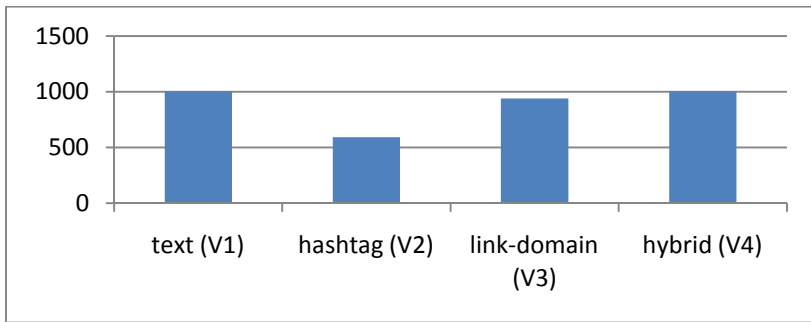


Fig. 1. Actual number of test users with at least one of their latest friends ranked by an algorithm

What can be noticed from figure 1 is that the hashtag based strategy (V2) produces the fewest number of ranks which can be explained by the fact that hashtag presence in tweets is relatively scarce compared to linked URLs and written free text. The strategy (V2) can be used for a set of users that regularly create hashtags in their messages but these entities cannot be solely relied upon. An interesting observation that is worth mentioning is that the hybrid approach (V4) is able to produce a rank of similarity greater than 0 as long as either (V1) or (V2) computes a non-zero value. Thus users that have representation in only one of the views (free-text or hashtag) can still benefit from the recommendations computed by the strategy based on the view. Consequently, the hybrid approach extends the coverage of recommendations.

To understand the viability of the proposed recommendations and to test the correlation between latest published tweets and added friends what is needed to be checked is where the relevant recommendations for a user (latest 5 friends) appear in the ranked list of all users. One of the most suitable measures for such evaluation purposes is recall which has its roots in information retrieval. The recall is the percentage of actually retrieved relevant recommendations and it increases with the number of recommended users. As the number of actual recommendations is usually not strictly defined beforehand, the recall at different levels of suggested items (users) is looked at. In our case all the users with built profiles are ordered in decreasing order of similarity to a test user which means that the recall at level 21779 (which is the total number of examined users) is 100% (or 1 if normalized to values between 0 and 1).

However, as our goal is to find whether the latest 5 friends are at the front of the recommendation list, the recall at levels below 100 is computed, i.e. the relevant friends are searched only among the first 100 recommended users. The figure 2 shows the recall curves for user profile views (V1) - (V4). The best performing contenders are the text-based and hybrid strategies.

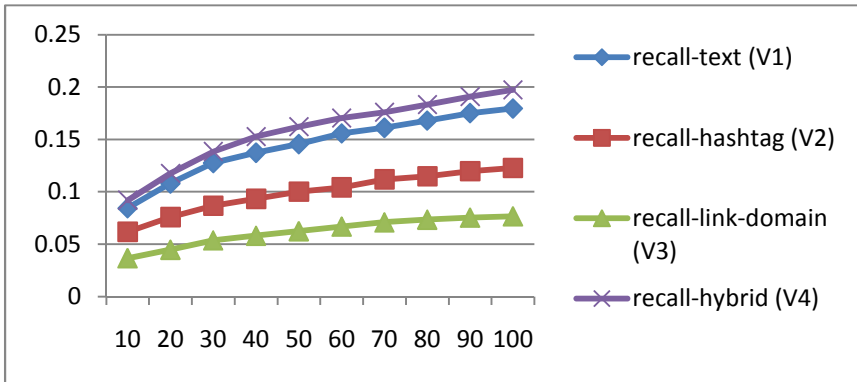


Fig. 2. Recall at various levels for the number of recommended users

The best among the considered strategies succeeds in retrieving 20% of the latest added friends among the first 100 users ordered by their similarity to a test user and 12% among the first 20 ranked users. This is an indication that indeed at least some of the latest chosen friends to be followed are picked by a user because they have similar topic interests as reflected by the content of their most recent tweets. It can also be inferred that such a content-based strategy is reasonably viable for recommending users to follow.

The figure 2 also illustrates the relative predictive power of the strategies compared to one another. Among the isolated strategies (V1), (V2) and (V3), the best one in terms of recall is (V1). Despite the fact that links and hashtags are becoming increasingly common as a way of expressing concepts and preferences in Twitter, the free text of the tweets remains the strongest indicator of the topic interests of a user. Furthermore, the link-domain-based strategy (V3) does not seem to be an extremely promising candidate for making recommendations compared to the other strategies. The result can be explained by the fact that although users with similar interests are expected to visit similar sites, the number of different domains linked by users in tweets is not so great. To be more specific on average 18 domains are cited in the latest 200 tweets of a user. Besides, a large number of the visited sites provide news in various areas of everyday life. In this case the site domain in its own right is insufficient to determine the topic of interest for users. The whole link must be followed and the content of the linked resource must be analyzed for a better representation of what the user cares about.

Another important observation that should be highlighted is that the hybrid approach performs better than all other strategies at all levels of recall. This confirms the intuition that when considering the user activity the more of it is employed to model preferences, the better the final user representation is. To further analyze the accuracy of the recommendation techniques the Mean Average Precision is calculated for each one (figure 3).

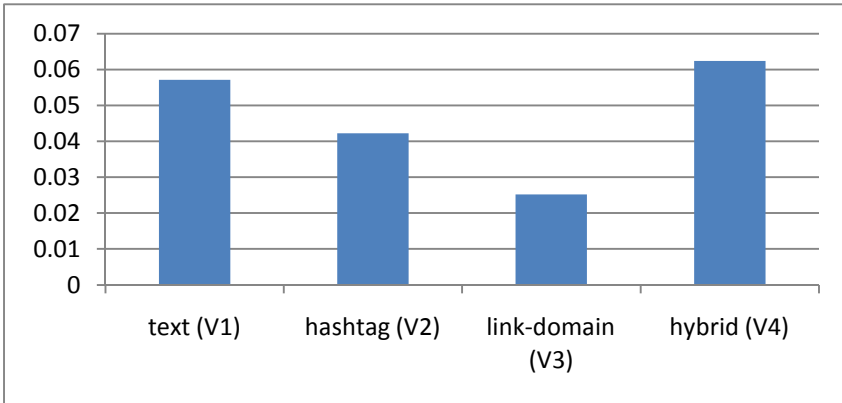


Fig. 3. Mean Average Precision for the different recommendation techniques

The Mean Average Precision is a single-figure measure across all levels of recall that has been found to have especially good stability and to be a reliable source for representing algorithmic accuracy. It is the mean of the average precision among a set of queries. Let q_j be a user from the test set Q and $\{d_1, d_2, d_3, d_4, d_5\}$ be the set of relevant recommendations (latest friends) for user q_j . Let R_{jk} be the set of top recommended users for following until user d_k is found in the suggested friend list. Then the Mean Average Precision is expressed by formula 8.

$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{5} \sum_{k=1}^5 Precision(R_{jk}) \tag{8}$$

The precision is the percentage of the relevant recommendations among the suggested ones. For each test user the precision of retrieving each of their latest friend is calculated at the level of the position of the friend in the global ranking of users. That is to say if friend f appears at position N in the ranked list, the precision at level N is computed. The precision for the latest 5 friends is averaged for a user and the mean of all averaged values is computed to produce a single floating point value between 0 and 1. It is worth pointing out that since only 5 friends for a given user (out of the set of 21779 excluding the user’s other friends) are classified as relevant the values for the precision are naturally low.

Figure 3 shows that the hybrid approach has the highest accuracy again. It also demonstrates a rather intriguing outcome – the hashtag-based strategy (V2) is much

closer to the leaders in terms of precision than it is in terms of recall. On the other hand, the hashtag-based scheme (V2) according to figure 1 is the one to provide the fewest similarity results greater than 0. This means that if the latest friends are indeed ranked by (V2) with a similarity greater than 0 then they are much more likely to be in the top recommendations in the list of suggested users. So, explicitly defined notions in the form of hashtags are candidates for a good topic-of-interest predictor when they are present in the user's tweets content.

6 Conclusions

In this paper an evaluation of several content-based strategies for modeling a Twitter user's profile has been provided with the aim of investigating the viability of techniques for recommending potential friends for following. The recommendations serve as a basis for the verification of the assumption that recent Twitter user activity in tweets is indicative of the latest friend choices made. The evaluation method used is non-standard in the sense that it has not been observed to be used in other related work.

Among the 4 evaluated recommendation strategies the best one in terms of coverage and recall proves to be the hybrid one (V4) which combines the higher precision results from the hashtag-based algorithm (V2) (compared to the link-based (V3)) and the coverage of actual made recommendations for friends of (V1). The hybrid strategy succeeds in taking advantage of the availability of more user-generated content by considering both the free text and hashtags published in users' tweets. As on average 20% of the latest added friends are retrieved among the first 100 recommendations by the hybrid strategy, the correlation between latest added friends and tweet content is confirmed. Besides, here only hashtags and free text have been employed.

The improvement of the hybridization approach in terms of accuracy suggests that more of the unused tweet content should be incorporated for a complete representation of the users' preferences. The analysis of the content of the externally linked resources is the next step towards bringing the full power of recent user activity into the act of effective recommendations. What remains to be done is finding a hybridization strategy that captures all aspects of communication and tweet activity in Twitter. Yet this paper has been able to show that research in that area of hybridization is not in vain and is in fact the approach to be pursued in the quest for filtering relevant and interesting content for users.

References

1. Hannon, J., Bennett, M., Smyth, B.: Recommending Twitter Users to Follow Using Content and Collaborative Filtering Approaches. In: Proceedings of RecSys 2010, New York, USA, pp. 199–206 (2010)
2. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 1–12. Springer, Heidelberg (2011)

3. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.: An Empirical Study on Learning to Rank of Tweets. In: Proceedings of COLING 2010, pp. 295–303 (2010)
4. Armentano, M.G., Godoy, D.L., Amandi, A.: A topology-based approach for followers recommendation in Twitter. In: 9th Workshop on Intelligent Techniques For Web Personalization&Recommender Systems, ITWP 2011 (2011)
5. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and Tweet: Experiments on Recommending Content from Information Streams. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 1185–1194 (2010)
6. Zaman, T.R., Herbrich, R., Gael, J., Stern, D.: Predicting Information Spreading in Twitter. In: Workshop on Computational Social Science and the Wisdom of Crowds (2010)
7. Noel, J., Sanner, S., Tran, K.N., Christen, P., Xie, L., Bonilla, E.V., Abbasnejad, E., Penna, N.D.: New Objective Functions for Social Collaborative Filtering. In: Proceedings of the 21st International Conference on World Wide Web, New York, USA, pp. 859–868 (2012)
8. Armentano, M.G., Godoy, D., Amandi, A.: Towards a Follower Recommender System for Information Seeking Users in Twitter. In: Proceedings of the IJCAI 2011: International Workshop on Social Web Mining, Barcelona, Spain (2011)
9. Garcia, R., Amatriain, X.: Weighted Content Based Methods for Recommending Connections in Online Social Networks. In: Workshop on Recommender Systems and the Social Web, Barcelona, Spain, pp. 68–71 (2010)
10. Burke, R.: Hybrid Web Recommender Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 377–408. Springer, Heidelberg (2007)
11. Phelan, O., McCarthy, K., Smyth, B.: Yokie - A Curated, Real-time Search & Discovery System using Twitter. In: 3rd Workshop on Recommender Systems and the Social Web, Boston, USA (2011)

Spam Fighting in Social Tagging Systems

Sasan Yazdani¹, Ivan Ivanov², Morteza AnaLoui¹,
Reza Berangi¹, and Touradj Ebrahimi²

¹ Iran University of Science and Technology, Tehran, Iran

² École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
sasan_yazdani@comp.iust.ac.ir, {ivan.ivanov,touradj.ebrahimi}@epfl.ch,
{analoui,rberangi}@iust.ac.ir

Abstract. Tagging in online social networks is very popular these days, as it facilitates search and retrieval of diverse resources available online. However, noisy and spam annotations often make it difficult to perform an efficient search. Users may make mistakes in tagging and irrelevant tags and resources may be maliciously added for advertisement or self-promotion. Since filtering spam annotations and spammers is time-consuming if it is done manually, machine learning approaches can be employed to facilitate this process. In this paper, we propose and analyze a set of distinct features based on user behavior in tagging and tags popularity to distinguish between legitimate users and spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a dataset of social bookmarks.

Keywords: Social tagging systems, Social spam, Spam detection, Spammers, User behavior, Tags popularity.

1 Introduction

Social systems (networks) allow users to store, share, search and consume content (resources) online. Tagging in social systems has become increasingly popular since the transition to Web 2.0, as it simplifies and eases search and retrieval of information, and allows users to access these information globally while interact and collaborate with each other. Tags can be assigned to different types of resources, such as images, videos, publications and bookmarks, making it a valuable asset to search engines on the Internet and in social tagging systems.

A few challenges have been identified in research community as important in social tagging systems, namely tag recommendation, tag propagation and tag relevance. For example, *tag recommendation* approaches suggest appropriate tags to resources (e.g., videos) in order to make it easy for users to search and access information in social systems [11]. In order to speed up the time-consuming manual tagging process, tags can be automatically assigned to images by making use of *tag propagation* techniques based on the similarity between image content (e.g., famous landmarks) and its context (e.g., associated geotags) [7]. Since user-contributed tags are known to be uncontrolled, ambiguous and personalized, one of the fundamental issues in tagging is how to reliably determine

the relevance of a tag with respect to the content it is describing [1]. The fact that tags are user-contributed enables spammers to pollute social systems with irrelevant or wrong information (spam) to mislead other users, and to damage the integrity and reliability of social systems. In general, spam on the Internet is created to trick search engines by giving the spam content higher rank in the search results for advertisement or self-promotion purposes. Various techniques have been proposed in the literature for combatting spam, for example, Google's PageRank [10] and TrustRank [20].

Tags play a vital role in social systems, since it is important that resources in these systems are assigned with relevant tags. Injection of irrelevant tags and inappropriate content in social systems can be performed mainly in two ways. First, spammers can use legitimate resources and assign irrelevant tags to them for the purpose of advertisement or self-promotion [3]. Second, spammers can use popular and high ranking tags to describe a spam resource and boost its rank [12]. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers.

In this paper, we propose a set of distinct features that can efficiently identify spam users in social tagging systems. The introduced features address various properties of social spam and users activities in the system, and provide a helpful signal to discriminate legitimate users from spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a dataset of social bookmarks.

The rest of the paper is organized as follows. Section 2 reviews the most recent related work. In Section 3, we propose a set of distinct features for spammer detection based on user behavior in tagging and tags popularity. Evaluation methodology and dataset are presented in Section 4. In Section 5, we compare several supervised learning approaches applied to the proposed features and analyze their performance. Finally, Section 6 concludes the paper with a summary and some perspectives for future work.

2 Related Work

The research work presented in this paper is related to different fields including tagging, tags characteristics, impact of spam and fighting spam in social systems. Therefore, the goal of this section is to review the most relevant work in the fields of spam impact on tagging and fighting against spammers in social tagging systems.

2.1 Tag Characteristics and Spam Impact in Tagging

Xu *et al.* [19] studied the characteristics of tags and categorized them into five groups: content-based which are used to describe the category an object belongs to, context-based which provide contextual information about the resource, attribute tags which point unnoticeable characteristic of a resource, subjective tags

which describe users point of view, and organizational tags that are personal like reminders and scheduler tags. Furthermore, the authors introduced criteria that must be fulfilled in order for a tag to be considered good. According to their study, a well-defined tag has properties like coverage of multiple facets of the resource, employing popular tags, excluding unlikely tags such as organizational or subjective tags.

Koutrika *et al.* [12] were the first to explicitly discuss methods of tackling spamming activities in social tagging systems. The authors studied the impact of spamming through a framework for modeling social tagging systems and user tagging behavior. They proposed a method for ranking content matching a tag based on taggers reliability in social bookmarking service Delicious. Their coincidence-based model for query-by-tag search estimates the level of agreement among different users in the system for a given tag. A bookmark is ranked high if it is tagged correctly by many reliable users. A user is more reliable if his/her tags more often coincide with other users tags. The authors performed a variety of evaluations of their trust model on controlled (simulated) dataset by populating a tagging system with different user tagging behavior models, including a good user, bad user, targeted attack model and several other models. Using controlled data, interesting scenarios that are not covered by real-world data could be explored. It was shown that spam in tag search results using the coincidence-based model is ranked lower than in results generated by, e.g. a traditional occurrence-based model, where content is ranked based on the number of posts that associate the content to the query tag.

2.2 Spam Fighting in Social Tagging Systems

Heyman *et al.* [5] classified anti-spam (or spam fighting) approaches into three categories: prevention-, rank- and identification-based. *Prevention-based* approaches employ series of mechanisms to keep spammers out of social tagging systems, such as CAPTCHA [16] and reCAPTCHA [17], or make it hard for spammers to pollute social system by restricting access, limiting number of resources a user can interact with, or requiring registration fee. Usually, prevention-based approaches are used as complementary defense systems to rank- or identification-based approaches. *Rank-based* approaches are very common in search by query scenarios and are used to demote spam content in order to return most legitimate resources on top of search results. *Identification-based* (or detection-based) approaches create a model from users' information, activities and interactions to efficiently detect and filter spam users (or content) from social tagging systems.

Bogers *et al.* [3] proposed an approach to identify spammers in social bookmarking systems such as BibSonomy and CiteULike. The approach is based on user language models assuming that spammers and legitimate users use different language jargons when posting. To detect spam users, they learned a language model for each post, and then measured its similarity to the incoming posts by making use of Kullback-Leiber (KL) divergence. The spam status of a new post takes the status of the most similar language model. Status of a user is determined by grouping all

users posts. This approach was evaluated on BibSonomy dataset for spam detection, proposed at ECML PKDD Discovery Challenge 2008 [6].

Krause *et al.* [13] employed a machine learning approach to detect spammers in BibSonomy. They investigated a framework for detecting spammers. The authors assumed that spammers usually use different strategies for polluting social bookmarking systems such as creating several accounts, publishing a particular post several times, and using semantically diverse tags to describe a bookmark and teaming up with other spammers to give good votes to each other. The authors investigated features considering information about a users profile, location, bookmarking activity and semantics of tags. By making use of these features, and naïve Bayes, support vector machine (SVM) classifiers, logistic regression and J48 decision trees, they were able to distinguish legitimate users from malicious ones. This study represents a good foundation for future machine learning spam detecting approaches.

Markines *et al.* [14] proposed six different tag-, content- and user-based features for automatic detection of spammers in BibSonomy. The authors used features representing the probability of a tag being spam, number of advertises per post and number of valid resources per user posts. It was shown that “TagSpam” feature (tag diversity in posts) is the best predictor of spammers among all other features, because spammers tend to use certain “suspect” tags more than legitimate users. Although their work showed promising results, most of the proposed features rely on an infrastructure to enable access to the content, and must be recalculated periodically to remain reliable. Therefore, the feasibility of the proposed features depends on the circumstances of a particular social tagging system.

Although BibSonomy is the most popularly explored domain for spam fighting, there are researchers who developed techniques for other social systems, such as Delicious, YouTube, MySpace or Twitter. Ivanov *et al.* [8] surveyed recent advances in techniques for combatting noise and spam in social tagging systems, classified the state-of-the-art approaches into a few categories and qualitatively compared and contrasted them.

3 Distinct Features

In this section, we first present a model of a social tagging system and then introduce a set of distinct features to distinguish between legitimate users and spammers in social systems.

Social tagging systems allow users to assign tags to resources shared online in order to enrich a resource with metadata and facilitate search for a particular resource, as previously explained in this paper. General model of a social tagging system is represented as a hyper-graph structure called *folksonomy* where the set of nodes consists of three kinds of objects: users, resources and tags, and hyper-edges connect these objects based on their relations [2]. The folksonomy can be defined with a quaternary structure $F = (U, T, R, P)$, where U represents the set of users u in the system, T is the set of tags t posted by users, R shows

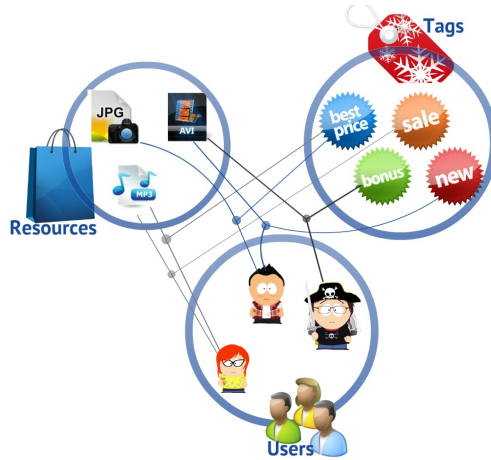


Fig. 1. An example of folksonomy representing a social tagging system with 3 users, 4 tags, 3 resources and 5 posts

the set of resources r and P defines the relation existing between tags, users, and resources. A relation linking a user, a tag and a resource represents a post. A post p in folksonomy can be represented with a triple $p = (u, r, T_u)$ which relates a user u who associated a resource r with a set of n tags $T_u = \{t_1, t_2, \dots, t_n\}$. Figure 1 shows an example of folksonomy with 3 users, 4 tags and 3 resources.

Distinguishing between legitimate users and spammers in social tagging systems can be regarded as a classification problem. The most important part in any classification problem is the extraction of a good set of features from data. Features should represent data well to achieve good classification rate. Features are used to reduce the dimensionality of data while keeping important and relevant information. After studying the BibSonomy user behavior, we introduce 16 distinct features for each user from the evaluation dataset. Each user is represented with a feature vector consisting of 16 features which can be used by any known classifier to fight spam. In the following, we describe the proposed features in details, discuss the observation behind them and explain how to extract them out of a folksonomy.

3.1 LegitTags/SpamTags

We studied users behavior in BibSonomy and found out those spammers and legitimate users tend to use different languages for their posts. Spammers often use a fraction of legitimate user vocabulary, mostly popular tags, to gain higher ranks. Apart from this fact, they have a very distinctive jargon which is barely used by legitimate users. Based on these observations, we propose two features: *LegitTags* and *SpamTags*. *LegitTags* calculates the number of tags a user has posted which are mostly used by legitimate users. However, spammers also have habit to use popular tags that are previously posted by legitimate

users. Therefore, we introduce a feature *LegitTags* which defines the probability that a particular tag is used only by legitimate users. Let U_t be the set of all users in a social tagging system who associated at least one resource with a tag t , T_u be the set of all tags posted by a user u , S_t be a subset of spammers in U_t and L_t be a subset of legitimate users in U_t . Then, the feature *LegitTags* for user u can be calculated as follows:

$$LegitTags_u = \frac{1}{|T_u|} \sum_{t \in T_u} \delta(u, t), \quad (1)$$

where $\delta(u, t)$ returns 1 if $|S_t|/|U_t|$ is less than a predefined threshold Th_{Legit} , otherwise it returns 0. Analogously, a feature *SpamTags* is defined as:

$$SpamTags_u = \frac{1}{|T_u|} \sum_{t \in T_u} \sigma(u, t), \quad (2)$$

where $\sigma(u, t)$ returns 1 if $|L_t|/|U_t|$ is less than a predefined threshold Th_{Spam} , otherwise it returns 0. Optimal threshold values for Th_{Legit} and Th_{Spam} are experimentally found, and for our evaluation dataset they are set to 0.21 and 0.13, respectively.

3.2 Tags Popularity Based Features

One characteristic of spammers is that they tend to use popular tags when annotating online resources to gain higher rank in a search by keyword scenario [12], as already discussed in Sections 1 and 2. Based on this finding, we propose six features which address the popularity of tags shared in a social tagging system, namely, *LegitPopularity*, *SpamPopularity*, *TagPopularity*, *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*.

For a particular tag t , we define a feature *LegitPopularity* as the number of times users in L_t used tag t in their posts. In an analogous way, features *SpamPopularity* and *TagPopularity* represent the number of times tag t was assigned to resources by users in S_t and U_t , respectively.

We propose three additional features representing tags popularity, namely *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*. They represent the number of users in L_t , S_t and U_t who assigned tag t to at least one resource, respectively.

3.3 User Activity Based Features

User activity based features take advantage of user's posting behavior in a social system to better discriminate between legitimate users and spammers. These features are explained in the following and summarized in Table 1. All features are computed for each user separately.

Feature *AverageTagsPerPost* shows the average number of tags a user assigned to different resources. The rationale behind this feature is that posts from legitimate users usually have more tags describing resources compared to

Table 1. Summary of user activity based features. All features are computed for each user separately.

| Distinct feature | Description |
|-----------------------------------|---|
| <i>AverageTagsPerPost</i> | Avg. no. of tags a user assigned to different resources |
| <i>AverageDistinctTagsPerPost</i> | Avg. no. of unique tags a user assigned to different resources |
| <i>NewTags</i> | No. of unprecedented tags a user added to the global dictionary of tags |
| <i>Legit2Spam</i> | Ratio between no. of legitimate and spam tags assigned by a user |
| <i>TagsPerUser</i> | Total no. of tags a user assigned to different resources |
| <i>DistinctTagsPerUser</i> | Total no. of unique tags a user assigned to different resources |
| <i>Posts</i> | No. of posts shared by a user |
| <i>DistinctTagRatio</i> | Ratio between no. of unique tags and total no. of tags assigned by a user |

posts shared by spam users. With the same rationale, we introduce a feature *TagsPerUser*, defined as the total number of tags a user assigned to different resources.

Based on our observation that spammers tend to use different popular tags for different posts and, at the same time, the intersection between sets of tags in two arbitrary posts from one spammer is none or very small, we introduce a feature called *AverageDistinctTagsPerPost*. This feature measures the average number of unique tags a user assigned to different resources. With the same rationale, we present two other features: *DistinctTagsPerUser*, defined as the total number of unique tags a user assigned to different resources, and *DistinctTagRatio*, which represents the ratio between number of unique tags and total number of tags assigned by a user.

Furthermore, number of new tags introduced by spammers to the global dictionary of tags is relatively higher than number of tags introduced by legitimate users. Based on this fact, we introduce a feature *NewTags*. This feature is defined as the number of unprecedented tags a user added to the global dictionary of tags.

We present here two other user activity based features. A feature *Legit2Spam* represents the ratio between number of legitimate and spam tags assigned by a user, while a feature *Posts* is defined as the number of posts shared by a user.

Discussion on the performance of all proposed features on discriminating legitimate users from spammers is presented in Section 5.

4 Evaluation

In this section, we present a dataset and classification metrics used to evaluate the set of proposed features.

Table 2. Statistics of the original dataset (ECML PKDD Discovery Challenge 2008) and a reduced dataset used for evaluation

| Statistics of datasets | Original dataset | | | Evaluation dataset | | |
|------------------------|------------------|------------|------------|--------------------|---------|---------|
| | Legitimate | Spam | Total | Legitimate | Spam | Total |
| No. of users | 2,467 | 29,248 | 31,715 | 500 | 500 | 1000 |
| No. of resources | 401,250 | 2,060,707 | 2,461,957 | 172,452 | 65,378 | 237,830 |
| No. of tags | 816,197 | 13,258,759 | 14,074,956 | 477,794 | 473,544 | 951,338 |
| Avg. posts per user | 162 | 70 | 77 | 344 | 131 | 238 |
| Avg. tags per user | 330 | 453 | 506 | 955 | 947 | 951 |
| Avg. tags per post | 2 | 7 | 6 | 3 | 8 | 4 |

4.1 Dataset

We used dataset collected from BibSonomy. BibSonomy is a social tagging system that allows users to share bookmarks and publication references. The system is aimed for researches and academic institutions which require a system without irrelevant information and commercial content. Therefore, this system has a rigorous policy against spammers. Moderators in this system manually find and remove spammers from the system [3]. If a user is labeled as a spammer, his/her posts will be no longer visible to other users. Spammer posts will not be removed from the system and this fact gives an illusion to spammers that they are still able to pollute the system.

We used a public dataset released by BibSonomy as a part of the ECML PKDD Discovery Challenge 2008 on Spam Detection in Social Bookmarking Systems [6]. Table 2 summarizes statistics of the dataset. This dataset consists of around 32,000 users who are manually labeled either as spammers or legitimate users, user-contributed tags and resources (bookmarks) which can be either web pages or BibTeX files. However, as shown in the second column of Table 2, an important skewness is present in this dataset since a majority of the users are spammers. This means that if a classifier labels all users as spammers, we would achieve a classification accuracy of over 0.92. Therefore, we selected randomly a subset of users (500 legitimate users and 500 spammers) to achieve a balance with respect to the number of users. Statistics of the dataset used for evaluation in this paper is shown in the third column of Table 2.

4.2 Classification Metrics

After having extracted proposed features from the evaluation dataset, several supervised classification methods, such as support vector machine (SVM), AdaBoost and decision trees, were applied on the extracted features to classify users as legitimate or spammers. Given the ground truth and the predicted labels, a confusion matrix is created and the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are computed.

Different metrics are used to evaluate the proposed features. The accuracy of the classification when shown solely is not a good indicator of a classifier behavior, and therefore, we calculated some complementary measures to thoroughly evaluate the proposed features. In addition to the classification accuracy defined as $\frac{TP+TN}{TP+FP+FN+TN}$, we calculated: (1) false positive rate (FPR) as $\frac{FP}{FP+TN}$, (2) precision (P) as $\frac{TP}{TP+FP}$, (3) recall (R) as $\frac{TP}{TP+FN}$, (4) F-measure as $\frac{2 \cdot P \cdot R}{P+R}$, and (5) area under receiver operating characteristics (AUC ROC) which represents the probability that an arbitrary legitimate user is ranked higher than an arbitrary spammer. Finally, we determined Matthews Correlation Coefficient (MCC) [15] to validate our result. As a less known performance metric, we explain it here in more details. MCC is a performance quality measure used in two-class classification problems. It is often used as a performance metric in bioinformatics. MCC is defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (3)$$

MCC has values between -1 and +1, where +1 indicates perfect classification (prediction), -1 shows total disagreement between prediction and observation, and 0 represents a random classification.

5 Discussion

In this section, we discuss the prominence of the proposed features for detection of spammers. First, performance of each feature separately is estimated and then some of them are aggregated to improve the classification performance. Finally, performance of different classifiers are compared and analyzed. All performance criteria were evaluated by making use of classifiers in Weka [18], a software library of most distinguished machine learning algorithms. Evaluation is performed using 10-fold cross-validation and default values for all parameters in Weka.

Figure 2 shows how well each of the proposed 16 features discriminates spammers. A decision stump classifier in Weka is applied on extracted features and the performance of each proposed feature is measured as accuracy, AUC ROC and F-measure. As we can see from the accuracy metric, each feature is able to correctly classify at least 60 % of users. Feature *LegitTags* has the best performance with more than 0.91 of accuracy in classification, and it is followed by *SpamTags*, *DistinctLegitPopularity* and *Legit2Spam* with 0.87, 0.76, 0.73 of accuracy, respectively. For classification of randomly selected users, as it can be seen from AUC ROC, again *LegitTags* and *SpamTags* have the best performance with 0.96 and 0.93 of AUC ROC. F-measure follows the trend of accuracy and AUC ROC, showing that *LegitTags* and *SpamTags* are the adequate features. Having considered all these measures, we can conclude that after *LegitTags* and *SpamTags*, tags popularity based features are the best performing set of features.

Feature *LegitTags* has the ability to very well separate spammers from legitimate users when fed solely into the classifier, as discussed previously in this

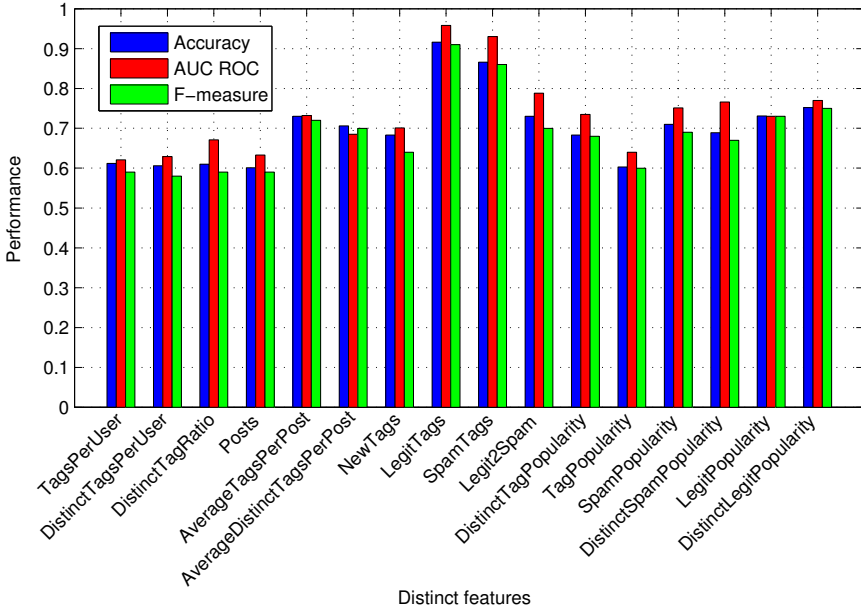
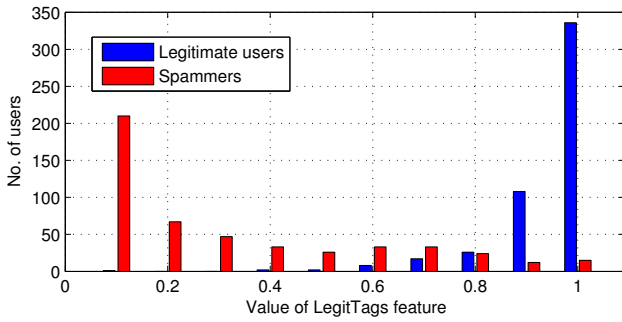


Fig. 2. The performance of each proposed feature plotted as accuracy, AUC ROC and F-measure

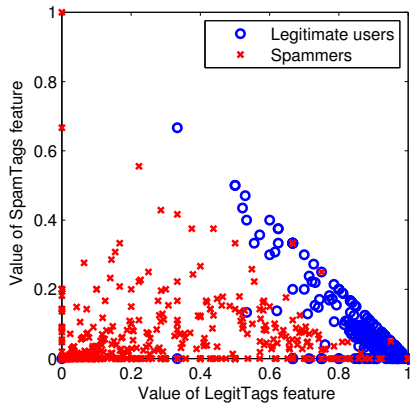
section. Therefore, we explore the performance of this feature in more details. 10-bins histogram of *LegitTags* values calculated from the evaluation dataset is shown in Figure 3 (a). When this feature is combined with the second best performing feature *SpamTags* and feature values are plotted in the feature space, we obtain the distribution shown in Figure 3 (b). These distributions give a visual intuition for how well feature *LegitTags* alone or combined with other feature separates two types of users. We can clearly see that the distributions of legitimate users and spammers can be easily separated by a simple threshold, for case (a), or line, for case (b). Therefore, linear discrimination classifiers are enough for spammers detection when using *LegitTags* and *SpamTags* features.

After *LegitTags* and *SpamTags*, tags popularity based features are the most powerful set of features, as shown in Figure 2. To further evaluate these features, we applied a standard discrimination function, the χ^2 statistics. The χ^2 (chi-square) statistics measures the goodness and powerfulness of features used for classification [9]. Again, we used Weka to apply this discrimination function. Figure 4 shows the consistent ranking of our six tags popularity based features to discriminate spammers from legitimate users.

It is well known that classification accuracy can be significantly improved by aggregating weak features rather than feeding different features separately into a classifier [4]. We can see from Figure 2 that each tag popularity and user activity based features have less than 0.75 and 0.73 of accuracy, respectively. Nevertheless, combination of these features results in a performance improvement.



(a)



(b)

Fig. 3. Discrimination power of the feature *LegitTags* to separate two types of users, when: (a) used alone, (b) combined with the feature *SpamTags*. Figure (a) represents the histogram of *LegitTags* values, and Figure (b) shows projection of *LegitTags* and *SpamTags* values in the feature space attempting to separate legitimate users (blue circles) from spammers (red crosses).

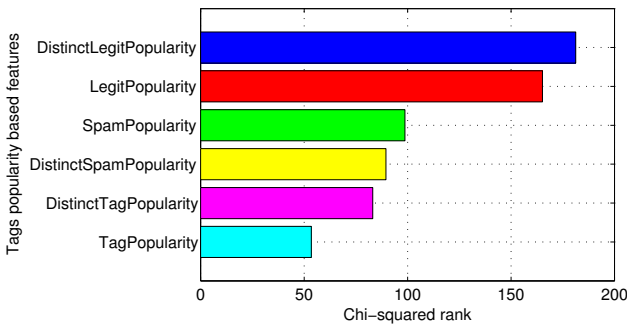


Fig. 4. Chi-squared ranking for all tags popularity based features

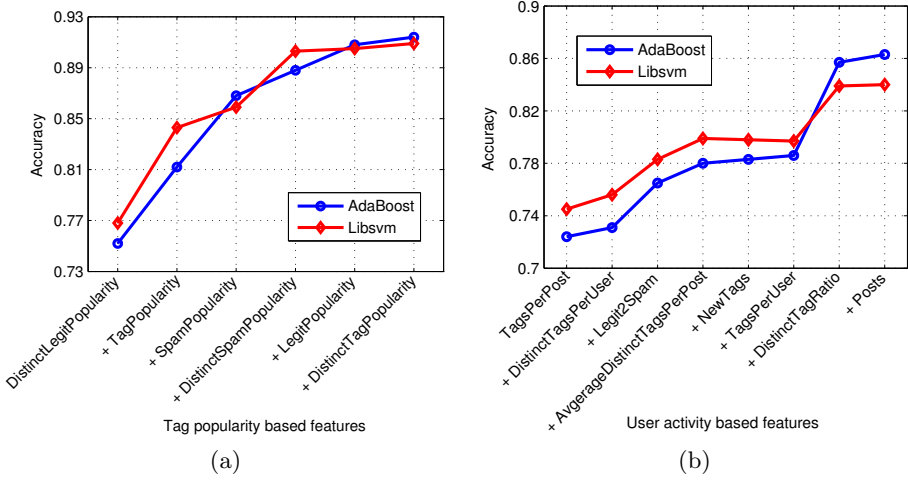


Fig. 5. Enhancement in the classification performance by aggregating: (a) all tag popularity based features, and (b) all user activity based features

Figures 5 (a) and (b) show how classification performance can be improved by separately aggregating tag popularity based features and user activity based features. Results are shown for two classifiers, namely AdaBoost and LibSVM. By combining all tag popularity based features we can improve classification accuracy from 0.75 to 0.91, while aggregating all user activity based features the accuracy increases from initial 0.73 to 0.86.

Finally, the proposed features are fed into more than 40 different classifiers and their performance in classification is evaluated. We used Weka to train classifiers with our features and to measure performance. Diverse classifiers are used, such as decision trees, neural networks and LibSVM, in order to have different perspectives on discriminative functions in feature space. Furthermore, ensemble classifiers [4] such as AdaBoost, bagging and rotation forest, were employed to have a comprehensive evaluation. The top 10 performing classifiers are reported in Table 3. Results show that AdaBoost was the best classifier for the evaluation dataset. It performs well with 0.987 of accuracy and only 0.013 of FPR. LibSVM and rotation forest classifiers have slightly lower accuracy of 0.986 and 0.981, with 0.014 and 0.019 of FPR, respectively. As noted by Markines *et al.* [14], in a deployed social spam detection system it is more important that FPR is kept low compared to high accuracy, because misclassification of a legitimate user is a more consequential mistake than missing a spammer. Other researchers, who proposed different features from the whole or partial dataset of ECML PKDD Discovery Challenge 2008, obtained similar results, for example, Markines *et al.* [14] were able to reach 0.979 of accuracy and 0.013 of FPR, while Bogers *et al.* [3] got 0.9799 of classification accuracy.

Table 3. Top classifiers created in Weka. Evaluation is performed using 10-fold cross-validation. The best performing classifier and metric values are highlighted in **bold**.

| Weka classifier | Accuracy | FPR | R | P | F-measure | AUC ROC | MCC |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| AdaboostM1 | 0.987 | 0.013 | 0.994 | 0.980 | 0.987 | 0.993 | 0.974 |
| Libsvm | 0.986 | 0.014 | 0.978 | 0.994 | 0.986 | 0.993 | 0.973 |
| RotationForest | 0.981 | 0.019 | 0.981 | 0.978 | 0.980 | 0.993 | 0.962 |
| SMO | 0.979 | 0.021 | 0.979 | 0.979 | 0.979 | 0.991 | 0.958 |
| RBFNetwork | 0.975 | 0.025 | 0.965 | 0.986 | 0.975 | 0.993 | 0.95 |
| Bagging | 0.974 | 0.026 | 0.974 | 0.974 | 0.974 | 0.996 | 0.948 |
| Decorate | 0.973 | 0.029 | 0.970 | 0.968 | 0.968 | 0.990 | 0.930 |
| FT | 0.972 | 0.028 | 0.966 | 0.972 | 0.970 | 0.985 | 0.944 |
| MultiBoostAB | 0.971 | 0.029 | 0.970 | 0.972 | 0.971 | 0.987 | 0.942 |
| MLP | 0.971 | 0.029 | 0.959 | 0.984 | 0.971 | 0.982 | 0.942 |

6 Conclusions

In this paper, we presented different features suitable for fighting spam in social tagging systems. The problem of having trustworthy tags associated to resources is important in social systems, because of their increasing popularity as means of sharing interests and information. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers.

We proposed 16 distinct features based on user activity in posting and tags popularity. The prominence of the proposed features in distinguishing between legitimate users and spammers is discussed. We measured the performance of each feature solely and showed that *LegitTags* feature, defined as the probability that a particular tag is used only by legitimate users, performed the best. We also showed that aggregation of features leads to the improvement in the classification performance. Finally, performance of different classifiers was compared. The results are promising. The best classifier achieved accuracy of 0.987 with false positive rate of 0.013 in discriminating legitimate users from spammers.

As a future study, we will explore more sophisticated features which are able to deal with dynamics of trust, by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling of phenomenon in real-world applications.

Acknowledgment. This work was supported by the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2).

References

1. Xirong, L., Snoek, C., Worring, M.: Learning tag relevance by neighbor voting for social image retrieval. In: Proc. ACM MIR, pp. 180–187 (2008)
2. Benz, D.K., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The social bookmark and publication management system BibSonomy. VLDB Journal 19(6), 849–875 (2010)
3. Bogers, T., Van den Bosch, A.: Using Language Models for Spam Detection in Social Bookmarking. In: Proc. ECML/PKDD Discovery Challenge, pp. 1–12 (2008)
4. Duda, R., Hart, P.: Pattern classification and scene analysis. Wiley (1973)
5. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing 11(6), 36–45 (2007)
6. Hotho, A., Benz, D., Jäschke, R., Krause, B.: ECML PKDD Discovery Challenge (2008), <http://www.kde.cs.uni-kassel.de/ws/rsdc08>
7. Ivanov, I., Vajda, P., Jong-Seok, L., Goldmann, L., Ebrahimi, T.: Geotag propagation in social networks based on user trust model. MTAP 56(1), 155–177 (2012)
8. Ivanov, I., Vajda, P., Jong-Seok, L., Ebrahimi, T.: In tags we trust: Trust modeling in social tagging of multimedia content. IEEE SPM 29(2), 98–107 (2012)
9. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proc. ICTAI, pp. 338–391 (1995)
10. Rogers, I.: The Google PageRank algorithm and how it works (2002)
11. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag Recommendations in Folksonomies. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenić, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 506–514. Springer, Heidelberg (2007)
12. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: Proc. ACM AIRWeb, pp. 57–64 (2007)
13. Krause, B., Schmitz, C., Hotho, A., Stumme, G.: The anti-social tagger: Detecting spam in social bookmarking systems. In: Proc. ACM AIRWeb, pp. 61–68 (2008)
14. Markines, B., Cattuto, C., Menczer, F.: Social spam detection. In: Proc. ACM AIRWeb, pp. 41–48 (2009)
15. Matthews, B.W.: Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta 405(2), 442–451 (1975)
16. Von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: CAPTCHA: Using Hard AI Problems for Security. In: Biham, E. (ed.) EUROCRYPT 2003. LNCS, vol. 2656, pp. 294–311. Springer, Heidelberg (2003)
17. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-based character recognition via web security measures. Science 321(5895), 1465–1468 (2008)
18. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005), <http://www.cs.waikato.ac.nz/ml/weka>
19. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: Proc. ACM WWW, pp. 1–8 (2006)
20. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with TrustRank. In: Proc. VLDB, pp. 576–587 (2004)

The Multidimensional Study of Viral Campaigns as Branching Processes

Jarosław Jankowski¹, Radosław Michalski², and Przemysław Kazienko²

¹ Faculty of Computer Science, West Pomeranian University of Technology,
Szczecin, Poland

`jjankowski@wi.zut.edu.pl`

² Institute of Informatics, Wrocław University of Technology,
Wrocław, Poland

`{radoslaw.michalski,kazienko}@pwr.wroc.pl`

Abstract. Viral campaigns on the Internet may follow variety of models, depending on the content, incentives, personal attitudes of sender and recipient to the content and other factors. Due to the fact that the knowledge of the campaign specifics is essential for the campaign managers, researchers are constantly evaluating models and real-world data. The goal of this article is to present the new knowledge obtained from studying two viral campaigns that took place in a virtual world which followed the branching process. The results show that it is possible to reduce the time needed to estimate the model parameters of the campaign and, moreover, some important aspects of time-generations relationship are presented.

Keywords: viral campaigns, diffusion of information, branching process, social network analysis, virtual worlds.

1 Introduction

There are variety of models to describe the diffusion of information: starting from epidemic models [24], [2], through the Bass model [5], adoption model [8], [29] and variety of models combining network and user properties [14], [23], [7], [3]. Another model, the super-diffusion [26], [27], [31], may be called the fastest diffusion across all the previously mentioned. A good comparison of information diffusion models may be found in [28] and more recent research regarding new models is presented in [16].

Yet another model, the branching process [21] is since recently used for analysing diffusion of information extending the ability to describe how the information may flow through the social network [22], [25], [17], [18]. Even some suggestions may be found that the branching model is more adequate to describe the information diffusion, especially while compared to the disease spreading models [18]. However, it is not a question of preference while fitting real-world data to a particular model – the goodness of fit decides which model a particular viral campaign follows. The outcomes of having a model and its parameters discovered are that one may benefit from

this general model properties and implications which make the analysis of the campaign easier. Generally, it is preferred to know which model a particular campaign follows while it happens, because in that case it would be easier to predict further behaviour of the information diffusion process. However, in that case, the branching process introduces some limitations described in Section 2.

This paper focuses on analysing two campaigns conducted in virtual world environment which followed the branching process model to present valuable outcomes in terms of reducing the need for having all the campaign data to calculate branching process parameters. To extend this result in terms of time aspects, authors decided to deepen the analysis towards the relationship between the time and generations in this model which led to better knowledge about how particular branches of the model are being developed in time, what may even allow to predict on-going campaigns results.

The structure of this paper is as follows. The next section of this paper describes the problem, which is followed by the related work analysis. Section 4 presents the experiment setup and the description of the analysed campaigns. Experimental studies results are presented in Section 5 with the conclusions and future work directions presented in Section 6.

2 Problem Description

The branching process in terms of information diffusion may be basically described as a process where an individual may spread the information to a number of consequents. Starting from a number of seeds understood as the first generation the information is forwarded towards next generations, creating a tree of information traversal. The information diffusion ends when there will be no further infections, that means reaching the all of the susceptible users. The nature of the branching process, especially the fact that it is not based on time but on generations, makes the whole process a bit harder to interpret on a time basis, because the number of users infected in a particular generation changes over time. And as the basic equations of the branching process are calculating the number of infected in the next generation basing on the previous one, the chance to estimate the parameters while the campaign is on-going is very weak, unless there exists a certainty that the number of infected users in previous generations would not change, which is rather unlikely in real campaigns. That results in constant underestimation of the overall number of infected users.

Basing on the above limitations of the branching model, authors of this paper decided to examine whether is there any chance to estimate the total number of infected users in the viral campaign while using only a partial information of all infections. And in that case the partial information means the complete information about only part of consecutive generations starting from the first one. So, to describe the problem in more formal way, the task is to estimate the p , N and λ parameters for the model [22] by using real-world data in the way that only the complete information about particular number of generations is available beginning with the first generation.

The question arises why would one benefit from such an approach if it means that the analysis would be performed *a posteriori*? To name only the one major argument, the task to estimate the campaign parameters is to find a model which fits all the branching process generations as good as it is possible. So, in that case, it is necessary to find in a three-dimensional space a set of parameters which minimize the error of fit across all generations. If the approach proposed by the authors succeeds, the calculation time needed for estimating the parameters will reduce still providing good estimation. However, as it is described in Section 4, authors decided also to evaluate how are particular generations changing over time. And if it will be seen that all the generations needed for the model parameters estimation stabilize before the campaign ends, it will lead to the conclusion that the parameter estimation may be performed earlier as well.

So, despite these limitations, why is the branching process becoming more popular in modelling the diffusion of information? As described in [22], most of the models base on aggregated information about the total number of infections. In that case the advantage of the branching process is that its approach allows to analyse the epidemics on different level focused on individual reproduction rate, what may lead to extending the knowledge about the diffusion of information.

3 Related Work

The concept of fitting real life data to a particular model plays an important role in statistics. Goodness of fit tests are used to fit variety of data to the existing models, and social network analysis also often reaches for those tests. For instance, real life social networks are to be fitted to models [15], variety of analyses are performed to check how particular social network properties fit power law distribution [11], not mentioning on new model generation and fitting [30]. The benefit of having a particular process fit to the model is that in that case it is easier to generalize the observed data and predict the future behaviour of the information diffusion.

As it was stated in the previous Section, the experiments conducted by authors of this paper are regarding finding the optimal set of parameters for modelling the branching process in terms of limited knowledge about the branching process. In that case authors wanted to find the answer on the question whether is it possible to estimate the parameters of the model by having only the information about a few generations. Most of the literature regarding estimation of the branching processes is related to supercritical processes [12], [13], however those approaches were more related to obtaining the distribution of offspring probability. In [22] authors analysed the possibility of use the branching process to model the diffusion of information and decided to estimate the model parameters in discrete time what differs from the approach presented in this paper. The other type of analysis authors of this paper perform is the study of the relationship between generations and time. An interesting case of continuous time branching processes was described in [20], where next generations were strongly related to the time due to biological reasons. Another work on the problem of

time-generation relationship in the branching process with regards to epidemics which also states that if the transitions are population dependent, the long-term prediction of these processes is an open problem is [19].

However, authors of this paper were unable to find similar to the presented approach in terms of modelling viral campaign by using consecutive generations and a single set of parameters, the deepened analysis of time-generations relationship was also not studied extensively by others.

4 Experimental Setup

Authors of this paper analysed two real-world campaigns from a virtual world environment with the goal to estimate the branching process parameters. However, the basic intuition in the branching process is that in every generation the model parameters will change due to increasing or decreasing interest in the campaign what may harden the task of predicting the final spread of the campaign. Authors decided to omit this problem by trying to estimate only a single set of parameters for the whole campaign.

The experiment setup was as follows: for the whole dataset number of infections in every generation was calculated. Next, starting from only the first generation branching model parameters were calculated and evaluated with the real data in terms of MSE errors of overall campaign reach and the cumulative MSE error calculated as the difference of real data reach and estimated reach for every generation. Next, this procedure was repeated for a model parameters estimated by using two consecutive generations starting from the first one, three generations and so on until the set of generations was equal to the number of generations in real-data. This procedure allowed to analyse how well the model built by using less number of generations is able to estimate the overall campaign reach and behaviour.

As it was already described in the Section 2, due to the changing number of infections for each generation in time, the proposed approach still requires to have the whole dataset to be applied. However, if the approach succeeds only a limited number of generations would be required to calculate the model parameters what decreases the overall processing time. But to get additional knowledge about the overall behaviour of the branching process, authors decided to extend the study by analysing how particular generations change over time. In that case if the experiment results will prove that the number of generations required to adequately model the data stabilize before the campaign ends, the additional outcome may be the ability to predict the campaign behaviour at earlier stage (on-going). So the second part of the research was devoted to analyse the time-generations relationship.

During the research, there were data from two viral actions with different characteristics from social platforms working in a form of virtual world. In both actions, users were spreading virtual goods like avatars using viral mechanism to their friends. The first viral action denoted as V_1 was based on sending gifts to friends and the senders' motivation to spread those gifts was not incentivized. The second action,

denoted as V_2 , was based on incentives and competition among users to spread visual elements of avatars among their friends.

The analysis of number of infections in time gives the knowledge about the dynamics of the campaign, however it is not delivering information about the structure of infections. Both campaigns had different specifics and by using aggregated models based on time dimension only, there is an additional analysis to show structures of infections needed. Next, both campaigns were analysed by using the proposed approach based on generations and parameters describing their characteristics. A generation is defined in terms of viral marketing as a number of transmissions required to reach a member along a chain of communication initiated by a single seed [4]. The approach based on generations can capture structures of transmissions which is not possible by the cumulative analysis based only on infections over the time. In the earlier research, three main parameters from epidemic theory were used for modelling the characteristics of viruses spreading which can be applicable to viral marketing campaigns as well [21]. Contagion parameter denoted as p describes the probability of transferring viral message by an infective. Epidemic intensity λ represents the number of customers reached. Epidemic threshold parameter ETP defined as $p*\lambda$ describes the progression of epidemics. Becker defined relation between characteristics of campaign and epidemic threshold parameter as sub-critical ($ETP < 1$), super-critical ($ETP > 1$) and critical ($ETP = 1$) [6]. D. B. Stewart et al. presented conceptual framework for viral marketing [25]. The mathematical model presented by the authors for viral campaigns modelling was based on deterministic model discussed by J.C. Frauenthal [9], used by R.M. Anderson and R.M. May for modelling epidemic [1], extended later by G. Fulford et al. [10].

5 Results

The empirical research was conducted by using the above described approach. The main goal was to conduct extended analysis of viral campaign using the approach based on branching processes and verify the ability to predict the viral campaign model by analysing two real datasets. In Figure 1 and Figure 2 the cumulative number of infections in the analysed time period (days) for both campaigns is presented.

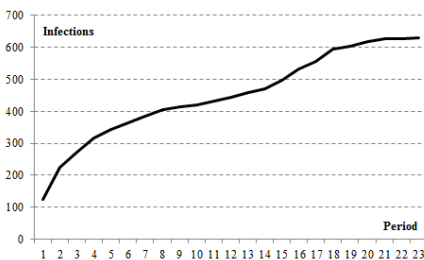


Fig. 1. Cumulative number of infections - V_1

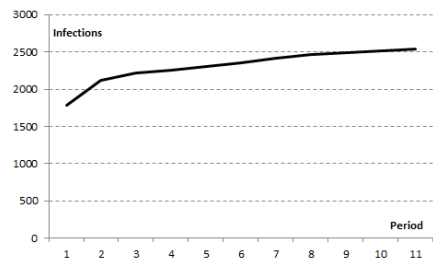


Fig. 2. Cumulative number of infections - V_2

Following the earlier approaches, the analyses of two viral campaigns were performed and p , λ and ETP parameters were computed and showed in Table 1 for both of campaigns.

Table 1. Number of infections and parameters of analysed campaigns

| | <i>G</i> | <i>Infected</i> | <i>Cumulative</i> | <i>Decisions</i> | <i>Infections sent</i> | <i>p</i> | λ | <i>ETP</i> |
|-------|----------|-----------------|-------------------|------------------|------------------------|----------|-----------|------------|
| V_1 | 1 | 1 | 1 | 1 | 11 | 1.0000 | 11.0000 | 11.0000 |
| | 2 | 11 | 12 | 10 | 49 | 0.9091 | 4.9000 | 4.4545 |
| | 3 | 49 | 61 | 26 | 106 | 0.5306 | 4.0769 | 2.1633 |
| | 4 | 106 | 167 | 42 | 123 | 0.3962 | 2.9286 | 1.1604 |
| | 5 | 123 | 290 | 41 | 90 | 0.3333 | 2.1951 | 0.7317 |
| | 6 | 90 | 380 | 33 | 79 | 0.3667 | 2.3939 | 0.8778 |
| | 7 | 79 | 459 | 20 | 41 | 0.2532 | 2.0500 | 0.5190 |
| | 8 | 41 | 500 | 11 | 43 | 0.2683 | 3.9091 | 1.0488 |
| | 9 | 43 | 543 | 12 | 40 | 0.2791 | 3.3333 | 0.9302 |
| | 10 | 40 | 583 | 14 | 38 | 0.3500 | 2.7143 | 0.9500 |
| | 11 | 38 | 621 | 7 | 13 | 0.1842 | 1.8571 | 0.3421 |
| | 12 | 13 | 634 | 3 | 4 | 0.2308 | 1.3333 | 0.3077 |
| | 13 | 4 | 638 | 1 | 1 | 0.2500 | 1.0000 | 0.2500 |
| | 14 | 1 | 639 | 0 | 0 | 0.0000 | 0.0000 | 0.0000 |
| V_2 | 1 | 9 | 9 | 8 | 187 | 0.8889 | 23.3750 | 20.7778 |
| | 2 | 187 | 196 | 52 | 552 | 0.2781 | 10.6154 | 2.9519 |
| | 3 | 552 | 748 | 115 | 782 | 0.2083 | 6.8000 | 1.4167 |
| | 4 | 782 | 1530 | 105 | 450 | 0.1343 | 4.2857 | 0.5754 |
| | 5 | 450 | 1980 | 55 | 251 | 0.1222 | 4.5636 | 0.5578 |
| | 6 | 251 | 2231 | 32 | 137 | 0.1275 | 4.2813 | 0.5458 |
| | 7 | 137 | 2368 | 18 | 47 | 0.1314 | 2.6111 | 0.3431 |
| | 8 | 47 | 2415 | 5 | 27 | 0.1064 | 5.4000 | 0.5745 |
| | 9 | 27 | 2442 | 4 | 51 | 0.1481 | 12.7500 | 1.8889 |
| | 10 | 51 | 2493 | 4 | 6 | 0.0784 | 1.5000 | 0.1176 |
| | 11 | 6 | 2499 | 3 | 4 | 0.5000 | 1.3333 | 0.6667 |
| | 12 | 4 | 2503 | 2 | 0 | 0.5000 | 0.0000 | 0.0000 |

As the above table shows, the epidemic parameters are changing over the time and describing campaign with a single set of parameters may be a challenging task. Due to these changes, it is difficult to predict the next stages of the campaign by using data from earlier periods. In the analysed campaign V_1 ETP for the first generation was equal 11 while in the second generation it was only 40.49% of earlier value being reduced to 4.4545. After few generations of decreasing, it went up to 1.0488 at generation number eight. For campaign V_2 , even bigger changes were identified (possibly because of the incentives), especially between the first generation with $ETP=20.78$ reduced to 2.95 in the second generation. In the next stage of research, the analysis of change of the parameters was performed for both of campaigns. Campaign V_1 was identified as super-critical during generations G_1, G_2, G_3, G_4 and G_8 while V_2 can be treated as super-critical for generations G_1, G_2, G_3 and G_9 . An interesting result is that characteristics of the second campaign show that incentives were not effective to

increase number of generations characterized as super-critical according to *ETP*, so the campaign reach in terms of number of generations was similar. The other problem identified during the analysis is related to changes of data for each generation with time periods, as it was described in Section 2. If the analysis is performed in the first period, after the second period the results from earlier analysis are useless in terms of prediction, because of changes in all existing generations. Computations must be performed on the whole data and cannot be based on incremental approach. Changes in the number of infections for campaign V_1 are showed in the Table 2 for the first ten days of campaign. In the first period of time (the first day of the campaign), 19,56% of all infections where registered within ten generations. During the second day no additional infections occurred for G_1 and G_2 but additional infections are found for generations G_3 - G_{10} . At this period, the first two infections are registered for generation G_{11} . Generations G_{12} and G_{13} are not built until period P_5 and P_6 .

Table 2. Changes of number of infections in generations over the time

| <i>G</i> | <i>Campaign period (in days)</i> | | | | | | | | | |
|----------|----------------------------------|--------|-------|-------|-------|-------|-------|-------|-------|-------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 29 | 6 | 5 | 1 | 0 | 2 | 3 | 0 | 1 | 0 |
| 4 | 40 | 25 | 13 | 8 | 0 | 0 | 3 | 0 | 2 | 0 |
| 5 | 22 | 31 | 12 | 7 | 3 | 1 | 1 | 3 | 3 | 2 |
| 6 | 14 | 7 | 11 | 3 | 5 | 7 | 2 | 0 | 0 | 1 |
| 7 | 3 | 12 | 2 | 0 | 3 | 4 | 0 | 4 | 3 | 3 |
| 8 | 2 | 7 | 0 | 0 | 4 | 0 | 1 | 3 | 1 | 0 |
| 9 | 2 | 6 | 3 | 5 | 4 | 2 | 5 | 2 | 0 | 0 |
| 10 | 2 | 4 | 1 | 4 | 5 | 2 | 3 | 3 | 1 | 0 |
| 11 | 0 | 2 | 1 | 15 | 3 | 0 | 1 | 3 | 0 | 0 |
| 12 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 19.56% | 15.65% | 7.51% | 6.73% | 4.38% | 3.13% | 3.13% | 3.13% | 1.72% | 0.94% |

The cumulative number of infections for chosen generations is presented in the Figure 3. It shows that the dynamics of increase of infections in generations is changing over the time. In the example presented, for generation G_4 during the first four periods the increase of number of infections was observed and during the next period it stabilizes. The situation changes slightly at period 16 when the next increase is observed. For earlier generations situation is more stabilized. For generations G_5 and G_6 , growth is observed until the 21st period, but highest changes were identified in periods 1-5.

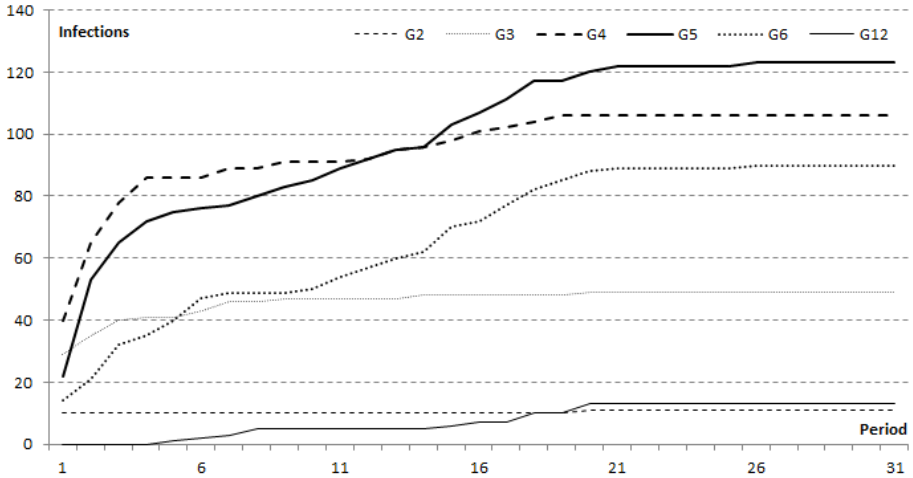


Fig. 3. Cumulative number of infections for selected generations in campaign V_1

In the next step, for both campaigns the dynamics of new generations creation was analysed and the results are presented in Figure 4. The figure presents the moment of the first occurrence of the particular generation. Even though the whole data set for campaign V_1 is based on 31 days and V_2 on 11 days, the results showed that during the first two hours of campaign ten generations were created for campaign V_1 and seven for campaign V_2 . Dynamics of creation of new generations was higher for campaign V_2 with incentives and during first twenty five minutes six generations were created while for first campaigns six generations were created during sixty minutes. It shows that campaigns go very fast in depth (vertically) and after that they start to grow horizontally.

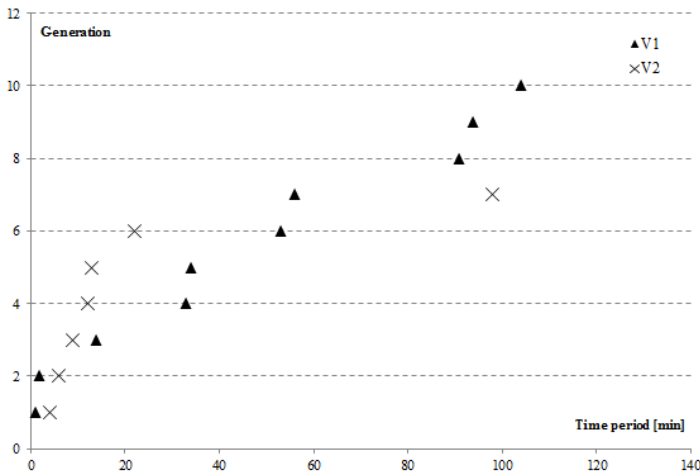


Fig. 4. The dynamics of generations' creation for campaigns V_1 and V_2

The presented analysis show how the generations-based approach can be used to analyse characteristics of viral campaign. Volatile data and change over the time make it difficult to build a prediction model using data from past periods due to changes in generations. The results presented show that estimation of parameters for viral campaigns based on generations approach using contagion parameter and epidemic intensity needs computations with every time period after data increase in all generations. Apart from this it is difficult to estimate parameters because of changes at all generation. The most convenient way would be describing campaign with a single set of p and λ parameters for the whole campaign. It would be easy under the assumption that those parameters are stable during whole campaign. In the next stage experimental results of the proposed method are presented for searching the best fitting model and estimating the campaign reach over consecutive generations. The statistics for the campaigns were used as a reference and the main goal was to estimate campaign reach and to generate a model describing campaign performance with the minimal possible set of generations. For each stage of the campaign, starting from the first generation a branching model is built and parameters are adjusted to find the best fit. Estimations were computed for both campaigns and all generation sets, starting from a set consisting of one generation towards the set combined of all generations, which, in fact, was the reference model. Results for campaign V_1 are showed in Figure 5.

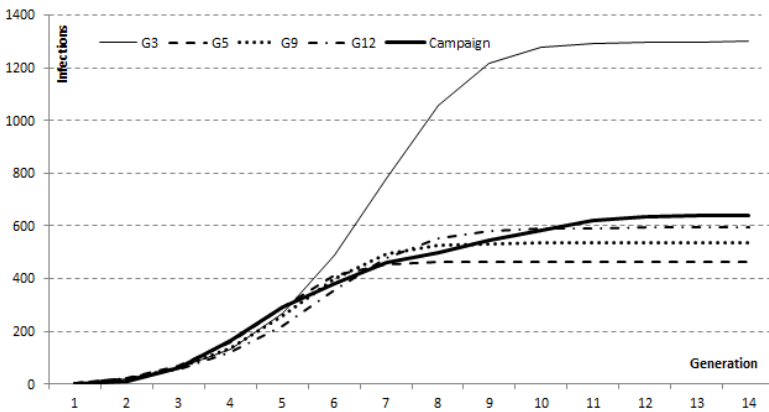


Fig. 5. Estimations for campaign V_1

The estimation computed at generation three resulted in estimation of campaign reach at level of 1299.10 while real campaign reach was 639. Computations performed after the fifth generation resulted in an estimated reach at level 461.97 with 27.70% reach error comparing to the full campaign dataset. Estimation after generation G_9 resulted in 16.35% reach error and after G_{12} - 7.23%. The results for the campaign showed that at the sixth generation G_6 , it was possible to estimate the model acceptable in terms of minimizing the MSE over all generations. As it may be seen, the model parameters gave the opportunity to predict the reach of the campaign accurately. Similarly, the quality of estimation was computed by using mean square error and there was a model selected giving the lowest error for all generations evaluated

for campaign V_2 . Computations after third generation predicted a total of 3867 reach, for the sixth predicted the reach was 2000 and for ninth while the observed campaign reach was 2536. In Table 3, the detailed errors computed for estimated model parameters are showed and the differences between reach predicted from model and for campaign V_1 .

Table 3. Mean square errors and reach errors for campaign V_1

| G | Period MSE | Campaign MSE | Estimated reach | Reach error | Reach error [%] |
|-----|--------------|-----------------|-----------------|---------------|-----------------|
| 1 | 0.00 | 207295.54 | 5.15 | 633.85 | 99.19% |
| 2 | 0.16 | 140916.39 | 142.38 | 496.62 | 77.72% |
| 3 | 26.68 | 223639.37 | 1299.10 | 660.10 | 103.30% |
| 4 | 47.97 | 146480.08 | 1103.70 | 464.70 | 72.72% |
| 5 | 56.66 | 10083.63 | 461.97 | 177.03 | 27.70% |
| 6 | 142.19 | 10798.72 | 456.67 | 182.33 | 28.53% |
| 7 | 153.49 | 10798.72 | 456.67 | 182.33 | 28.53% |
| 8 | 1103.10 | 8900.25 | 475.36 | 163.64 | 25.61% |
| 9 | 454.36 | 3250.94 | 534.50 | 104.50 | 16.35% |
| 10 | 647.85 | 3250.94 | 534.50 | 104.50 | 16.35% |
| 11 | 1198.52 | 1363.36 | 592.79 | 46.21 | 7.23% |
| 12 | 1241.85 | 1363.36 | 592.79 | 46.21 | 7.23% |
| 13 | 1303.94 | 1363.36 | 592.79 | 46.21 | 7.23% |
| 14 | 1353.71 | 1353.71 | 604.17 | 34.83 | 5.45% |

The results in the Period MSE column show errors computed when comparing model chart for selected number of generations with real data. The column Campaign MSE shows error after comparing the model for all generations with the campaign data. Results for campaign V_1 show that it is possible to build the model of the campaign with acceptable reach error by using only data from five generations (35% of all generations). The reach error decrease for both of the campaigns is illustrated in Figure 6.



Fig. 6. The relationship of number of generations used for estimation and the reach error

6 Conclusions and Future Work

The presented research showed multidimensional approach to viral campaign analysis by using branching processes as a model and view on campaigns based on generations. By comparing to the time-based analysis, it is possible to catch some deepened characteristics and interesting results. Recently some research was focused on applications of branching processes in viral marketing but it was mainly focused on building models of campaigns on whole datasets and changes within generations were not discussed.

The research focused on technical analysis of the distribution of media without taking into account the social aspects. This approach was targeted to separate the components of generalized characteristic viral campaign from the social factors that may be unique to the analysed environment. This approach, however, does not exclude ability to use social network characteristics and take into account the distribution network and the attributes characterizing the participants in the campaign.

This analysis based on branching processes approach delivered information about different specifics of both campaigns. Generations give the possibility to analyse structure of infections and make it possible to observe dynamics on each level. Campaigns performance can be compared and the results obtained can be used to evaluate effectiveness and detect drop or increase in the campaign dynamics. It showed different dynamics of changes at generation level for both campaigns.

Apart from extended data analysis presented in this research, the method of building branching model with parameters based on best fit model not using data from all stages of campaign was also presented. The presented approach makes it possible to predict campaign reach without analysing campaign parameters for each generation. Results based on real campaigns showed that it was possible to estimate the campaign reach after fifth generation while whole dataset had fourteen generations. This approach is useful for situations when changes in the parameters make it difficult to describe the whole campaign with only two parameters which are stable for all stages like contagion and epidemic intensity. In terms of studying campaigns which are ongoing, the results from the proposed method are dependent on stabilization of infections in generations used for computations and the proposed approach may be used as a predicting one only after required generations will stabilize. That means generations should be included in the set used for the estimation of parameter when no dynamic growth is observed in number of infections. Apart from selecting the best fit model, the proposed approach can be used to build the knowledge base on campaign structures and instead of comparing the campaign to the model real campaigns data can be used as well.

Despite extending the scientific knowledge in the topic, this sort of knowledge may be found out as valuable for practitioners. Especially the time-generation analysis results show that at the very beginning we are able to see how deep the campaign will go – in that case the early knowledge about the possible reach of the campaign may give the campaign managers additional time to start the campaigns in different social network areas as well.

The presented research opens new research questions which may be explored further. For performance and quality of predictions a method can be developed to predict changes in generations in the next periods to detect a moment of time when stabilization is expected to include generation in the set for computations. Factors and characteristics of campaign affecting community exploration with generations and relations between the number of infections in each generation were also found to be an interesting area for the next research, as well as research based on simulation data which shows for what kind of networks' and parameters' models the proposed approach is suitable.

Acknowledgments. This work was partially supported by fellowship co-financed by the European Union within the European Social Fund, the Polish Ministry of Science and Higher Education, the research project 2010-13.

References

1. Anderson, R.M., May, R.M.: The logic of vaccination. *New Scientist*, 410–415 (1982)
2. Anderson, R.M., May, R.M.: *Infectious Diseases of Humans: Dynamics and Control*, 1st edn. Oxford Science Publications. Oxford University Press (1992)
3. Ba, S., Pavlou, P.: Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior, *Social Science Research Network Working Paper Series* (2006)
4. Bampo, M., Ewing, M.T., Mather, D.R., Stewart, D.B., Wallace, M.: The effects of the social structure of digital networks on viral marketing performance (2008)
5. Bass, F.M.: A new product growth model for consumer durables. *Management Science* 15, 215–227 (1969)
6. Becker, N.G.: *Analysis of Infectious Disease Data*. Chapman & Hall, London (1989)
7. Bolton, G.E., Katok, E., Ockenfels, A.: How effective are electronic reputation mechanisms? an experimental investigation. *Manage. Sci.* 50(11), 1587–1602 (2004)
8. Centola, D., Macy, M.W.: Complex contagion and the weakness of long ties. *American Journal of Sociology* 113(3), 702–734 (2007)
9. Frauenthal, J.C.: *Mathematical Modelling in Epidemiology*. Springer, New York (1980)
10. Fulford, G., Forrester, P., Jones, A.: *Modelling with Differential and Difference Equations*. Cambridge University Press, Cambridge (1997)
11. Guo, L., Tan, E., Chen, S., et al.: Analyzing Patterns of User Content Generation in Online Social Networks. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 369–377 (2009)
12. Heyde, C.C.: Remarks on Efficiency in Estimation for Branching Processes. *Biometrika* 62(1), 49–55 (1975)
13. Heyde, C.C.: On Estimating the Variance of the Offspring Distribution in a Simple Branching Process. In: Maller, R., Basawa, I., Hall, P., et al. (eds.) *Selected Works of C.C. Heyde*, pp. 276–288. Springer, New York (2010)
14. Holme, P., Newman, M.E.J.: Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* 74(5) (2006)
15. Hunter, D.R., Goodreau, S.M., Handcock, M.S.: Goodness of Fit of Social Network Models. *Journal of the American Statistical Association* 103, 248–258 (2008)

16. Iribarren, J.L., Moro, E.: Affinity Paths and Information Diffusion in Social Networks. *Social Networks* 33, 134–142 (2011)
17. Iribarren, J., Moro, E.: Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Phys. Rev. Lett.* 103, 038702 (2009)
18. Iribarren, J.L., Moro, E.: Branching Dynamics of Viral Information Spreading. *Phys. Rev. E.*, 84, 046116 (2011)
19. Jacob, C.: Branching Processes: Their Role in Epidemiology. *International Journal of Environmental Research and Public Health* 7, 1186–1204 (2010)
20. Klein, B., Macdonald, P.D.M.: The Multitype Continuous-Time Markov Branching Process in a Periodic Environment. *Advances in Applied Probability* 12(1), 81–93 (1980)
21. Kolmogorov, A.N., Dmitriev, N.A.: Branching stochastic processes, *Doklady Akad. Nauk U.S.S.R.* 56, 5–8 (1947)
22. van der Lans, R., van Bruggen, G., Eliashberg, J., Wierenga, B.: A viral branching model for predicting the spread of electronic word of mouth. *Marketing Science* 29(2), 348–365 (2010)
23. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. *ACM Trans. Web* 1(1) (2007)
24. Norman, B.: *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London (1975)
25. Stewart, D.B., Ewing, M.T., Mather, D.R.: A Conceptual Framework for Viral Marketing. In: *ANZMAC 2009* (2009)
26. Tadic, B., Thurner, S.: Information Super-Diffusion on Structured Networks. *Physica A* 332, 566–584 (2004)
27. Tsallis, C., Bukman, D.: Anomalous Diffusion in the Presence of External Forces: Exact Time-Dependent Solutions and their Thermostatistical Basis. *Phys. Rev. E.* 54, R2197–R2200 (1996)
28. Valente, T.: *Network models of the diffusion of innovations (quantitative methods in communication subseries)*. Hampton Press, NJ (1995)
29. Wu, F., Huberman, B.A.: *Social structure and opinion formation* (2004)
30. Yang, J., Leskovec, J.: Modeling Information Diffusion in Implicit Networks. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 599–608 (2010)
31. Zekri, N., Clerc, J.: Statistical and Dynamical Study of Disease Propagation in a Small World Network. *Phys. Rev. E.* 64, 56115 (2001)

Models of Social Groups in Blogosphere Based on Information about Comment Addressees and Sentiments

Bogdan Gliwa, Jarosław Koźlak, Anna Zygmunt, and Krzysztof Cetnarowicz

AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
`{bgliwa,kozlak,azygmunt,cetnar}@agh.edu.pl`

Abstract. This work concerns the analysis of number, sizes and other characteristics of groups identified in the blogosphere using a set of models identifying social relations. These models differ regarding identification of social relations, influenced by methods of classifying the addressee of the comments (they are either the post author or the author of a comment on which this comment is directly addressing) and by a sentiment calculated for comments considering the statistics of words present and connotation. The state of a selected blog portal was analyzed in sequential, partly overlapping time intervals. Groups in each interval were identified using a version of the CPM algorithm, on the basis of them, stable groups, existing for at least a minimal assumed duration of time, were identified.

Keywords: social network analysis, groups, blogosphere, sentiment.

1 Introduction

An important problem in the analysis of social media is to identify the real relations between users in the best possible way, which allows us to identify groups that best reflect reality considering majority of existing significant interactions between entities and their emotional (sentimental) characteristics.

Nowadays, blogs play a significant role in the exchange of information on different subjects and the forming of opinions. A very important element of blogs is the possibility of adding comments, which facilitate discussions. Comments may be written in relation to posts or other comments and may have a different content and emotional attitude. Blogosphere is very dynamic, thus the relationships between bloggers are very dynamic and temporal: the lifetime of posts is very short.

In the research on blogosphere, different interactions between users are used for constructing models for analysis. This paper concerns the analysis of number, sizes and other characteristics of groups identified in the blogosphere using a set of models identifying social relations. These models differ regarding the method of classifying the addressee of the comments (they are either the post author

or the author of a comment on which this comment is directly addressing) and a sentiment calculated for comments considering the statistics of words present and connotation.

Taking into consideration the sentiment while analysing groups allows us to identify groups built by interactions having different degrees of positive, neutral or negative sentiment. Qualification of differences between such groups may be important not only for sociological research, but also for identification of kind of influence and its consequences, applied for example to choice of advantageous marketing politics or identification of influential users who spread verbal violence and hatred.

2 Research Domain Overview

2.1 Models of Blogosphere

The research concerning the analysis of blogosphere, produced constructions of different models of parts of blogosphere. One can observe that the character of these models is strictly dependent on the kinds of analysis for which they are created, e.g. identification of key users and groups.

For such applications, it is possible to distinguish universal models, which embrace both the representation of the character of given nodes and the links between them, the models focusing on the classification of nodes without considering the strength of the links between given pairs of nodes and models focusing mostly on neighborhoods of nodes and not taking the characteristic features of individual nodes into consideration.

In [1] several graph structures related to blogs are distinguished: a blog network (formed by linked blogs), post network (formed by linked posts) and blogger network (formed by linked bloggers). The authors consider different methods of identification of links between nodes: (i) hyperlinks to other blogs existing on the blogs, (ii) every pair of nodes, whose distance is smaller than a given constant ϵ are connected by links, (iii) number of k nodes nearest to a given node is connected to it, (iv) all blogs are connected by edges with weights expressing similarities of given blogs.

Another important factor of the models is the dynamics of existing links and their weights in time. In [11], focused on the analysis of the evolving blog groups, the similarity relations between blogs were expressed, which led to considering them as members of the same group. In [4] the authors proposed a method (community factorization) for representation of structures and temporal dynamics of blog groups. In [2] a model for the identification of influential bloggers is presented, which took into consideration the time of interactions and when the given post ceased to be influential, causing new interactions to represent links between blogs.

2.2 Groups in Social Networks

There are many definitions of groups (communities, clusters), mainly according to the area in which they were created. So it is difficult to find in literature an

unequivocal definition of a group, acceptable to everybody [16]. A group can be treated as a dense subset of vertices in a network, which are loosely connected with vertices outside the group. In practice, in complex social networks, groups are not isolated and individuals can be, in a given time, members of many groups. Many methods of finding groups (overlapping or not) have been proposed. In [5] there are detailed descriptions of the most popular methods and algorithms. Every group can be described by several parameters, e.g. density (ratio of the number of links within the group to the maximum possible number of links), stability (the ratio of the number of people, present in both group to the number of all group members), cohesion (ratio of the average strength of links between the members to the average strength of their links with people outside the group).

Due to the nature of the blogosphere (the user may be a member of various discussion groups), the most useful are the algorithms finding overlapping groups. The most prominent representative of this group of methods is CPM algorithm [13,12] where groups are defined as sub-graphs consisting of a set of connected k -cliques. With the increase of parameter k the smaller and more dis-integrated groups arise [13] and there is a suggestion that values of $k = 3, \dots, 6$ seem to be the most appropriate.

2.3 Sentiment Analysis

Emotions are an integral component of statements in social media, especially on blogs or forums. Different groups of users can discuss the same topics in a completely different atmosphere, supporting each other or disagreeing. For each such statement, we can assign a value expressing an emotional attitude: positive, negative, neutral, objective or bipolar [17].

A large increase in interest in problems of analysis of sentiment can be seen around 2001. Some reasons for such interest in this research area are shown in [14]: the development of advanced methods of analysis of natural language, which were already mature enough that it can be successfully applied in practice, more and easier availability of test data that were suitable for such analyzes (mostly available on the WWW) and the increasing demand for intelligent applications.

The term “sentiment analysis” (also used later interchangeably with “opinion mining”) was initially pertained to “automatic analysis of evaluative text and tracking of the predictive judgments” and was closely associated with analyzing market sentiment. Later, the term was rather treated as classifying reviews according to their polarity: either positive or negative. Nowadays the term refers to “computational treatment of opinion, sentiment, and subjectivity in text” [14]. Sentiment analysis is closely related to natural language processing. Analysis of sentiment generally consists of several steps ([17]): part-of-speech tagging (division into language tokens), subjectivity detection (determining the statement as subjective or objective) and polarity detection (for subjective statements evaluate their polarity). There are different techniques and statistical methodologies to evaluate sentiment.

The main difficulty in assessing the sentiment is that it is context-sensitive. Currently, the increasingly popular use of sentiment analysis is the analysis of political blogs [7], and more recently Twitter [15] due to the high amounts of opinions, sentiments and emotions.

2.4 Sentiment Analysis in Domain of Social Networks and Group Identification

The general idea of finding groups in a social network (e.g. blogosphere) is to identify a set of vertices, communicating to each other more frequently than with vertices outside the group, regardless of the expressed emotional potential. Simply counting the number of comments and the weight of edges connecting two users does not distinguish situations when a user writes a comment in support of the ideas expressed by another in a post and when he disagrees with the writer of the post he/she is commenting.

In [18] authors focus on group detection based on links and sentiment – they were finding non-overlapping clusters that share similar sentiment. The researchers claim that this is the first work on sentiment group detection. In this work, they propose two methods of finding such communities. The first method assumes that sentiment can be either positive or negative. In the second method, the range of sentiment is divided into intervals and group users into groups according to the specific differences in the ranges of values of sentiment.

The problem of sentiment based clustering was used directly for the analysis of the blogosphere in [10]. The authors proposed an algorithm called hyper-community detection and they used two methods: content-based hyper-community detection and sentiment-based hyper-community detection. In the first, they extracted topics from blog content, while the second method used sentiment information (from mood tags or emotion words used in posts).

In paper [3], the authors use sentiment analysis with social network approach in the context of radicalisation, searching terrorists in some specific groups from the Youtube portal [1]. They tried to find out whether a chosen group was populated by radicals who could convince others to their beliefs and whether males or females are more radical. Sentiment analysis was used to define the level of radicalization of their comments containing some chosen keywords and social network analysis – to extract key members in the group and to compare some network characteristics between a male and a female group. In article [9] authors tried to predict the success in the Oscar Awards based on analysis of communication on IMDb portal [2]. They used sentiment analysis as a tool to define positivity of the user's opinions about movies – authors searched for positive keywords that were extracted based on their betweenness centrality. The researches took advantage of social network analysis by weighting user posts according to the importance, expressed by betweenness centrality, of users that wrote them and treating most influential users as people who can possibly create trends.

¹ www.youtube.com

² The Internet Movie Database – www.imdb.com

3 Dynamic Models of Social System

Our model of social system, which first version was presented in [8], is adapted to the analysis of the characteristics of groups, their formation, dynamic, reasons and predicted character of future evolution. The state of the system is analyzed in subsequent time intervals called time slots. For each such interval the interactions taking place between entities are analyzed, and groups identified. It is assumed, that the groups may overlap.

For the identification of the groups the Clique Percolation Method [13,12] in the version for a directed graph with weights is used. Then, among such identified groups the stable groups are discovered, using SGCI (Stable Group Changes Identification) algorithm [19,20,6]. The concept of stable groups was introduced due the dynamic character of blogosphere, where groups may change very rapidly, and for our analysis of the evolution of blogosphere the most interesting are groups which last for a longer time. The condition that a group is considered as a stable group is to identify in the next time slots groups with similar sets of members, evaluated using the Jaccard measure modified by us (expressed as a ratio of size of intersection of the pair of considered groups to the size of one of the groups from them - the larger value of such a ratio is considered as the modified Jaccard measure). The group is stable if it has such similar groups at least during the minimum assumed number of time slots.

The model is described in two parts – the first (described in section 3.1) concerns the fundamental elements of the model – entities and interactions among them (more details in section 3.2), and the second (section 3.3) – the organization with social system (section 3.4) and groups.

3.1 Fundamental Model of Social System

Dynamic model of social system $Soc(t)$, describes its state in the time slot t :

$$Soc(t) = (N(t), X(t), \zeta, I(t), Org(t)) \quad (1)$$

where:

$N(t)$ – set of entities building a social system,

$X(t)$ – vectors of values of measures calculated for the entities from the set N ,

$X_{N_i}(t)$ represents a vector of measures of the entity N_i for the time slot t ,

ζ – function, which assigns values of a vector of measures to entities N ,

$I(t)$ – set of interactions, consists of all the interactions between entities, together with the times they took place, their type, sets of involved entities and their roles in the interaction, the content and/or sentiment of the exchanged information,

$Org(t)$ – organization of the social system, described in section 3.3.

3.2 Interactions between Bloggers

Applying the model to the analyzed blogosphere domain and the analyzed problem of group identification, we can distinguish the following kinds of interactions

between entities: commenting on posts, commenting on a comment, static links in blogs or posts to another blog/post, logins/nicks of bloggers mentioned in the content of post or comment. The identification of some of these mentioned interaction types burdened by the varying level of uncertainty, whether the assignment was correct or not. In our work we are focusing on the interactions caused by commenting on posts of other users or by commenting on previously written comments to posts. These interactions have varying characters which make them useful while analyzing the dynamics of groups and for a significant part of them it is possible to correctly identify who is being addressed.

The representation of the individual interaction, assumed by us, is as follows:

$$i_t = (N_i, N_j, N_p, t_z, k, s) \quad (2)$$

where: N_i – interaction initiator (writer of post or comment), N_j – the addressee of the comment (sometimes not specified), N_p – author of post to which the comment/interaction is written, t_z – given time slot, k – type, which may be post, comments to post, comments to comment, s – sentiment value, expressed in the bounded interval $[-1, 1]$.

3.3 Organization of Social System

The organization of social system Org is expressed using the following elements:

$$Org(t) = (R(t), \psi, GT(t), \gamma, G(t), \xi, XG(t), \zeta^g) \quad (3)$$

$R(t)$ – social relation, shaped as the results of interactions taking place,

ψ – function which builds social relations R between a pair of entities, on the basis of interaction taking place between them,

$$R(a, b, t_z) = \psi(I(a, b, t_z)) \quad (4)$$

Equation (4) shows social relation between users a and b in the time slot t_z , ψ returns a strength of the relation expressed as a positive real number.

$GT(t)$ – set of identified temporary groups,

γ – a function which assigns entities to fugitive groups, $\gamma : N \times R \rightarrow GT \times \{0, 1\}$

The used method of the classifications of nodes to groups is as follow: for each time slot, the fugitive groups are identified on the basis of the version of the CPM algorithm, calculated for a directed graph with weights.

$G(t)$ – set of identified stable groups, Groups are considered as stable, when their life span equals at least l_{tmin} (which is set in the tests as equal to 3).

ξ – function which identifies stable groups among fugitive ones, $\xi : GT \rightarrow G$,

$XG(t)$ – vectors of values of measures calculated for the groups by ζ^g , $XG_{Gr_i}(t)$ represents a vector assigned to a group Gr_i which may be temporary (element of G) or stable (element of GT),

ζ^g – a function which calculates values of defined vectors of measures for temporary or stable groups and assigns it to $XG(t)$.

3.4 Building of Social Relations

In our model of social relations two main factors are considered: the frequency of interactions between nodes and the sentiment of interactions. The sentiment of the interaction may be classified into one of three groups: positive interaction, negative interaction or neutral (indifferent) interaction, on the basis of content analysis and strength of positive or negative connotation of words appearing in the comment.

In this work the following versions of the ψ function are distinguished:

- ψ_{pn} – considers all comments as addressed to the author of post, does not take sentiment of comments into consideration,
- ψ_{cn} – scores comments which have a defined addressee of the comments as addressed to this addressee and not to the post author, if it is not possible to identify the addressee, the comment is scored as addressed to the post author, sentiment is not taken into consideration,
- ψ_{cs} – scores comments which have a defined addressee of the comments as addressed to this addressee and not to the post author, if it is not possible to identify the addressee, the comment is scored as addressed to the post author, sentiment is taken here into consideration, and either relations caused by each kind of the sentiment (positive, negative, neutral) are considered separately or average values of the sentiment for every existing links are calculated, making this link to appear only in that adequate kind of sentiment model. The following subversion can be distinguished:
 - $\psi_{cs,p}$, $\psi_{cs,n}$, $\psi_{cs,i}$ (sentiment counting models) – in the given models, only interactions with positive (cs,p), negative (cs,n) or neutral (cs,i) sentiment are considered, for every pair of users interactions with each sentiment are scored separately,
 - $\psi_{cs,p+i}$ – similar to previous ones, interactions with positive or neutral sentiment are taken into consideration together, the interactions with negative sentiment are omitted,
 - $\psi_{cs,p}^a$, $\psi_{cs,n}^a$, $\psi_{cs,i}^a$ (sentiment mean models) – the average value of sentiment for a given ordered pair of users is taken into consideration, the directed relation between two users may be assigned only to one of these (which means positive, negative and neutral) models,
 - $\psi_{cs,p+i}^a$ – similar to previous ones, but considers links with both positive or neutral average sentiment.

4 Application of Models to Group Identification and Analysis

4.1 Description of Experiments

Data Set. The analyzed data set contains data from the portal www.salon24.pl which consists of blogs (mainly political, but also have subjects from different

areas). The data set consists of 26 722 users (11 084 of them have their own blog), 285 532 posts and 4 173 457 comments within the period 1.01.2008 - 31.03.2012. The analyzed period was divided into time slots, each lasting 30 days. The neighboring slots overlap each other by 50% of their length and in the examined period there are 104 times lots.

The large graph from all time slots consists of 26 053 nodes and 663 098 edges. Nodes in this graph are the users - both the owners of blogs and people only commenting on other posts. The number of nodes in the graph is lower than the overall number of active authors (26 722) in the given period, because some posts did not have any comments. Thus their authors cannot appear in this graph, unless they had commented on others or had any of their posts commented on.

Data Set Preparation. We decided to remove edges with weights below 2 to eliminate some noise and to reduce calculation time. After removing such edges, the number of nodes was equal to 15 578 (59.8% of initial number of nodes) and the number of edges to 311 718 (47% of the initial number of edges). When we are considering the number of connections as the number of edges multiplied by their weights, then the removed edges constitute 8.42% of such connections.

To extract groups from networks we used CPMd version (for directed graphs) of CPM from CFinder³ tool, for different k in ranges 3 to 5.

Sentiment Calculation. The sentiment for posts and comments was calculated using a tool developed at the Luminis Research company⁴. Their method is based on searching words from analyzed text in a dictionary and counting sentiment for found ones. The dictionary is manually built and contains about 37 000 words (including about 4000 positive and negative words together – the others are the neutral ones). Each word in the dictionary has a weight in the range $-1; 1$ - negative values determine negative sentiment, positive – positive one and neutral words have a weight equal to zero (intensity of positive or negative sentiment depends on assigned value - the closer value to 1 or -1, the greater the intensity of the sentiment is). Then the sentiment values for found words in the dictionary are summed and using the sum value, the number of positive, negative and neutral words in analyzed text the final sentiment value is calculated (based on heuristic equation with mentioned values). The final value describing the overall sentiment is between -1 and 1, but thresholds for negative, neutral and positive sentiment need adjusting. This can be done by analyzing some texts (part of texts earlier marked by algorithm) by human, manually assigning sentiment values (positive/negative/neutral) for them, next comparing these values with algorithm ones and finally setting appropriate thresholds.

In order to adjust thresholds for sentiment values, we analyzed about 150 random texts and based on this analysis we set the following thresholds: negative (< 0), neutral ($0 - 0.3$), positive: (> 0.3).

³ www.cfindex.org

⁴ www.luminis-research.com

4.2 Comparison of Post and Comments Models

During the analysis of the groups emerging in the blogosphere it is very important to identify, at first, the real characters of the interactions taking place, especially who is sending and receiving them.

In the case of comments, although they are assigned to a given post, in reality they often refer directly to an earlier entry commenting in this post. In the blog portal salon24 we analyzed, the identification of the receiver of the comments is not that evident as they are only assigned to the post and the commenter can only refer to the name of the bloggers whose comment they are commenting on. But, this is not done in an automatic way, the blogger is only able to do it by appropriately writing the subject of their comment (by writing “@bloggername” there). It is not always common practice, and if not specified, the writer of any post is considered as a receiver of that comment.

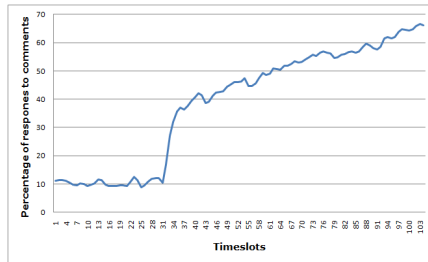


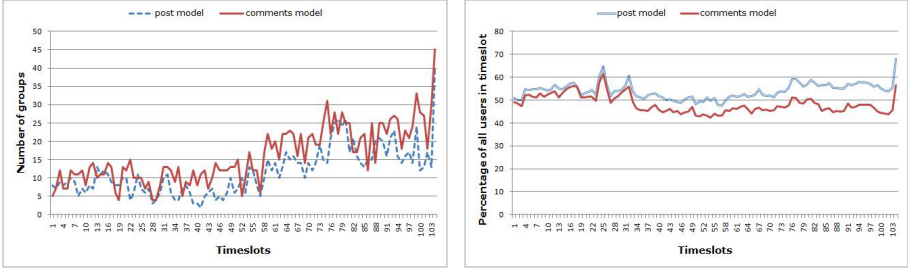
Fig. 1. Percentage of responses of type comment-comment to all responses

For all 4 173 457 comments we identified 1 953 571 as comments that are responses to other comments (about 50 %). In fig. 1 a noticeable increase in the percentage of comments having the receiver specified in such a way in time may be seen, so in the majority of cases it is possible to correctly consider that information in the model, what increases the accuracy of the represented interactions between bloggers and the subsequent emerging social relations.

Such assumptions are confirmed by the fact that in the new model (comments model ψ_{cn}) more groups were identified (see fig. 2a) than in old one (post model ψ_{pn}), a smaller part of user are not assigned to any groups (see fig. 2b).

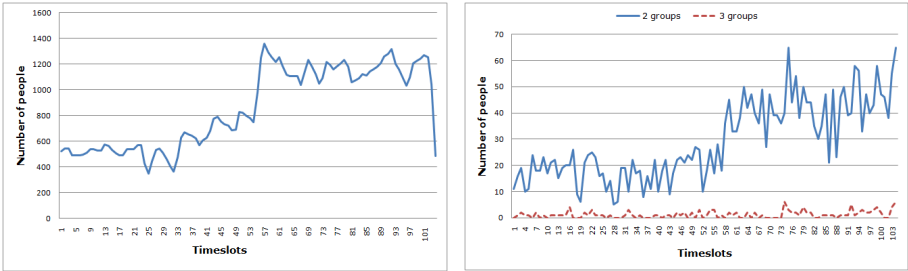
In figs. 3a and 3b, the numbers of users belonging to one, two or three stable groups in each interval for $k=3$ are specified. The figure presents mentioned belongings only in the comments model, but in the post model diagram is very similar. We can notice that these numbers increase, mostly because of the increase of the popularity of the portal and the significance of political events taking place.

In tab. 1a there are presented the total numbers of stable groups with different sizes, calculated for k equal 3, 4 and 5, for models based on comments assigned to post author (ψ_{pn}) and previous comments authors (ψ_{cn}). The most significant differences are obtained for low sizes of groups. Usually, models with comments give more groups, because of higher quantity of different links in these models.



(a) Number of groups in timeslots. (b) Users not belonging to any stable group.

Fig. 2. Comparison between post and comments models for $k=3$



(a) 1 group (b) 2 and 3 groups

Fig. 3. Membership of people to groups for $k=3$ in comments model

In tab. 11b one can see, that comments model gives us more stable, dense and cohesive groups what is confirmed by their mean values. The comment model gives more different connected pairs of bloggers both inside the group which influence increase of density and cohesion.

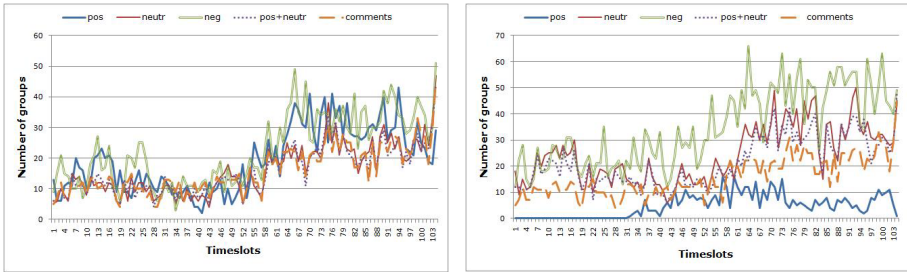
4.3 Comparison of Sentiment Models

In the next analysis we focused our attention on comparing models with comments without (ψ_{cn}) and with sentiment (different versions of (ψ_{cs}) function, described in section 3.4) for $k=3$.

In fig. 4a and 4b the negative groups are dominating, but for groups in model with average sentiment (in fig. 4b - $\psi_{cs,n}^a$), stronger negative interactions are necessary to form them. One can notice, that such relations build well-shaped groups with strongly connected members. Such behavior seems to be natural in the politic blogs, especially discussing controversial, emotion inspiring/arousing subjects. It is worth noting that negative relation between bloggers does not need to signify that the first blogger has a negative attitude regarding the second one, but that during the discussed subject they express negative emotions caused by another blogger or the general situation.

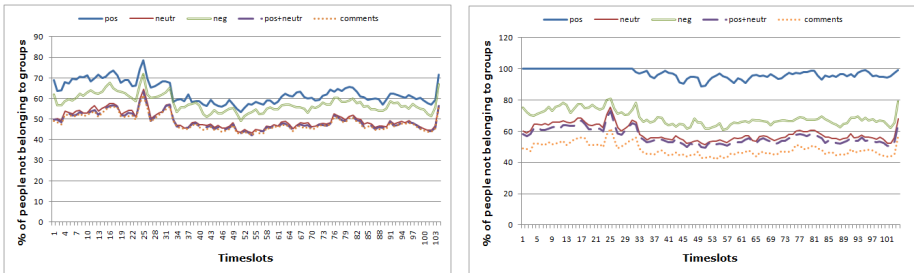
Table 1. Comparison between posts and comments models

| (a) Stable group sizes | | | | | | | (b) Mean values for stable groups | | | | |
|------------------------|------|----------|------|----------|------|----------|-----------------------------------|----------|-------|-------|-------|
| Size | k=3 | | k=4 | | k=5 | | Measure | Model | k=3 | k=4 | k=5 |
| | post | comments | post | comments | post | comments | | | | | |
| 3 | 992 | 1350 | 0 | 0 | 0 | 0 | Stability | post | 0.100 | 0.081 | 0.099 |
| 4 | 81 | 147 | 1373 | 1358 | 0 | 0 | | comments | 0.133 | 0.098 | 0.106 |
| 5 | 25 | 32 | 235 | 210 | 966 | 1059 | Density | post | 0.459 | 0.489 | 0.511 |
| 6 | 7 | 10 | 52 | 54 | 213 | 205 | | comments | 0.598 | 0.631 | 0.657 |
| 7 | 2 | 4 | 19 | 21 | 74 | 63 | Cohesion | post | 73.7 | 36.5 | 29.8 |
| 8 | 1 | 3 | 13 | 6 | 39 | 35 | | comments | 157.9 | 46.0 | 41.9 |
| 9 | 3 | 1 | 3 | 10 | 26 | 26 | | | | | |
| 10 | 0 | 0 | 5 | 4 | 20 | 13 | | | | | |
| 11-50 | 0 | 0 | 40 | 58 | 50 | 121 | | | | | |
| 51-100 | 0 | 0 | 1 | 0 | 14 | 10 | | | | | |
| 101-200 | 0 | 0 | 4 | 5 | 30 | 22 | | | | | |
| > 200 | 104 | 104 | 98 | 99 | 57 | 69 | | | | | |



(a) Sentiment counting model. (b) Sentiment mean model.

Fig. 4. Comparison of number of groups in slots in sentiments models for k=3



(a) Sentiment counting model. (b) Sentiment mean model.

Fig. 5. Comparison of percent of users not belonging to any stable group in sentiment models for k=3

In fig. 5a and fig. 5b one can see a significant difference between models when counting each kind of sentiment interactions separately and using the average value of the sentiment. In mean comments model interactions with positive and negative sentiment canceling each other out and the obtained average is close to

0, for this reason there are significantly more persons belonging to the groups constructed for neutral average sentiment ($\psi_{cs,i}^a$). It confirms the predictions that a model with an average value of sentiment identifies only radical sentiments in the case of positive and negative relations.

Table 2. Comparison of stable groups sizes between sentiment models for k=3
 (a) Sentiment counting model (b) Sentiment mean model

| Size | pos | neutr | neg | pos+neutr | comments | Size | pos | neutr | neg | pos+neutr | comments |
|---------|------|-------|------|-----------|----------|---------|-----|-------|------|-----------|----------|
| 3 | 1606 | 1359 | 1855 | 1329 | 1350 | 3 | 282 | 2071 | 2780 | 1873 | 1350 |
| 4 | 220 | 147 | 256 | 149 | 147 | 4 | 87 | 238 | 458 | 201 | 147 |
| 5 | 52 | 33 | 74 | 36 | 32 | 5 | 25 | 58 | 139 | 57 | 32 |
| 6 | 19 | 12 | 14 | 10 | 10 | 6 | 21 | 23 | 62 | 16 | 10 |
| 7 | 16 | 2 | 11 | 4 | 4 | 7 | 15 | 10 | 22 | 9 | 4 |
| 8 | 5 | 2 | 8 | 2 | 3 | 8 | 6 | 3 | 18 | 0 | 3 |
| 9 | 4 | 3 | 4 | 2 | 1 | 9 | 6 | 3 | 7 | 1 | 1 |
| 10 | 2 | 1 | 3 | 0 | 0 | 10 | 6 | 2 | 12 | 3 | 0 |
| 11-50 | 10 | 3 | 13 | 2 | 0 | 11-50 | 48 | 12 | 20 | 6 | 0 |
| 51-100 | 0 | 0 | 0 | 0 | 0 | 51-100 | 3 | 0 | 2 | 0 | 0 |
| 101-200 | 14 | 0 | 2 | 0 | 0 | 101-200 | 0 | 1 | 28 | 1 | 0 |
| > 200 | 90 | 104 | 102 | 104 | 104 | > 200 | 0 | 103 | 72 | 103 | 104 |

Table 3. Comparison of stable groups parameters (mean values for all stable groups in time slots) between sentiment models for k=3

| (a) Sentiment counting model | | | | | (b) Sentiment mean model | | | |
|------------------------------|-----------|---------|----------|--|--------------------------|-----------|---------|----------|
| Model | Stability | Density | Cohesion | | Model | Stability | Density | Cohesion |
| pos | 0.114 | 0.538 | 89.8 | | pos | 0.229 | 0.448 | 34.8 |
| neutr | 0.130 | 0.59 | 157.3 | | neutr | 0.087 | 0.545 | 104.8 |
| neg | 0.117 | 0.557 | 135.4 | | neg | 0.087 | 0.526 | 61.4 |
| pos+neutr | 0.135 | 0.593 | 157.4 | | pos+neutr | 0.097 | 0.554 | 116.6 |
| comments | 0.133 | 0.598 | 157.9 | | comments | 0.133 | 0.598 | 157.9 |

Analyzing the total number of groups with different sizes depending on model and character of polarization (tab. 2), one can notice that counting separately the groups in each model, the sentiment counting models give much more positives groups then sentiment mean models, but significantly less for negative and neutral groups.

In the sentiment mean model the most stable groups were obtained for positive sentiment (tab. 3), it may be caused by the fact, that the number of these groups is low (as can be seen in tab. 2). The method of the identification of relations used in this model gave only groups exchanging very positive content, such specific groups are characterized by a high stability of memberships. For remaining models, measures of groups for sentiment mean models are lower or much lower than for the sentiment counting models, so they identify groups less dense, less stable and less separated from the environment. In sentiment mean model there is a lot less connections between nodes than in sentiment counting model, so it may explain smaller values of density.

5 Conclusion

The paper introduces a set of developed models describing social networks, taking into consideration different kinds of interactions and sentiment polarization. Models were applied to the analysis of stable groups, identified in the selected blog portal. The introduced set of models can help in systematization of the problem domain and allow us to identify research directions and relations between them.

Several experiments were conducted which delivered new, detailed information about a character and behavior of groups of users on the portal. The method of identification of stable groups in blogosphere was improved which allowed us to obtain more stable, dense and cohesive groups. In new model (comments model) lower number of users did not belong to any group. Introduction of the sentiment as an interaction attribute allowed to observe different characteristic behaviors of groups with different polarization. Positive sentiment groups are formed around not controversial topics while negative sentiment groups are associated with controversial matters and possibly quarrels.

The presented solutions will be applied to analyze other blog portals and different kinds of social media, for example microblogs. The next works will embrace: improving the quality of the sentiment analysis, key bloggers identification and analysis of their memberships in given groups. We are going to integrate presented sentiment models with our research on group dynamics and prediction of group evolution, as well as the identification of the most significant, strongly linked members of the group, constituting group cores. Another direction is to associate models based on sentiment with extended description of groups which considers the most popular discussed subjects identified by analysis of tags or post and comment content.

Acknowledgments. The authors thank P. Maciołek who provided and allowed the use of the algorithm and tools for analysis of sentiment of texts in Polish language.

References

1. Agarwal, N., Liu, H.: *Modeling and Data Mining in Blogosphere*. Morgan & Claypool Publishers (2009)
2. Akritidis, L., Katsaros, D., Bozaris, P.: Identifying influential bloggers: Time does matter. In: *Procs. of the 2009 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology, WI-IAT 2009*, vol. 1, pp. 76–83. IEEE Computer Society, Washington, DC (2009)
3. Birmingham, A., Conway, M., McInerney, L., O'Hare, N., Smeaton, A.F.: Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In: *Proc. of the 2009 Int. Conf. on Adv. in Social Network Analysis and Mining*, pp. 231–236. IEEE Comp. Soc., Washington, DC, USA (2009)
4. Chi, Y., Zhu, S., Song, X., Tatemura, J., Tseng, B.L.: Structural and temporal analysis of the blogosphere through community factorization. In: *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2007*, pp. 163–172. ACM, New York (2007)

5. Fortunato, S.: Community detection in graphs. *Phys. Rep.*, ch. 486 (2010)
6. Gliwa, B., Saganowski, S., Zygmunt, A., Bródka, P., Kazienko, P., Koźlak, J.: Identification of group changes in blogosphere. In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, August 26-29*. IEEE Computer Society (2012)
7. Gryc, W., Moilanen, K.: Leveraging textual sentiment analysis with social network modelling: Sentiment analysis of political blogs in the 2008 u.s. presidential election. In: *Procs. of the "From Text to Political Positions" Workshop (T2PP 2010)*. Vrije Universiteit, Amsterdam (2010)
8. Koźlak, J., Zygmunt, A.: Agent-based modelling of social organisations. In: *International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2011, Korean Bible University, Seoul, Korea, June 30-July 2*, pp. 467–472. IEEE Computer Society (2011)
9. Krauss, J., Nann, S., Simon, D., Fischbach, K., Gloor, P.: Predicting movie success and academy awards through sentiment and social network analysis. In: *ECIS 2008 Proceedings* (2008)
10. Nguyen, T., Phung, D.Q., Adams, B., Tran, T., Venkatesh, S.: Hyper-community detection in the blogosphere. In: *Proceedings of Second ACM SIGMM Workshop on Social Media, WSM 2010*, pp. 21–26. ACM, New York (2010)
11. Ning, H., Xu, W., Chi, Y., Gong, Y., Huang, T.S.: Incremental spectral clustering by efficiently updating the eigen-system. *Pattern Recogn.* 43(1), 113–127 (2010)
12. Palla, G., Abel, D., Farkas, I.J., Pollner, P., Derényi, I., Vicsek, T.: k-clique percolation and clustering. In: *Bollobás, B., Kozma, R., Miklós, D. (eds.) Handbook of Large-scale Random Networks*. Springer (2009)
13. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818 (2005)
14. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2) (2008)
15. Shamma, D.A., Kennedy, L., Churchill, E.F.: Statler: Summarizing media through short-messaging services. In: *CSW 2010*. ACM, USA (2010)
16. Tang, L., Liu, H.: Graph mining applications to social network analysis. In: *Aggarwal, C., Wang, X. (eds.) Managing and Mining Graph Data*. Springer (2010)
17. Tromp, E., Pechenizkiy, M.: Senticorr: Multilingual sentiment analysis of personal correspondence. In: *Proc. of ICDM 2011 Workshops*. IEEE Press (2011)
18. Xu, K., Li, J., Liao, S.S.: Sentiment community detection in social networks. In: *Procs. of the 2011 iConference, iConf. 2011*, pp. 804–805. ACM, NY (2011)
19. Zygmunt, A., Bródka, P., Kazienko, P., Koźlak, J.: Different approaches to groups and key person identification in blogosphere. In: *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011, Kaohsiung, Taiwan, July 25-27*, pp. 593–598. IEEE Computer Society (2011)
20. Zygmunt, A., Bródka, P., Kazienko, P., Koźlak, J.: Key person analysis in social communities within the blogosphere. *J. UCS* 18(4), 577–597 (2012)

Dark Retweets: Investigating Non-conventional Retweeting Patterns

Norhidayah Azman, David E. Millard, and Mark J. Weal

Web and Internet Science, Department of Electronics and Computer Science,
University of Southampton, UK
{nba08r,dem,mjw}@ecs.soton.ac.uk

Abstract. Retweets are an important mechanism for recognising propagation of information on the Twitter social media platform. However, many retweets do not use the official retweet mechanism, or even community established conventions, and these “dark retweets” are not accounted for in many existing analysis. In this paper, a comprehensive matrix of tweet propagation is presented to show the different nuances of retweeting, based on seven characteristics: whether it is proprietary, the mechanism used, whether it is directed to followers or non-followers, whether it mentions other users, if it is explicitly propagating another tweet, if it links to an original tweet, and what is the audience it is pushed to. Based on this matrix and two assumptions of retweetability, the degrees of a retweet’s “darkness” can be determined. This matrix was evaluated over 2.3 million tweets and it was found that dark retweets amounted to 12.86% (for search results less than 1500 tweets per URL) and 24.7% (for search results including more than 1500 tweets per URL) respectively. By extrapolating these results with those found in existing studies, potentially thousands of retweets may be hidden from existing studies on retweets.

Keywords: retweets, tweet propagation, dark retweets, Twitter, microblogs.

1 Introduction

There have been several studies investigating the propagation of tweets, focusing on retweeting as the main mechanism of propagation. Existing work has traditionally looked at both conventional retweeting mechanisms, such as Twitter’s proprietary retweeting mechanism, and manually inserted retweet markers, such as “RT” and “via.”

Retweets form an important part of tweet propagation research, from conversational patterns [1] to overall retweet ratios [2,3]. However, there are several different nuances to the act of propagating a tweet, mainly due to the different mechanisms and features involved in propagating a tweet. This paper aims to describe these different nuances of retweeting by deconstructing the action into separate characteristics. This paper also introduces “dark retweets”, describes

the different assumptions of retweetability, and also discusses the possible impact they may have on future tweet propagation studies. Our research shows that there are many different ways a user may propagate information across Twitter, and there may be proportions of propagating tweets which may have been missing from existing work.

Section 2 discusses several existing papers on retweets and in particular those which are not explicitly marked as retweets. Section 3 outlines the matrix of tweet propagation, based on seven characteristics: whether it is proprietary, the mechanism used, whether it is directed to followers or non-followers, whether it mentions other users, if it is explicitly a retweet, if it links to an original tweet, and what is the audience it is pushed to. Section 4 focuses on dark retweets and the different assumptions of retweetability. The following Section 5 presents the evaluation of this matrix, from descriptions of the experimental setup to discussions of the results found.

2 Research Background

Since the introduction of Twitter and microblogs, researchers have been focusing on patterns of propagation across Twitter. boyd et. al [1] was one of the earliest studies focusing on retweets. Two datasets were used; one being a random sample of tweets taken in 5-minute intervals, and the other being a sample of around 203,000 retweets. From the random sample, the study claimed 3% were retweets, and that the existence of URLs increases the retweetability of that tweet. This paper also acknowledges the existence of tweets which contained texts which were similar to previously published tweets, yet do not contain conventional retweet markers. However, due to the difficulty in determining the provenance of these tweets, they were not focused upon within this study.

Several researchers have tackled this problem by making assumptions about the existence of propagation paths between subsequent tweets based on the timestamps of those published tweets and the content similarity between them. The study that is being presented in this paper is most similar to the work done by Adar et. al [4], which investigated the existence of implicit URL links within the blogosphere. Similarly, the work by Matsumura et. al [5] also included the assumption that if a collection of blogs which contained the same URL or trackback also contained the same terms, then it was assumed that the first blog in that collection was influencing the subsequent blogs. The paper written by Galuba et. al [6] described the F-cascade within tweets. It involved users who seemed to have copied a URL that was previously tweeted by someone they follow.

Similarly, in the work done by Wu et. al [7], the phrase “reintroduction of content” was used to describe intermediary tweets which are similar to previously published tweets but also do not contain conventional retweet markers. Their dataset consisted of tweets which only had URLs in them. In this paper, retweets and reintroductions were treated equivalently, with no separation between the two. The same approach was taken by Bakshy et. al [8], who studied

influence prediction within Twitter. Their metrics indicating influence were not restricted to just retweeting by including retweet markers within tweet texts. Their study also used the approach of using URLs as unique keys which group tweets together, thus all instances of tweets which include the URLs being focused on were considered as a “rebroadcast” of influence. Again, their paper did not differentiate between conventional and non-conventional retweets.

In the study by Nagarajan et. al [9], tweets “without indication of retweeting or making references to others” were initially classified as “other” tweets. The paper studied datasets based on three topics: Health Care Reform Debate, Iran Election, and the ISWC conference. Similarity engines were then used to retrieve tweets similar to the top 10 most frequent tweets in each of these three datasets. This allows tweets without explicit retweeting markers to be grouped together. The retweet patterns of these groups were then subsequently studied. The paper claimed that tweets for calls for action, collective groups and crowdsourcing domains are more likely to have more unmarked, unattributed retweets, as opposed to information sharing tweets.

Given the above studies, this paper presents a combination of the theory of implicit links as presented by Adar et. al [4] with the methodology of only using tweets containing URLs in them [7,8]. The hypothesis of this paper is that there exists a proportion of dark retweets, or tweets which are propagated without using conventional retweeting mechanisms. This may lead to hidden data which would not have been focused upon within existing studies on tweet propagation.

3 Matrix of Tweet Propagation

In this study, tweets are deconstructed using several characteristics. In the following descriptions of these characteristics, the abbreviations stated in parentheses are used in the matrix of tweet propagation shown in Table II.

Proprietary: The propagation of a tweet is considered proprietary (P) if it was published using methods that were built into the Twitter use case structure. For example, a retweet is considered proprietary if it was made using Twitter’s proprietary methods, either by *a*) clicking the retweet button on its official user interfaces (e. g. web page, mobile apps), or *b*) third party apps utilizing the Twitter API’s proprietary retweeting method.

Propagation mechanism: Tweets can either be propagated as a retweet (‘Rt’), a reply (‘@’), or a direct message (DM).

Follower or non-follower: The propagated tweets can be made by either a follower (F) or a non-follower (nF). This relates to the relationship between the author of the originating tweet and the person propagating that tweet. In this column, a follower relationship is marked as ‘1’ in Table II, while a non-follower relationship is marked as ‘0’.

Mentions other users: A mention exists in a tweet if its text contains other people’s Twitter usernames in them.

Explicit: A tweet is considered to be explicitly propagated by a user if a retweet marker such as ‘RT’ or the ‘@’ reply marker was written explicitly in the tweet text. Proprietary retweets/replies and manually marked retweets/replies are considered explicit, while those without any retweet/reply markers are considered as implicit. For example, “*Done! RT @User_X Sign this petition! http://bit.ly/SmgF*” would be considered as an explicit retweet, while “*@User_Y Please sign this petition: http://bit.ly/SmgF*” would be considered as an explicit reply.

Links to original tweet: If a propagating tweet contains metadata that links to the originating tweet, then the originating tweet’s unique ID is stored. The Twitter API automatically stores this metadata when its proprietary retweet or reply mechanism is used.

Tweet pushed to: all or some people: This denotes the difference between the visibility of a retweet and a reply. Retweets are pushed onto the timelines of all the followers of the retweeter. In Table 1, a ‘11’ value means the tweet is pushed to all and some of the authors’ followers. This visibility changes for replies addressed to a specific Twitter user. They are only pushed to the timelines of mutual followers of the reply creator and the person being addressed to. For example, if User A makes a reply to User B, then the reply will only appear on the timelines of those who follow both Users A and B. In theory, it is possible for anyone to see this reply by looking up User A’s personal page on Twitter, which lists all the tweets made by User A. However, this requires extra effort from those who don’t follow Users A nor B, hence it is assumed that there exists a state where a tweet is visible only to some people but not all. This is marked as ‘01’ in Table 1.

Using these seven characteristics, a binary matrix was constructed to illustrate all possible combinations of these characteristics. This process resulted in a 2^{10} matrix, containing 1024 rows. Each row was then manually evaluated to identify if it is possible for any single tweet to possess the combination of characteristics as recorded in that row. Table 1 shows the valid rows after this evaluation was completed. The abbreviations used under the Categories column in Table 1 come from the characteristics described in Section 3. The categories in Table 1 were made mainly by grouping the rows according to the characteristics of Proprietary, Mechanism and Follower/Non-follower. For example, PRtF denotes proprietary (P) retweets (Rt) made by followers (F), while @nF denotes a non-proprietary reply (@) made by a non-follower (nF). In Table 1, there are rows which are coloured in three shades of grey. The lightest shade of grey corresponds to Orphan Retweets and Replies, which will be described in Section 3.5. The rows which are coloured in the two darkest shades of grey correspond to Dark Retweets, which will be described in Section 4.

3.1 Original Tweets and Mentions

Table 1 shows 18 different groups. The base group is Original Tweets, which a) do not seem to have been made using any proprietary retweeting or replying

Table 1. Matrix of Tweet Propagation

| Categories | Proprietary | Mechanism | | Explicit | F/nF | Link to original | Mentions other users | Push | |
|---------------------------|-------------|-----------|------|----------|------|------------------|----------------------|------|------|
| | | Rt | @ DM | | | | | All | Some |
| Original Tweet | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PRtF | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| PRtnF | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| Rt@F | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| Rt@nF | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| RtF | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| RtnF | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| P@F | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| P@nF | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| P@RtF | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 |
| P@RtnF | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| @F | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| @nF | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| @RtF | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| @RtnF | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| PDMF | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| Orphan Rt (Ori Not Found) | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| Orphan @ (User Not Found) | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |

mechanisms, *b*) do not seem to be explicitly propagating another tweet, and *c*) therefore do not have any links to originating tweets nor users.

There also exists the Original Mentions group, which is similar to Original Tweets, but contains mentions in its texts. In this paper, the rows for Original Mentions were omitted from Table 1 for brevity, and Original Mentions were grouped together with Original Tweets, under the assumption that Original Mentions behave similarly to Original Tweets. This is because they also do not seem to have used any retweeting or replying mechanisms.

3.2 Explicitness and Links to the Originating Tweet

Tweets made using proprietary retweeting or replying methods are considered to cause two other characteristics to be true, namely Explicit (explicitly propagating another tweet), and Links to Original (contain metadata that links to the tweet that is being retweeted or replied to). Therefore these proprietary retweets and replies are marked in Table 1 with the value of 1 under the Explicit and Link to Original columns.

Non-proprietary tweets are also considered to be explicit only if they include retweet and/or reply markers within their texts.

3.3 Multiple Mechanisms in Tweets

Several categories include two mechanisms, such as Rt@ and @Rt. Although the main common factor between these categories is the existence of multiple mechanisms when creating these tweets, there are distinct differences between these groups according to the order of the mechanisms used.

The Rt@ category was created specifically for retweets that were made using Twitter's proprietary replying mechanism. Manually typing in retweet markers in front of copied and pasted tweets has been the traditional way of creating retweets before Twitter's proprietary retweeting mechanism was created. Manual retweets allow users to modify the text of the tweet in order to add responses or other new content into the retweet. This modification ability does not exist with Twitter's proprietary retweeting mechanism, which propagates tweets in its original form. A completely manual retweet – where the user manually types in 'RT @User_B' and then copies User B's tweet – would not contain any metadata that links to another tweet, which is opposite to all proprietary Twitter retweets or replies.

However, there exists certain retweets which are not marked by the Twitter REST API as being made using Twitter's proprietary retweet mechanism. Even so, they still contain metadata linking to originating tweets. On further inspection, these tweets were found to be retweets that were manually created after the proprietary replying mechanism was used. For example, User A would like to retweet some text written by User B, but instead of clicking on the 'Retweet' button, User A clicks the 'Reply' button next to User B's tweet. This action causes User A's input textbox for new tweets to be automatically filled with '@User_B', and this allows User A to copy and paste User B's tweet, prefix 'RT'

or other retweet markers in front of the whole text, or modify the text slightly and prefix it with ‘MT’ (modified tweets). This retweeting style would not be classified by the Twitter API as a proprietary retweet, therefore in Table II the Rt@ categories contain 0 under the Proprietary column.

For @Rt categories, these tweets were intended to become replies, where the tweet texts begin with a mention to another Twitter user. However, the tweet texts also contain retweet markers such as ‘RT’ or ‘via’. These @Rt categories are particularly interesting because the reach of these replies are not similar to a normal retweet. Section III has already discussed the visibility of retweets and replies. This difference in reach may have an implication to future retweet propagation studies.

3.4 Limited Visibility of Direct Messages

The PDMF category in Table II concerns direct messages (DM) which can only be accessed by the parties involved in private interactions. Due to this private nature of DMs, we could not study DM propagations in more detail.

3.5 Orphan Retweets and Replies

As seen in Table II, Orphan Rt and Orphan @ categories exist due to certain missing elements.

A retweet is considered as an Orphan Rt if the Twitter API labels it as a proprietary retweet, but the metadata related to the author of the originating tweet is missing. On further manual checks on a separate trial dataset, it was found that this is because the tweet that is being retweeted no longer exists. Interestingly, the Twitter API response does not delete the metadata linking to the unique ID of the deleted tweet, but returns an empty response for the originating author’s metadata instead. In Table II, the Link to Original column is marked with 1 but Mentions Other Users is marked with 0.

Similarly, an orphan reply (Orphan @) exists when the person being replied to (the username prefixed at the start of the tweet text) no longer exists. When orphan replies are looked up via the Twitter API, the metadata for linking to originating tweets and also originating authors become unavailable. In Table II, the Link to Original and Mentions Other Users are both marked with 0s.

Due to the unique characteristics of these orphan categories, they are grouped separately to all the other categories in Table II.

4 Dark Retweets

A dark retweet is defined as a retweet which is propagated using non-conventional retweeting methods. The term “dark” is used to denote how undetectable or invisible a retweet is. In Table II, several rows were shaded in different shades of grey. The darkness of the shades signify the degree of difficulty in detecting the retweet. There are two main assumptions of retweetability, consisting of varying degrees:

Assumption #1. Is Tweet A a retweet of Tweet B? This assumption becomes more concrete if Twitter API returns a looked up tweet as a proprietary retweet, and metadata of the originating tweet exists. This assumption also becomes moderately concrete for non-proprietary retweets if retweet markers such as ‘RT’ or ‘via’ exists within the tweet text. However, this latter assumption is still debatable – is this manually marked, non-proprietary retweet referring to an original tweet that does indeed exist?

Assumption #2. Who was the originating author? This assumption becomes more concrete if a tweet looked up via Twitter API results in metadata identifying the author of the tweet that is being retweeted. This assumption also becomes moderately concrete for non-proprietary retweets if they contain the username of the perceived author of the originating tweet. However, this latter assumption is still debatable – is the manually mentioned user the correct author of the originating tweet?

The more we have to assume about these two questions, then the “darker” a retweet becomes. For example, orphan retweets, as described in Section 3.5, are assumed to be retweets due to Twitter API’s metadata that claims this to be true. However, the originating tweets no longer exist, therefore violating Assumption #1.

In the case of copied tweets with no attributions or retweet markers, they may not be considered as retweets because there is no evidence that suggests the existence of an originating tweet that is being referred to (which relates to Assumption #1), nor any identifying information of an originating author (which relates to Assumption #2). However, there have been several studies which have documented the existence of tweets which propagate across Twitter without using retweet markers nor giving proper attribution to originating authors [17,9]. In Table 1, retweets which are not explicitly marked as retweets are considered as dark retweets. The darkest retweets are shown by the darkest-shaded rows, where the Explicit, Link to Original and Mentions Other Users columns are all marked with 0s.

Replies are also considered within the context of dark retweets, because a tweet reply is still an explicit way of propagating tweets, albeit without using conventional retweeting mechanisms. Manual inspections of URLs which return lots of replies seem to show that there are authors who send out multiple replies of similar tweets, rather than making a general retweet. The advantage of this is that a user is able to send a tweet directly to a non-follower. Therefore in the context of tweet propagation, replies are also included in this study. In this paper, tweets are considered as dark retweets depending on its degree of “darkness” compared to other conventional retweets.

Referring to Table 1, the darkest shade of grey denotes that both Assumptions 1 and 2 are difficult to presume within these types of tweets, whilst the second darkest shade denotes that only one of the two assumptions are difficult to presume. However, for brevity, this paper combines the rows coloured by the two darkest shades of grey into one category which represents Dark Retweets.

The following section outlines the proportion of visible, dark and orphan retweets, based on the shades of grey illustrated in Table [1](#).

5 Matrix Evaluation

Based on the matrix that was described in Section [3](#), a study was done to evaluate the proportions of tweets which fell within the different categories. The objective of this evaluation was mainly to deduce the extent of dark retweets, involving tweets which propagate without conforming to conventional retweeting methods. These dark retweets may be missed out by existing research of tweet propagation which focus only on retweets made by proprietary Twitter methods or manual insertions of retweet markers in tweet texts.

The hypothesis of this experiment is that dark retweets do exist, and therefore suggests that existing studies on tweet propagation may be missing some further hidden data.

5.1 Experimental Setup

An experiment was run over 2,348,936 (2.3 million) tweets, spanning over 49 days (12/05/12–29/06/12). These tweets were collected at random from Twitter’s Streaming API on its Spritzer setting, which provides a random sample containing 1% of current tweets being published globally in real time.

The dataset used in this study only focused on tweets which had a URL in them. This approach is similar to the one used by Wu et. al [\[7\]](#), as their dataset had a similar restriction as well. To get these tweets, the Streaming API was used to collect random URLs. Then, all the tweets that contained each of these URLs were collected using the Twitter Search API. Using the search results, the tweets are then classified by querying the Twitter REST API for more details of each tweet. This includes the follower/following information for each particular Twitter user that gets seen within these collected tweets.

A suite of Python scripts were created to perform two main tasks; data collection and data processing. During data collection, the scripts access Twitter’s real-time Streaming API and collect unique URLs from random tweets. They then use Twitter’s Search API to collect all the tweets containing each of the unique URLs collected. Due to Twitter’s rate-limiting policy, the search results were limited to the most recent 1500 tweets, or tweets published in the last 7 days if the 1500 limit did not get reached.

During the experiment, several URLs returned exactly 1500 search results via Twitter’s Search API. These URLs were classified as URLs which may have more than 1500 search results, but further results could not be retrieved due to Twitter’s rate-limiting policy. Section [5.3](#) will discuss in more detail the ramifications of these limits with respect to the outcomes of this research.

5.2 Evaluation Results

Based on the matrix of tweet propagation as described in Section [3](#), an experiment was run to observe the proportions of visible, dark and orphan retweets

in the above dataset of 2.3 million tweets. Out of this dataset, 820,318 (34.92%) were classified as either visible, dark or orphan retweets. For search results less than 1500 tweets per URL, 15,840 URLs were analyzed. This total increased to 16,976 URLs when the other rate-limited URLs were included.

The results are shown in this paper in the form of pie charts (Figure 1) and line graphs (Figure 2). As shown in Figure 1, Visible Rts form the biggest proportion of retweets overall, but interestingly the proportion changes when search results from rate-limited URLs were included. When the dataset is restricted to search results less than 1500 tweets per URL, dark retweets only accounted for 12.86% of all retweets, but this value increased to 24.7% when the restriction did not apply.

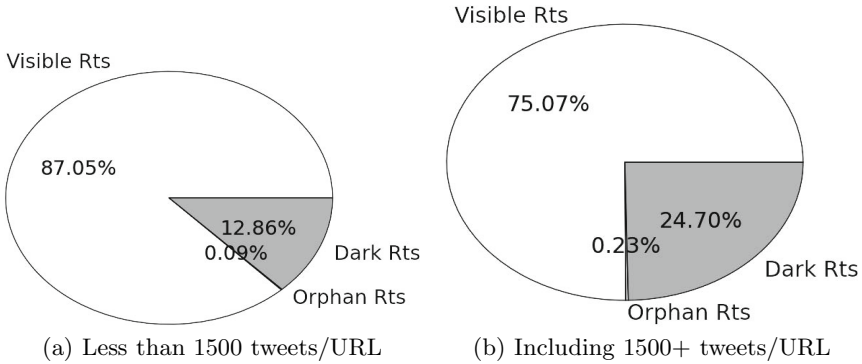


Fig. 1. Pie charts of visible, orphan and dark Rts according to size of search results (total tweets) per URL

Table 2 shows the proportions of visible, dark and orphan tweets as found in Figure 1. This table contains more than the 18 categories outlined in Table 1 because the categories of RtF and RtnF are split into visible and dark groupings. Table 2 shows that these dark RtF and RtnF categories accounted for 0% tweets. Further tweet content analysis would be helpful in capturing these two categories.

For the purpose of this paper, only the URLs with up to 1500 search results were considered. Section 5.3 will later present the reasoning for this approach, in addition to further discussions of the impact of Twitter Search API's rate limit on findings such as the above. Figure 2 shows line graphs which plot the proportions of visible, dark and orphan retweets over the total tweets found per URL. These graphs appear to show a positive correlation between total tweets and the proportions of both visible and dark retweets. However, there seems to be no strong correlation between total tweets and orphan retweets. To verify these observations, the Kendall tau-b correlation co-efficients, represented by τ_b , were calculated for the relationships between visible, dark and orphan retweets against the total tweets found per URL. A value of 1 for τ_b would signify a perfect positive correlation between two variables. A value of -1 would signify a similar perfect association but for a negative correlation. A value of 0 would denote that no correlation was found. Calculations made showed that at $\tau_b = 0.739$, visible

Table 2. Proportions of Tweets Found Based on Table 1

| Tweets | | All % | Rts only % |
|---------|----------|-------|------------|
| Ori | | 65.08 | - |
| Visible | PRtF | 12.97 | 37.14 |
| | PRtnF | 8.15 | 23.33 |
| | Rt@F | 0.00 | 0.00 |
| | Rt@nF | 0.13 | 0.37 |
| | RtF | 3.11 | 8.90 |
| | RtnF | 1.56 | 4.46 |
| | P@RtF | 0.00 | 0.00 |
| | P@RtnF | 0.08 | 0.23 |
| | @RtF | 0.15 | 0.42 |
| @RtnF | 0.08 | 0.23 | |
| Dark | RtF_d | 0.00 | 0.00 |
| | RtnF_d | 0.00 | 0.00 |
| | P@nF | 3.82 | 10.93 |
| | P@F | 0.00 | 0.00 |
| | @F | 1.64 | 4.71 |
| | @nF | 3.16 | 9.06 |
| Orphan | OrphanRt | 0.00 | 0.00 |
| | Orphan@ | 0.08 | 0.23 |

retweets seem to have a largely positive correlation against total tweets per URL. Meanwhile, dark retweets have a moderately positive correlation against total tweets per URL ($\tau_b = 0.453$), while orphan retweets have a very small positive correlation against total tweets per URL ($\tau_b = 0.120$).

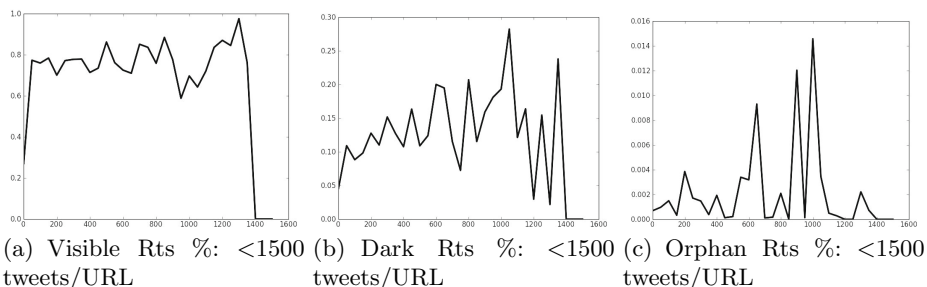


Fig. 2. Proportions of visible, orphan and dark Rts according to size of search results (total tweets) per URL

5.3 Twitter Search API’s Limit of 1500 Tweets

As explained in Section 5.2, two distinct findings emerged when the dataset used was separated into two, namely search results which were not rate-limited, and

those which included them. The proportions of dark retweets for search results less than 1500 tweets per URL amounted to 12.86% of all retweets. However, this value increases to 24.7% when rate-limited URLs were also taken into account.

The findings for URLs that were rate-limited were not focused upon in this paper. This is because we could only see an incomplete subset of the tweets which would have included these URLs. This limitation means that it was more likely for us to underestimate the proportion of retweets that were found. This is true particularly for dark retweets, which could only be detected if the complete timeline of all related tweets were found. Only then could the tweets be ordered based on each tweet's timestamps. Since the ability to see complete timelines is compromised by the 1500 search limit, therefore it is possible that we would underestimate the amount of dark retweets found, and overestimate the amount of original tweets found. Therefore, any estimation for dark retweets would be a conservative estimate.

Since the proportion of dark retweets seems to rise substantially when the rate-limited URLs are included (from 12.86% of all retweets to 24.7%), then this seems to suggest that dark retweets are more prevalent amongst rate-limited URLs. On manual inspection, it seems that the search results for rate-limited URLs tend to contain tweets published by spambots, or autosenders tied to web services. The importance of recording dark retweets may be slightly diminished when they are considered to be more popular amongst spambots. However, it would be difficult to automatically discount all tweets containing rate-limited URLs as spam.

5.4 Impact on Existing Research

The existence of dark retweets may impact the findings of other existing studies on tweet propagation. In this paper, 12.86% of dark retweets were found for URLs which returned up to 1500 search results each. Table 3 shows the extrapolation of our findings with two other papers, namely those made by Suh et. al [10] and Cha et. al [2]. These figures were derived by matching the proportions found in these papers with the findings as presented in this paper.

The main difference between the work by Suh et. al [10] and this paper is that the dataset used in our research is restricted to tweets containing URLs only. Therefore if we were to extrapolate the proportions of retweets with URLs in them, as found by Suh et. al [10], then 0.47%, or 347,800 tweets, would potentially be hidden from their dataset.

This extrapolation becomes more difficult to make when the retweets dataset is not restricted to tweets containing URLs only, such as the case with Cha et. al's study [2]. However, in their paper, it was claimed that 92% of the retweets they found contained URLs in them. The proportion of retweets that they found was not given in their paper, therefore we could only extrapolate the proportion of retweets with URLs as stated by them, amounting to 92% of all retweets. From this percentage, 13.6% of those retweets could be extrapolated as dark retweets, therefore potentially hidden from Cha et. al's evaluation.

There are various studies which use retweets as the basis of their studies, but considering that our studies have restricted the dataset that we used to tweets containing URLs only, then it is not possible to be certain that any extrapolations will hold.

Table 3. Extrapolation of our Findings with Existing Research

| Studies | | Suh | Cha |
|-----------------------------|------------------|-------------------|-------------------|
| Existing work | Dataset size | 74 million | 1.76 billion |
| | Original tweets | 88.85% | - |
| | All Rts | 11.15% | - |
| | Rts with URLs | 28.4% of retweets | 92% of retweets |
| Extrapolation from our work | Original tweets | 96.36% | - |
| | Visible retweets | 3.17% | 92% of retweets |
| | Dark Retweets | 0.47% | 13.6% of retweets |
| | Hidden Retweets | 347,800 | - |

The impact of this hidden data is the current focus of ongoing research work. One such focus is on the perceived reach of a retweet. In the work done by Kwak et. al [11], it was claimed that an average retweet could reach 1000 users, irrespective of who the originating author is. Considering the different visibility properties of dark retweets, their existence may or may not affect this number. Further research could reveal the actual impact of dark retweets on tweet propagation studies such as the above.

6 Conclusion

This paper has presented a comprehensive matrix of tweet propagation, which was deconstructed into seven different characteristics, namely proprietary, mechanism, follower/non-follower, mentions other users, explicit, links to original tweet, and push audience. The paper has discussed the different nuances of tweet propagation, such as the availability of metadata linking to originating tweets. The different visibilities of retweets and replies also meant that the reach of these tweets were different. In addition, several retweets seemed to be using multiple retweeting mechanisms, which impacts the total reach of their tweets. Orphan retweets and replies also exist due to the subsequent deletion of originating tweets and users at a later time. The concept of dark retweets was introduced, based on two assumptions of retweetability pertaining to a retweet’s provenance; an originating tweet exists and it was made by an originating author.

The experiment that was run to evaluate the matrix of tweet propagation showed that over 2.3 million tweets, dark retweets amounted to 12.86% (for search results less than 1500 tweets per URL) and 24.7% (for search results including more than 1500 tweets per URL) respectively. By extrapolating the results found in several existing papers [10,2], the potential hidden data as a result of dark retweets amounted to between 0.47–13.6%.

Several threads of future work is planned for this research, such as conducting a deeper statistical analysis on the frequencies found in Table 2, by looking at the stability of percentages found over time. This is in addition to identifying other external factors impacting these percentages, such as location, working hours, external offline events, etc. Another future work thread involves quantifying the actual impact of dark retweet detection on existing studies on tweet propagation.

References

1. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: Hawaii International Conference on System Sciences, pp. 1–10. IEEE Computer Society Press, Los Alamitos (2010)
2. Cha, M., Haddadi, H., Benevenuto, F., Gunmadi, K.P.: Measuring user influence in Twitter: The million follower fallacy. In: International AAAI Conference on Weblogs and Social Media Fourth International AAAI Conference on Weblogs and Social Media. AAAI (2010)
3. Mustafaraj, E., Metaxas, P.: From obscurity to prominence in minutes: Political speech and real-time search. In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, NC: US (2010)
4. Adar, E., Zhang, L., Adamic, L.A., Lukose, R.M.: Implicit structure and the dynamics of blogspace. In: Workshop on the Weblogging Ecosystem, WWW 2004, New York, NY (2004)
5. Matsumura, N., Yamamoto, H., Tomozawa, D.: Finding influencers and consumer insights in the blogosphere. In: International AAAI Conference on Weblogs and Social Media Fourth International AAAI Conference on Weblogs and Social Media. AAAI (2010)
6. Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., Kellerer, W.: Outtweeting the twitterers- predicting information cascades in microblogs. In: 3rd Workshop on Online Social Networks, Boston, MA, USA. USENIX (2010)
7. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW 2011). ACM (2011)
8. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Identifying ‘influencers’ on Twitter. In: Fourth ACM International Conference on Web Search and Data Mining. ACM (2011)
9. Nagarajan, M., Purohit, H., Sheth, A.: A qualitative examination of topical tweet and retweet practices. In: International AAAI Conference on Weblogs and Social Media Fourth International AAAI Conference on Weblogs and Social Media. AAAI (2010)
10. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: IEEE Second International Conference on Social Computing, pp. 177–184. IEEE (2010)
11. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA, pp. 591–600. ACM (2010)

Studying Paths of Participation in Viral Diffusion Process

Jarosław Jankowski¹, Sylwia Ciuberek², Anita Zbieg³,
and Radosław Michalski²

¹ Faculty of Computer Science, West Pomeranian University of Technology,
Szczecin, Poland

`jjankowski@wi.zut.edu.pl`

² Institute of Informatics, Wrocław University of Technology, Poland
{`sylwia.ciuberek, radoslaw.michalski`}@pwr.wroc.pl

³ Institute of Psychology, University of Wrocław,
Wrocław University of Economics, Poland

`anita.zbieg@gmail.com`

Abstract. Authors propose a conceptual model of participation in viral diffusion process composed of four stages: awareness, infection, engagement and action. To verify the model it has been applied and studied in the virtual social chat environment settings. The study investigates the behavioural paths of actions that reflect the stages of participation in the diffusion and presents shortcuts, that lead to the final action – the attendance in a virtual event. The results show that the participation in each stage of the process increases the probability of reaching the final action. Nevertheless, the majority of users involved in the virtual event did not go through each stage of the process but followed the shortcuts. That suggests that the viral diffusion process is not necessarily a linear sequence of human actions but rather a dynamic system.

Keywords: information diffusion, online social networks, participation model, multistage analysis.

1 Introduction

The studies that direct attention to diffusion of innovation process [25], [7], [32], social influence mechanism [10], [21], [24], [12], [23] social contagion and epidemics outbreaks [5], [6] or cascades of influence patterns [34] investigate a similar phenomenon: a propagation, transmission and adoption of information (content, opinions, behaviours, emotions) within a network of social relations. As the information generated and shared online is gaining steadily in importance [28], more and more researchers are trying to deal with the power of electronic contagions. They are especially interested in social networking sites [26], [9] being recently the most popular online activity that has outnumbered e-mail actions [20], [1]. Because there is a great need to understand mechanisms and factors crucial for the spread of information, researchers search for new ways on how to study the phenomenon. The results from research areas related to dynamics and mechanisms of social transmission and

adoption are successfully adopted to word of mouth process investigation [11], [33], [22] and viral seeding strategies examination [17], [4]. Nevertheless, the empirical network studies seem to be promising, relatively little has been done in this area [31].

The research presented in this paper is targeted to online social platforms with the ability to capture different forms of users' behaviours: communications, activity and transfers among users. Most of the research in the field of information diffusion and viral contagion is addressed to participants' characteristics, the structures of network they are embedded in or/and characteristics of information that is transmitted. Rare are studies in which attention is paid directly to the behaviours. The main motivation in the current research is to observe human action systems in more detailed way -by analysing different forms of behaviours related directly and indirectly to viral campaign and stages of the participation for the viral action. The study specifies several social behaviours related to the diffusion process and identifies several stages of participation in viral diffusion, starting from activity before receiving or sending viral information and behaviours after infection towards reaching a final diffusion goal.

2 Related Work

The process of information contagion among individuals and their further participation in particular viral actions can be observed and studied in three-folds[22]. Researchers focus on personal characteristics of people engaged in social contagion, their needs and motivations. Vital for the diffusion process are also social factors such as the characteristics of other people influencing an individual, attributes of the channels through which information flows and attributes of social system in which individuals operate. And finally, researchers consider the characteristics of spread information to be important for the social contagion.

2.1 Personal Characteristics and Stages of Adoption

Under the study there are personal characteristics of an individual who passes the message further as well as the one who is exposed to the message. What matters for the virus propagation are the personality traits like extraversion and innovativeness [4], authority of the sender, activity of the receiver [36] and similarity of sender and receiver demographics traits [29]. Moreover, people share information motivated by the need to be part of a group, but the need to be individualistic and stand out from a crowd is reported as a second reason [18]. Vital for virus propagation is also the need to be altruistic and the need for personal growth [18].

Multistage Models of Engagement

The change of individual's opinion or behaviour is studied as the process of perceptual adaptation to a new stimuli, e.g. information, opinion, product or technology. The behavioural success of adoption depends on the cognitive processes that engage individual's attention as well as on personal motivations and emotions that lead to the

interest. A few stages of the adoption have been distinguished [25], [8]. To adopt the received message, an individual first needs to pay attention or simply identify and notice information which is called the *awareness stage*. In the next stage of adoption (*interest stage*) any interest about the news is required. Interest leads to the engagement and active learning about the news by e.g. studying, searching, discussing, using or sharing it. Finally, the necessary information about the news, the decision of adoption or rejection is made (final *decision stage*). The process was first reflected in the theory originally developed by Rogers [25] and successfully adopted and simplified to WOM marketing e-mail contagion research in the form presented above [8]. It has also been used as the background for the model and computer simulation of the decision to participate in viral marketing campaigns [33].

2.2 Social Factors and Mechanisms of Adoption

Nevertheless, the change of behaviour is preceded by intra-individual cognitive process, the perception and behaviour of a person is also influenced by inter-individual social context a person is embedded in. The innovation is communicated among the peers, friends, acquaintances and other members of a larger social group. As Rogers defines diffusion as "the communication of an innovation over time through certain channels among a social system" [25], the innovation is adopted due to interpersonal interaction within the social network.

Social Influence, Imitation and Social Learning

Social learning theory [3] explains how people learn within a social context and assumes that the behaviour of an individual is influenced by other peoples' behaviour as well as by the personal characteristics of an individual. Moreover, behaviour of an individual influences the behaviour of others in a similar way that others influence an individual, what is called the reciprocal determinism. For the change of behaviour some important stages are required. First, an individual while observing a new behaviour needs to pay attention to characteristics of the behaviour to be able to imitate it. Next, the retention stage is needed, and this means that an individual is able to remember details of the behaviour. It is also possible for an individual to reproduce or imitate the behaviour and organize their own responses in accordance with the observed behaviour of others, which usually improves with practice. To do so, an individual must have the motivation or some other incentives [3]. Without any motivation, even if the previous stages are present, no imitation occurs. On the other hand, the social influence phenomenon occurs when people intentionally and directly influence others, being their friends, co-workers, family, acquaintances, etc. When other people act on the targeted individual, the closer they are to a person (physically or psychologically) and the more they are - the strongest social influence occurs [21], [24]. Individuals have several motivations to follow other people. Social diffusion operates through spreading awareness and interest for an information, social learning about the benefits and risks or normative influence extending the validity and

legitimacy of the news. Informative influence occurs when there is lack of time or lack of other sources of knowledge and when the concerns that not adopting (when majority already adopted) can result in some disadvantages [30].

The Number of Connections and the Number of Adopting Friends

A two-step flow model of communications [19] considered a small group of people called influentials as important for social influence and diffusion process, as they directly influence many neighbours. Influentials can be people with many connections occupying a central position in the social network [17], [15]. However, there are studies reporting that the viral content received from individuals with many ties have bigger chance to be ignored [22], [35], [13]. There is also a field of discussion if people with many connections are more or less susceptible in influencing other people. As they participate in a network more frequently, they are more often exposed to anything that flows through the network including a virus or a viral content [17], [6]. On the other hand, individuals with many friends were observed as less likely to be influenced by others and this is what researchers explain as generally weaker tie strength formed by highly connected individuals [2].

Characteristics of Relations and Social Structures

Strength of the tie that connects two individuals is one of the most important characteristic having impact on the flow of information [22], [17], [4], [36]. Tie strength can be reflected by the number of common friends or triples formed by two individuals. Consumers are more likely to open e-mail messages sent from a person that they feel close to [8], [22] and stronger ties can increase the likelihood that the message will be passed along to others [26], [22]. Strong ties built on time spent together, emotional intensity, intimacy and reciprocal services between people [16], facilitate the flow of information. However, the weak ties serve as bridges creating short paths in the network [16] and allow us to reach a larger number of people in the network and to traverse greater social distance.

2.3 Characteristics of Transmitted Information

Vital for the awareness, usage and transmission of information is its relevance to the preferences of the receiver [22]. In Second Life community study [2], for the contagion it was vital, if the asset was popular or niche, as both types of assets were spreading through the network differently.

2.4 Limitation of Earlier Approaches and Motivation

Majority of the presented studies concentrate on characteristics of people engaged in the diffusion process, channels and structures through which an information is transmitted and the characteristics of the information. Relatively small work is addressed to the behaviours of participants and social stimuli that directly or indirectly come from behaviour of other people engaged in the diffusion process. The goal of

proposed model is to focus on those behaviours. Authors take into account micro (personal decision) and macro (social stimuli) perspective of human action in the context of viral diffusion process. The study captures participants' activity before infection, actions related to information spreading, activity after infection and reaching the goal of diffusion which in our case is the participation in the event.

While multistage models of participation in the diffusion of information attempt to fill the research gap and investigate participants behaviours, the observations of this kind focus mainly on the behavioural paths previously assumed theoretically and ignore other paths that can occur. The authors propose and verify the conceptual model of participation in the diffusion process, and the study is not limited to the observations of behavioural paths assumed in the model. The present work attempts to observe all possible behavioural paths that reflect the stages of engagement in the diffusion process and presents as well some behavioural shortcuts that lead to the main information diffusion goal which is the decision to participate in the event.

3 Conceptual Framework

The model is targeted to virtual avatar chat world where the interactions among communicating users are more similar to a real world environment. Virtual versions of real goods can be used and exchanged. Interest can be built for example by observing goods possessed by other users or by messaging the information. In Fig. 1, the exemplary stages of communication based on viral content and typical activities in the system are shown.

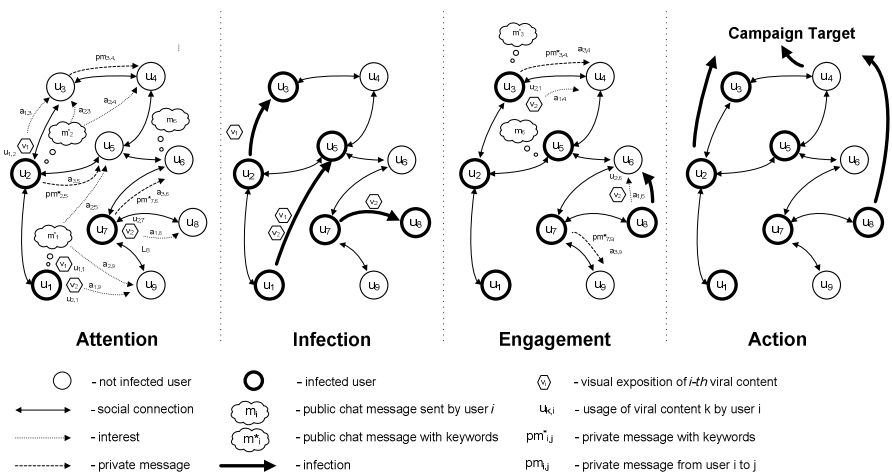


Fig. 1. Exemplary viral campaign in the online environment

Among users $U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8,u_9\}$ logged to the system, communication channels with messages m , pm may be identified. At the *awareness stage*, users are exposed to different ways of building attention to viral campaign integrated with chat

messages system. Both public and private messages can contain keywords related directly to viral action and are denoted as m^* and pm^* respectively. Communication containing keywords related directly to the campaign may build interest in action apart from the main viral visual components v_1 and v_2 . Users u_1, u_2, u_7 are infected earlier and they have visual components in their repository. User u_1 has the set of two viral components $C_1=\{v_1, v_2\}$, assigned user u_2 has one component $C_2=\{v_1\}$ and user u_7 has one component $C_7=\{v_2\}$. Messages containing keywords and visual components related to the campaign may be exposed to other users, building their interest and demonstrating the engagement in action. At this stage there are factors building attention for the user: using visual objects $a_{1,i}$, public messages containing keywords $a_{2,i}$ and private messages with keywords $a_{3,i}$. At the *infection stage* users are infected with viral content: user u_3 receives content v_1 from user u_2 , user u_5 receives content v_1 and v_2 from u_1 , and user u_8 receives content v_2 from u_7 . At the *engagement stage* infected users can engage in propagating information about campaign and increase their own interest in virus. Newly infected user u_3 exposes content v_2 to user u_4 increasing the current level of interest and sending private message $pm^*_{3,4}$ containing keyword used in the campaign. User u_8 exposes content v_2 to user u_6 and this results with the infection. User u_5 is not engaged in spreading the information about viral content and he/she sends public messages without any viral content. Finally, at Stage 4 (*action*) some of the infected users make up their decision and they reach the campaign target which can be defined as interaction required by the campaign organizer. Usually, only a fraction of users will reach this stage. In the exemplary process users u_2 and u_8 infected earlier with viral content, and user u_4 who was not infected with viral object but was exposed to the messages containing information about campaign, decided to perform the expected action. The example presented typical stages of communication from multiplayer online systems like games and virtual worlds. To generalize the approach, authors identified the characteristics of each phase of communication and characterized the process of participation in the campaign as composed of four stages: *attention*, *infection*, *engagement* and *action* presented in Fig. 2.

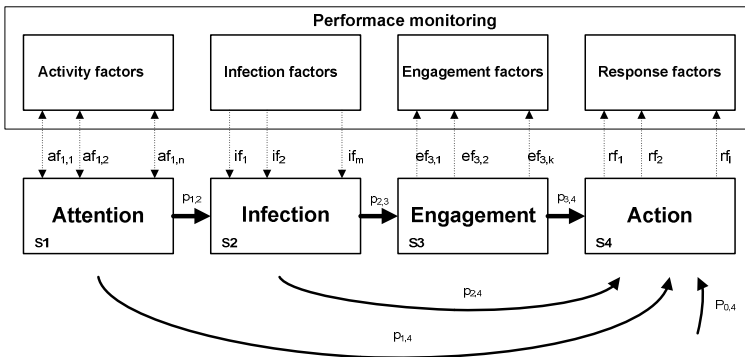


Fig. 2. A multistage model for viral campaign in online environment

At each stage there were factors specific to the particular stage. At the first stage related to building attention, a set of activity factors before infection $AF=\{af_{1,1},af_{1,2},\dots,af_{1,n}\}$ is measured. At this stage, users can send and receive both public and private messages. Interest in action is measured as a ratio of messages containing keywords to all messages. Some of the users from the first stage can be infected with different components of viral action at the second stage. And at this stage infection factors can be measured from set $IF=\{if_{1,1},if_{1,2},\dots,if_{1,n}\}$. After the infection took place, users can continue to engage themselves in the action but apart from messages with keywords they can spread viruses and use visual components to build the attention among other users. Factors measured at this level are included in a set $EF=\{ef_{1,1},ef_{1,2},\dots,ef_{1,n}\}$. At the last stage response factors showing engagement in relation to the main goal of campaign may be identified. A set of response factors during campaign is denoted by $RF=\{rf_1,rf_2,\dots,rf_k\}$. Apart from users engaged in the campaign and viral transmission, the activity of users that were not infected by viral content is captured, as they "can go" to last stage without earlier infection. To gather information about users' activity, the transition probabilities between stages may be computed. Probability $p_{2,4}$ represents the probability of response without engagement after infection, $p_{3,4}$ represents the probability of reaching the goal of campaign after engaging in the spread of information about the campaign using different forms of activities. The $p_{0,4}$ denotes the probability of reaching the campaign goal without infection and no interest revealed in the campaign. Between stages the probabilities are measured as $p_{1,2}$ and $p_{2,3}$. Users do not necessarily need to take actively part in all stages of the campaign to perform the required action and in our case attend the event. In the system, the reference factor showing typical activity may be defined. Reference activity shows the typical action in the system as well as the overall activity in the period analyzed like number of logins, time spent in system or communication activity. The defined measures can be used for the evaluation of viral campaign not only in terms of viral spread but also as engagement in different stages of campaign, thus comparing different approaches. In the next step, the empirical research based on the proposed approach is presented.

4 Empirical Results

To verify the proposed approach, authors used the dataset from online platform that connects functionality of multiplayer system and virtual world was used. The system combines function of chat and the entertainment platform where users are represented by graphical avatars. They have the opportunity to engage in the life of online community and to perform actions in different fields. Users have the access to virtual objects like avatars, clothes, different characters and special effects that can be distributed among users thought the viral transmission. The dataset was collected during the event with the main goal to motivate users to attend meeting in the system related to the virtual protest. The action was started five days prior to the event and two

visual viral elements were used in a form of avatar face and informative object holding in avatar hand, which was an expression of a pro-protest attitude. The keywords related directly to viral action were identified. In the period t there is the total number of 4910 unique users logged into the system. The viral content was initially delivered to 16 users selected from the most active group of protesters which became seeds for viral transmissions. The only way to obtain the items was by receiving it from other users.

To measure the campaign performance, a designed tracking system was used to capture all the actions related to the campaign by assigning them to the specific stages of campaign. The viral components were designed in a form of digital goods assigned to the user inventory. In the inventory, users can collect different elements of avatars or objects with special functionalities. After the viral infection, the digital object is added to the receiver inventory and it can be used in the visual chat environment. During the infection, the object is duplicated and a copy of the object is delivered to the receiver as well as stays at the sender's account. The tracking system was measuring the events related to sending objects and using the object at different stages of the process. The usage of the object was treated as an engagement and was represented by putting a mask on the avatar or using the transparent. Each time the mask or transparent was used, the system was saving the related actions with the assigned timestamps. Sequences of infections were used to build connection and to observe the diffusion of digital objects. Moreover, the tracking system was capturing keywords related to the campaign and assigning them as an additional measure of engagement. Also the information about users entering the dedicated chat room was collected. The obtained data was anonymized and no personal information about participants was used.

Among logged users 324 of them received viral content and 134 of them decided to send assets to friends which makes 41% of receivers engaged in forwarding messages. The monitoring of participation at each stage was based on the messages, usage of viral components and sending the viral content. The results showed that from all infected users 128 (39%) reached the goal of campaign and in the fifth day participated in the event. The event was visited by 197 users not infected directly by visual components. It shows that other factors affecting decisions and attitudes exist and not necessarily directly related to main viral campaign. The next part of the analysis was performed by using participation factors for the proposed multistage approach. The analysis was performed for activity within five days frame. To monitor engagement at this stage, factors showing communication activity containing keywords in relation to all messages sent by user during the monitoring period were defined. As observed in the campaign, most of the users' activity was recorded in the system and it was possible to track the activity before the infections. In the stage S1 (*awareness stage*) activity prior to infections is included, S2 (*infection stage*) contains activity related to infections and stage S3 (*engagement stage*) contains activity after infection prior to main event. Within stage S4 (*actions stage*) factors representing visiting event and activity during event were identified. The probabilities of transitions between stages were computed. Table 1 shows identified activity factors for all stages.

Table 1. Activity factors measured during campaign

| <i>Factor</i> | <i>Description</i> |
|----------------------------|--|
| S1:af_{1,1} | Messages containing keywords sent by non-infected users before event divided by all sent messages |
| S1:af_{1,2} | Messages with keywords received before event by non-infected users divided by number of messages received in this period |
| S1:af_{1,3} | Messages containing keywords received by infected users before infection divided by all messages prior to infection |
| S1:af_{1,4} | Messages containing keywords sent before infection by infected users divided by all messages prior to infection |
| S2:if_{2,1} | Viral component received |
| S2:if_{2,2} | Viral component sent |
| S3:ef_{3,1} | Messages sent with the keyword after infection before event divided by all messages sent in this period |
| S3:ef_{3,2} | Messages with keywords received after infection before event divided by number of messages received in this period |
| S3:ef_{3,3} | Viral component used after infection before event divided by all messages sent by user |
| S4:rf₁ | Visit to event |
| S4:rf₂ | Number of messages sent during event by infected users with keywords divided by all messages sent during event |
| S4:rf₃ | Number of messages sent during event by non-infected users with keywords divided by all messages sent during event |
| S4:rf₄ | Number of times viral component was used during event divided by number of messages sent by user during event |

Among all logged users 1298 communicated the keywords and 245 of them were infected ($p_{1,2}= 0.19$). 152 users after infection sent messages containing keywords($p_{2,3}= 0.47$). 73 of the users from Stage 3 visited the event, resulting in $p_{3,4}=0.48$. A number of 1021 users among not infected used keywords in messages and 118 of them visited the event ($p_{1,4}=0.12$). 106 users after infections did not use keywords in messages and 44 of them visited the event, resulting in $p_{2,4}=0.41$. 125 users after infection sent messages with keywords and 61 of them visited the event ($p_{3,4}=0.49$).3604 users (logged to system during five days)did not use keywords in messages and 79 of them visited event ($p_0=0.022$). Fig. 3shows the probabilities assigned to each stage of the process.

The transitions between the stages show that reaching each level of the process increases the probability of attending the main event. The probability of attending the event was increased from $p_0=0.022$ to $p_{1,4}=0.12$ if the engagement in sending messages was observed even without infection. Receiving the viral content increased the probability of reaching the final stage to $p_{2,4}=0.41$ even if no engagement was observed after the infection. It shows that the viral content was affecting the decision to join event and the probability increased after engagement and infection to $p_{3,4}=0.48$.

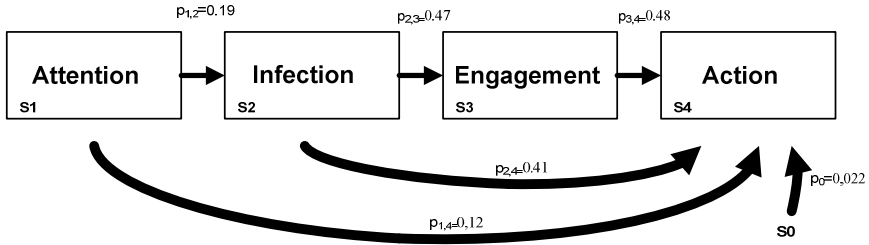


Fig. 3. Probabilities assigned to moving between stages

The differences between users of group G1 visiting the event and group G2 not visiting the event among all users logged to the system within the monitored period were found. A number of 4585 users did not visit the event while 325 visited. Table 2 shows the analysis based on the Mann-Whitney significance test.

Table 2. Differences between visitors and non-visitors for group G1 and G2

| Factor | Rank Sum | Rank Sum | U | Z | p-value | Valid N | Valid N |
|-------------------|----------|----------|----------|----------|----------|---------|---------|
| if _{2,1} | 1069235 | 10987270 | 473865.0 | 10.98194 | 0.000000 | 325 | 4585 |
| ef _{3,2} | 1011660 | 11044846 | 531440.5 | 8.65046 | 0.000000 | 325 | 4585 |
| af _{1,4} | 984894 | 11071612 | 558206.5 | 7.56659 | 0.000000 | 325 | 4585 |
| ef _{3,3} | 982105 | 11074400 | 560995.0 | 7.45367 | 0.000000 | 325 | 4585 |
| ef _{3,1} | 964057 | 11092449 | 579043.5 | 6.72281 | 0.000000 | 325 | 4585 |
| if _{2,2} | 940898 | 11115608 | 602202.5 | 5.78500 | 0.000000 | 325 | 4585 |
| af _{1,3} | 902094 | 11154412 | 641006.5 | 4.21366 | 0.000025 | 325 | 4585 |
| af _{1,1} | 880253 | 11176252 | 662847.0 | 3.32924 | 0.000871 | 325 | 4585 |
| af _{1,2} | 840438 | 11216067 | 702662.0 | 1.71696 | 0.085987 | 325 | 4585 |

Significant factors were related to the messages: received messages with keywords related to event after infection $ef_{3,2}$ and sending messages with keywords $af_{1,4}$ before infection. Before infection users interested in action were discussing about it and keywords related to event were used at this stage. Analysis show that engagement at early stage was resulting in higher interest to the main action. The next important factor $ef_{3,3}$ is showing that users in group G1 were more actively using viral components after the infection. The engagement in the event and factors affecting intensity of messages containing keywords during the event were also analysed. The results based on multiple regression are showed in Table 3.

Table 3. Factors affecting engagement during event

| <i>Factor</i> | <i>b*</i> | <i>Std. Err.</i> | <i>b</i> | <i>Std. Err.</i> | <i>t(315)</i> | <i>p-value</i> |
|-------------------------|-----------------|------------------|-----------------|------------------|-----------------|-----------------|
| Intercept | | | 0.067661 | 0.014063 | 4.811124 | 0.000012 |
| ef_{3,3} | 0.000578 | 0.062929 | 0.000953 | 0.103667 | 0.009191 | 0.992673 |
| ef_{3,2} | -0.001277 | 0.063362 | -0.001440 | 0.071463 | -0.020152 | 0.983935 |
| ef_{3,1} | 0.070580 | 0.054658 | 0.019943 | 0.015444 | 1.291298 | 0.197547 |
| if_{2,2} | 0.012411 | 0.062075 | 0.000573 | 0.002867 | 0.199935 | 0.841661 |
| if_{2,1} | 0.092130 | 0.067792 | 0.010677 | 0.007856 | 1.359003 | 0.175118 |
| af_{1,4} | 0.295254 | 0.059231 | 0.615257 | 0.123426 | 4.984810 | 0.000023 |
| af_{1,3} | -0.052056 | 0.053760 | -0.123527 | 0.127568 | -0.968319 | 0.333628 |
| af_{1,2} | -0.045759 | 0.050451 | -0.508764 | 0.560928 | -0.907004 | 0.365098 |
| af_{1,1} | 0.316411 | 0.050446 | 0.906451 | 0.144517 | 6.272283 | 0.000012 |

Two from measured factors resulted as statistically significant. As the most important factor $af_{1,1}$ was identified, representing the number of messages with the keywords sent by non-infected users in relation to all messages before the event. The second important factor was $af_{1,4}$ showing the engagement in the distribution of messages with the keywords by infected users before the infection. In the next stage, multiple regression analysis was performed only using infected users and their activity during the event. As factors affecting the engagement the results showed $af_{1,4}$ and $ef_{3,1}$, representing the engagement in the action before the event. The engagement before the infection was more important than after the infection. And finally, no statistically significant factors were identified as affecting af_4 and the usage of visual components during the event. This result confirms that users engaged in the first stages of the process were more willing to engage in the final event.

5 Discussion

The performed experimental studies showed that reaching each level of the diffusion process increased the probability of attending the main event. The behaviour before the event (receiving and spreading messages and viral components) was related directly to the activity during the event. Not only the number of infections received by user was increasing its interest in action but also the number of infections sent by a user. Using viral components by a person increased the interest in action of people being receivers of the information, as well as confirms sender's interest in action as the expression of some attitude (e.g. manifesting the opinion or convincing others to it) strengthens the attitude. Interestingly, users directly interested in the action even without receiving the viral components joined it. Probably because the diffused information was not specific for the studied online platform and diffused also in a real world. Hence, the behaviour of some participants were not necessarily influenced by stimuli that directly came from other users but have their own attitude acquired outside the platform.

Tracking the steps of participation in the diffusion process, authors tried to fit users' behaviour to the AIEA model, which describes the increase of participation leading to the final stage – action. Generally, the basic intuition suggests that the users which were the most engaged would follow to the action stage. As the results showed on Figure 3, this kind of intuition might be misleading because this group represents only a small fraction of all users in the last stage. This analysis showed that there is a significant importance of other paths in the model which may be understood as shortcuts from one of the previous stages towards the last stage. For instance, a high fraction of users took part in the event who were registered previously only in the attention stage (S0 → S4). In that case all the paths should be analysed with equal importance because the analysis showed that the majority of users participating in the final event were not necessarily the most engaged ones. This leads to the conclusion that in the online environment, the action performed by users may result in bigger changes if someone was not present in all stages. The other question arises: what is the role of users who are very engaged but not participating in the event? Although someone may ignore them, their role is invaluable in terms of influencing other users – if they would be passive in terms of infecting other users, the number of users in previous stages would be significantly lower. The vital role of users in the third stage is to increase the overall interest of users into the event and, as the analysis showed, even weak engagement of users in the whole multistage process will convert to the final action.

6 Summary

Social contagion studies are usually a simplification of a real world situation where decisions related to participation in viral diffusion are based on many different factors which are difficult to observe and monitor. The current study refers to some of them: verbal communication, visual aspects of infections when viral content is visible, interest to viral content revealed prior to infection and talks about it. The proposed model attempts to merge micro and macro dynamics of human behaviour. The former reflect intra-personal process of making the successive actions by a person, the latter pay attention to the inter-actions that occur between a person and other people.

The study showed that the stages in viral diffusion process are not necessarily sequential. By studying the additional paths (shortcuts) in the users' participation processes, the analysis showed that users mostly did not followed the basic path. That leads to the finding that viral diffusion process probably is not a linear sequence of actions but rather a dynamic system, while many models of diffusion assumes the infection as a part of a sequential process and many empirical studies covers only the observation of previously defined sequence of behaviours. That of course doesn't mean that we cannot find any sequential model that describes great a social phenomenon, but that empirical studies of social diffusion conducted with the interest to the dynamics of human actions can be an interesting way of further research. The proposed model and analysis may be used in empirical studies related to social diffusion process in online environments, as the studied functionality is similar to many popular online setting and the studied phenomenon is not only relevant to social informatics but can be applied to many fields related to social contagion: word of mouth and viral

marketing, the studies of public opinion and actions, information systems and organizational research.

Acknowledgments. This work was partially supported by fellowship co-financed by the European Union within the European Social Fund, the Polish Ministry of Science and Higher Education, the research project 2010-13.

References

1. Arndt, J.: Role of Product-Related conversations in the diffusion of a new product. *Journal of Marketing Research* 4(3), 291–295 (1967)
2. Bakshy, E., Karrer, B., Adamic, L.A.: Social influence and the diffusion of user-created content. In: *Proceedings of the 10th ACM Conference on Electronic Commerce, EC 2009*, pp. 325–334. ACM, New York (2009)
3. Bandura, A.: Social cognitive theory of mass communication. In: Bryant, J., Zillman, D. (eds.) *Media effects: Advances in theory and research*, pp. 121–153. Lawrence Erlbaum Associates, Mahwah (2002)
4. Chiu, H.C., Hsieh, Y.C., Kao, Y.H., Lee, M.: The Determinants of Email Receivers' Disseminating Behaviors on the Internet. *Journal of Advertising Research* 47(3), 524–534 (2007)
5. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* 357(4), 370–379 (2007)
6. Christakis, N.A., Fowler, J.H.: Social network sensors for early detection of contagious outbreaks. *PLoS ONE* 5(9), e12 948+ (2010)
7. Coleman, J., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. *Sociometry* 20(4), 253–270 (1957)
8. De Bruyn, A., Lilien, G.L.: A multi-stage model of word-of-mouth influence through viral marketing. *International Journal of Research in Marketing* 25(3), 151–163 (2008)
9. Emarketer Press Releases: Social Network Ad Revenues to Reach \$10 Billion Worldwide in 2013 (2011)
10. Festinger, L.: Informal Social Communication. *Psychological Review* 57(5), 271–282 (1950)
11. Godes, D., Mayzlin, D.: Firm-Created Word-of-Mouth communication: Evidence from a field test. *Marketing Science* 28(4), mksc.1080.0444-739 (2009)
12. Goel, S., Mason, W., Watts, D.J.: Real and Perceived Attitude Agreement in Social Networks. *Journal of Personality and Social Psychology* 99(4), 611–621 (2010)
13. Goel, S., Watts, D.J., Goldstein, D.G.: The Structure of Online Diffusion Networks. *Forthcoming in ACM EC 2012* (2012)
14. Golan, G.J., Zaidner, L.: Creative Strategies in Viral Advertising: An Application of Taylor's Six-Segment Message Strategy Wheel. *Journal of Computer-Mediated Communication* 13(4), 959–972 (2008)
15. Goldenberg, J., Han, S., Lehmann, D.R., Hong, J.W.: The Role of Hubs in the Adoption Process. *Journal of Marketing* 73(2), 1–13 (2009)
16. Granovetter, M.: The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory* 1, 201–233 (1983)
17. Hinz, O., Skiera, B., Barrot, C., Becker, J.U.: Seeding Strategies for Viral Marketing: An Empirical Comparison. *Forthcoming in Journal of Marketing* (2012)

18. Ho, J.Y.C., Dempsey, M.: Viral Marketing: Motivations to Forward Online Content. *Journal of Business Research* 63(9/10), 1000–1006 (2010)
19. Katz, E., Lazarsfeld, P.: *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers (2005)
20. Keenan, A., Shiri, A.: Sociability and social interaction on social networking websites. *Library Review* 58(6), 438–450 (2009)
21. Latané, B.: The psychology of social impact. *American Psychologist* 36(4), 343–356 (1981)
22. Liu-Thompkins, Y.: Seeding viral content: Lessons from the diffusion of online videos. Forthcoming in *Journal of Advertising Research* (2011)
23. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 415–444 (2001)
24. Nowak, A., Szamrej, J., Latané, B., Nowak, A., Szamrej, J., Latan, B.: From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review* 97, 362–376 (1990)
25. Rogers, E.M.: *Diffusion of Innovations*, 5th edn. Free Press, New York (2003)
26. Shu-Chuan, C., Kim, Y.: Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International Journal of Advertising* 30(1), 47–75 (2011)
27. Tarde, G.: *On Communication and Social Influence: Selected Papers*, Reprint edn., Apr. 15. *Heritage of Sociology Series*. University Of Chicago Press (2011) 1st edn. (1893)
28. Thackeray, R., Neiger, B.L., Hanson, C.L., McKenzie, J.F.: Enhancing Promotional Strategies Within Social Marketing Programs: Use of Web 2.0 Social Media. *Health Promotion Practice* 9(4), 338–343 (2008)
29. Trusov, M., Bodapati, A.V., Bucklin, R.E.: Determining Influential Users in Internet Social Networks. *Journal of Marketing Research* 47(4), 643–658 (2010)
30. van den Bulte, C., Lilien, G.L.: Medical Innovation Revisited: Social Contagion versus Marketing Effort. *American Journal of Sociology* 106(5), 1409–1435 (2001)
31. van den Bulte, C., Wuyts, S.: *Social Networks and Marketing*. Marketing Science Institute, Cambridge (2007)
32. Valente, T.W.: *Network Models of the Diffusion of Innovations*. Hampton Press, NJ (1995)
33. van der Lans, R., van Bruggen, G., Eliashberg, J., Wierenga, B.: A Viral Branching Model for Predicting the Spread of Electronic Word of Mouth. *Marketing Science* 29(2), 348–365 (2010)
34. Watts, D.J.: A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99(9), 5766–5771 (2002)
35. Watts, D.J., Dodds, P.S.: Influentials, networks, and public opinion formation. *Journal of Consumer Research* 34(4), 441–458 (2007)
36. Zbieg, A., Żak, B., Jankowski, J., Michalski, R., Ciuberek, S.: Studying Diffusion of Viral Content at Dyadic Level. In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1291–1297. IEEE Conference Publications (2012)

Paradox of Proximity – Trust and Provenance within the Context of Social Networks and Policy

Somya Joshi¹, Timo Wandhöfer², Vasilis Koulolias¹,
Catherine Van Eeckhaute¹, Beccy Allen³, and Steve Taylor⁴

¹ Gov2u, Greece

{somya,vasilis,catherine}@gov2u.org

² GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

timo.wandhoefer@gesis.org

³ Hansard Society, London, UK

beccy@hansardsociety.org.uk

⁴ IT Innovation, Southampton, UK

sjt@it-innovation.soton.ac.uk

Abstract. With social networks evolving and integrating within traditional policy domains, the question arises - do we have in our hands a tool for genuine participation, transparency and dialogue, or are the concerns surrounding privacy, trust, provenance and localization still haunting and shaping the arena? In this paper, we discuss this very question via the illustrative lens of the WeGov Project. We start by providing a critical rethinking of e-governance within the context of social media. We then move onto an in depth look at the WeGov project, its toolkit, end-user engagement strategies and methodologies. Finally we draw from our findings some critical insights into the impacts on and implications of such technologies for the policy-making environment. We conclude with a set of recommendations for future work in this area as well as a summary of key lessons learnt within this innovative initiative.

Keywords: Trust, Provenance, Social Networks.

1 Introduction

Recent developments in Information & Communication Technologies (ICTs) have reframed policy and governance contexts to allow for greater, more transparent citizen engagement on issues relating to participative, citizen-led political change & development. There is a need to provide both policy makers and citizens with access to the tools, information, and skills necessary to make informed decisions and to take full advantage of opportunities for community development in the policy-making process.

With regard to technological innovations that are enabling shifts in policy making, two ICTs most dramatically influence the recent explosion of the “social web”:

- Mobile Communications – extending Internet access through a new generation of mobile phones and handheld computers;

- Social Media – enabling individuals to easily upload their own content (text, photos, video) and to find (and discuss) the content generated by others; and closely linked to this is Online Social Networking – enabling people to maintain and to extend their personal and professional networks, as well as to facilitate the flow of information through these networks.

All the above categories are also converging significantly, adding another layer of complexity to research in this field. What is of particular interest here is to determine how increasingly, policy makers are gaining access to platforms enabling them to challenge the status quo of traditional governance, and to imagine what non-hierarchical participative governments might act like.

Interest in social networks has grown exponentially with the development and spread of online social network sites. Social network sites (SNSs) are “web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system” (Boyd & Ellison, 2007). While their key technological features are fairly consistent, the cultures that have emerged around SNSs are varied. Some sites cater to diverse audiences, while others attract people based on common language or shared racial, sexual, religious or nationality-based identities. Sites also vary in the extent to which they incorporate new information and communication tools, such as mobile connectivity, blogging and photo/video-sharing (Boyd & Ellison, 2007).

An impressive body of work already exists around the thematic strains of trust and provenance in social networks (Golbeck, J., 2006 and Zhang, Y., Chen, H., and Wu, Z., 2006), these have mostly taken the route of computational modeling (agent based and statistical analysis). We hope to complement this understanding via a qualitative route, whereby we base our analysis on interview data derived from in-depth interviews with policy makers as stakeholders. Their experiences, opinions and outlook deriving directly from engagement with the latest technical tools has provided us with rich insights into the trends and trials facing the integration of social networks into every day policy making.

In this paper we critically examine the value attributed to data collected from social media, in terms of localization as well as trust and provenance, from the point of view of policy makers. This stakeholder group is of particular importance we believe, as within the political arena of governance, citizen opinions and preferences have traditionally been vulnerable to manipulation to suit the vested interests of those in power. How this dynamic is influenced by current SNS evolution, forms the main thrust of our examination within this paper. To contextualize our inquiry we use the case of the WeGov project, which provides us with a unique lens through which to analyze pressing questions surrounding the meeting of social networking worlds and policy-making.

In particular we look at the WeGov toolbox, which is a web application to support policy makers’ everyday interactions with citizens on SNS. Politicians login with a username and password to see the landing page showing updated search and analysis results in a set of widgets that can be customized around topics and groups of their choice. Some widgets can be geographically restricted to the current location of the

politician and additional locations can be added to the WeGov toolbox. In addition to widgets on the landing page, the toolbox supports a search page with further functionalities and more detailed search and analysis results. Three different analysis components allow users and comments to be identified with respect to a particular topic: (i) the topic opinion analysis creates groups of words that represent the topics within a discussion; (ii) the user activity analysis predicts which posts are going to generate more attention; (iii) and the user behavior analysis classifies users according to their behavior and interactions within the SNS. During the design of the WeGov toolbox policy makers' have been engaged in almost all phases of the development process.

This paper starts with theoretical considerations concerning the rethinking of E-Government (& in particular the integration of social networks within the field of policy making). It presents the landscape of governance with a brief overview of what new forms offer over traditional ones. This is followed by a detailed description of the WeGov case. The successive steps in the development of the WeGov toolkit are presented and illustrated with the most important research results. We examine in light of our stakeholder engagement, the issues of localization as well as trust and provenance that emerge with the adjustments brought on by SNS use in policymaking. Finally, the validation and evaluation of the toolkit is discussed, together with conclusions and recommendations drawn from the research.

2 Rethinking Governance - The Landscape Then and Now

Reform in the field of public governance relates to several 'ideals' upheld within the umbrella term of "good governance". According to DFID, good governance is centered upon three main concepts (DFID, 2006):

- State capability: the ability to get things done, to formulate and implement policies effectively.
- Accountability: a set of institutionalized relationships between different actors that might help bring about responsiveness.
- Responsiveness: when a government or some other public authority act on identified needs and wants of the citizens.

It is these twin concepts of responsiveness and accountability that interest us the most. For we seek to critically examine both the trust and provenance attached by policy makers to dialogues with citizens on social networks, and their need to respond efficiently to their local constituency needs.

It is crucial to define what we mean when we refer to e-governance, given the broad nature of the term and its myriad uses. In this paper we look at e-governance as "the use by government agencies of information technologies (such as Wide Area Networks, the Internet, and mobile computing) that have the ability to transform relations with citizens, businesses, and other arms of government. These technologies can serve a variety of different ends: better delivery of government services to citizens, improved interactions with business and industry, citizen empowerment through access to information, or more efficient government management. The resulting

benefits can be less corruption, increased transparency, greater convenience, revenue growth, and/or cost reductions." (Retrieved at: <http://go.worldbank.org/M1JHE0Z280> on 29th Sept 2012.)

Proponents often promise the outcome of better government including improved quality of services, cost savings, wider political participation, or more effective policies and programs (Garson, 2004; Bourquard, 2003; Gartner, 2000). Others argue the ideal of E-Government has not accomplished the promise of more effective and democratic public administration (Jaeger, 2005; Garson, 2004). (Heeks, 2003) estimates that the failure rate of E-Government projects may be as high as 85%. Therefore, despite large investment, debate continues concerning the vision of E-Government for administrative reform. We take from these studies the idea of e-governance as a facilitator of efficiency and better utilization of resources, and critically examine this rhetoric in light of our experiences within the WeGov initiative.

It is proven that access to information and communication in its own right plays an important role in promoting good governance (Coffey, 2007). In a policy note, DCERN (Development Communications Evidence Research Network) concludes that if "we accept the view that governance requires an inclusive public space based on informed dialogue and debate – an environment in which voice and accountability are central – then it is clear, in theory at least, that communication must have a positive impact on good governance" (DCERN, 2007, p 5). The question that arises then is - Can social networks facilitate in creating this inclusive public policy space, or at least access to it, where dialogue and debate is key?

New possibilities offered by ICT give governments chances to rethink ways of working and providing services for citizens and businesses (Bekkers & Homburg, 2007; Heeks, 2003; Prins, 2001). In this changing world, government authorities simultaneously face two challenges: the importance of fulfilling the new needs and expectations of their citizens and the reality of reduced budgets (Bertot & Jaeger, 2008). The new service delivery must provide greater satisfaction with higher efficiency (West, 2004).

3 Trust and Provenance: How to Navigate Relationships between Policy Makers and Citizens in an Online World?

Any technical system that is brought into the policy-making environment can only work efficiently as part of the larger socio-technical system - i.e. the organization and its human actors (Checkland, 1999). Some authors claim that reported failures of systems to yield the expected productivity gains in organizations (Landauer, 1996) partially stem from a reduction in opportunities to build social capital (Resnick, 2002). Trust can be formed as a by-product of informal exchanges, but if new technologies make many such exchanges obsolete through automation, trust might not be available when it is needed. We find similar considerations in the field of sociology and public policy: the drop in indicators of social capital seen in modern societies in recent years has been attributed—among other factors—to the transformations of social interactions brought about by advances in communication technologies

(Putnam, 2000). Interactions that used to be based on long-established personal relationships and face-to-face interaction are now conducted over distance or with automated systems—a process known as ‘dis-embedding’ (Giddens, 1990). According to this view, by conducting more interactions over distance or with computers rather than with humans, we deprive ourselves of opportunities for trust building. If we are to realize the potential of new technologies for enabling new forms of interactions without these undesirable consequences, trust and the conditions that affect it must become a core concern of systems development.

In a similar vein the provenance of data emerging from online social networks is an issue of considerable concern for policy makers who find themselves balancing finely between the need to reach out and engage their citizens, while at the same time not fall prey to false inputs from either the press or people outside of their local constituency context. There are two main aspects of data provenance: ownership of the data and data usage. Ownership will tell the user who is responsible for the source of the data, ideally including information on the originator of the data. So for example, within the case of SNS, the policy maker will need to have a clear idea of where the feedback to a proposed law is coming from. Data usage on the other hand gives details regarding how the data has been used and modified and often includes information on how to cite the data source or sources. Data provenance is of particular concern with electronic SNS data, as data sets are often modified and copied without proper citation or acknowledgement of the originating data set. Within the remit of this paper we refer to provenance in terms of source traceability of inputs on SNS. In other words we look at who is posting the information and how this impacts trust within the network of relationships online.

3.1 The WeGov Context – Tool and Socio-technical Construct

WeGov - Where eGovernment meets the eSociety is a 33 months research project that has been funded with support from the European Commission under the SEVENTH FRAMEWORK PROGRAMME THEME ICT 2009.7.3 ICT for Governance and Policy Modeling. The WeGov project addresses the networking of citizens about politics, and with policy makers, through social networks like Twitter and Facebook. The approach chosen consists in developing a site, including tools that support the political decision-makers in the analysis of social networks.

The concept of the toolbox is like its name says a box with tools. The tools are accessible throughout the landing page within small and clearly widgets. Figure 1 shows a selection of these widgets. The first widget addresses the main requirement that were mentioned by stakeholders from the initial evaluation phase for monitoring the constituency or local regions. This widget works as storage for locations. End users can create locations and use them in combination of other functionalities. The geographically restriction is currently implemented with Twitter.

The topic analysis functionality (Cp. Figure 1, panel bottom left) reorganizes comments within groups of words that represent a subset of a discussion. (Naveed et al. 2011; Sizov, 2010) The purpose of discussion activity analysis is to predict which posts are going to generate more attention. In the WeGov toolkit the output of

this analysis is translated in top posts to watch (Cp. Figure 1, panel bottom right). The top users to watch are computed by adding the scores of the top posts for each user (Cp. Figure 1, panel bottom center). I.e., the top users are those who generate more top posts (post that are likely to generate higher levels of attention). (Rowe et al., 2011a; Rowe et al., 2011b) The purpose of user behavior analysis is to classify users according to their behavior and interactions within the SNS (Cp. Figure 1, panel top right). (Rowe et al. 2011a)

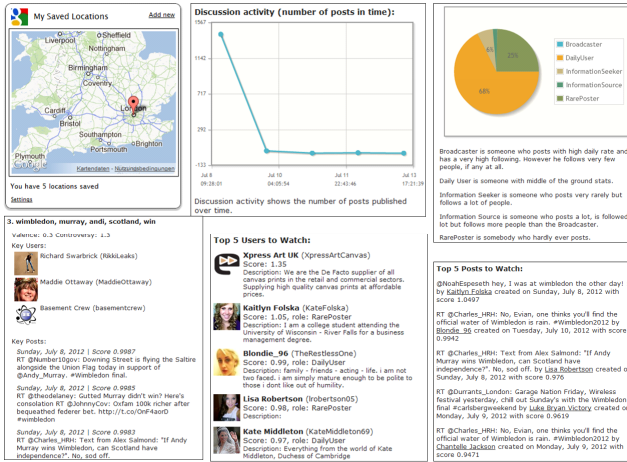


Fig. 1. WeGov toolbox component

3.2 Methodology – How We Set Up Our Experiments and How We Engaged Our Stakeholders?

On the one hand the WeGov toolbox is a research project that develops a website within the field of policy-making to support everyday life functionality for engaging policy makers with citizens on SNS. Hereby the challenge is the overlapping of the politicians’ needs and the technical feasibility of analysis components that are developed in the project. On the other hand the WeGov toolbox is a feasibility study for validating its approach of using automatic analysis components to engage with citizens on SNS. Therefore it is necessary to engage policy makers already in the design process of the analysis tools. The development process needs continuously iterations of combining the policy makers’ requirements with the technical feasibility from the viewpoint of analysis tool development throughout presenting and discussing software prototypes.

3.3 Socio-technical Plan

The aspired outcome of the WeGov project is a toolbox for the e-government to better engage with the e-society and to attenuate the gap between them. Therefore the project has developed a three-step model how the socio-technical process will apply.

The WeGov toolbox mostly considers members of parliament for validating its approach. The MP and the MP’s employees organize these channels as well. That’s why our methodology includes the “state” person (policy maker) and the state PM’s office (PM’s researcher). The state person can have different political layers. For the paper we focus on the EU Parliament, the German Bundestag and the city level. The following table shows an overview of engage stakeholders.

The initial user engagement was based on a conceptual approach, supported by diagrams and workflows, without the opportunity for hands-on testing by the policy maker, making it difficult to create a sustainable commitment to integrate the tool in their daily workflow. For further evaluation a prototype that can be given to the end users is required. Nevertheless, the outcome of this first engagement phase was very helpful to understand the politician’s every day workflow, including profiling the politician’s image, public relations, press work and the dialogue with citizens (Wandhöfer et al., 2011). We were able to communicate the project’s concept and describe the tools. This led to the end users describing situations where they could use the tools. These findings (e.g. which Twitter user to follow on “climate change”) led to the definition of additional use cases and to the enrichment of the initial ones. These use cases have formed the basis of the requirements for the second prototype of the WeGov toolkit. The user interactions also supplied new useful specifications and functionalities (e.g. SNS search with local relevance to engage with the constituency) for the toolbox development.

The image below captures our research design in a clear graphical form:

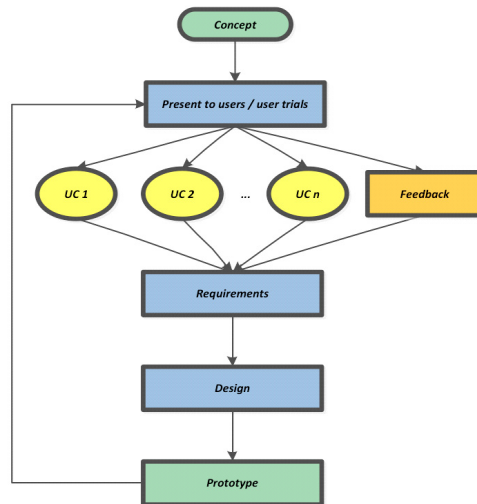


Fig. 2. WeGov research design

This iterative engagement with stakeholders on the projects evolution, progress and outcomes will enable the final results to be more grounded and externally verified by the current concerns of policy makers, their needs and expectations. The WeGov

stakeholder engagement model considers the good stakeholder engagement principles of transparency, meaningful dialogue, expectation-management, feedback and analysis within its practical execution (Good Stakeholder Engagement - Key Components of Stakeholder Engagement. URL: <http://goo.gl/hoq2L> (Retrieved May 2012).

For the purpose of this paper we look at 28 interviews that span the period of June 2011 to June 2012. These were conducted in two phases.

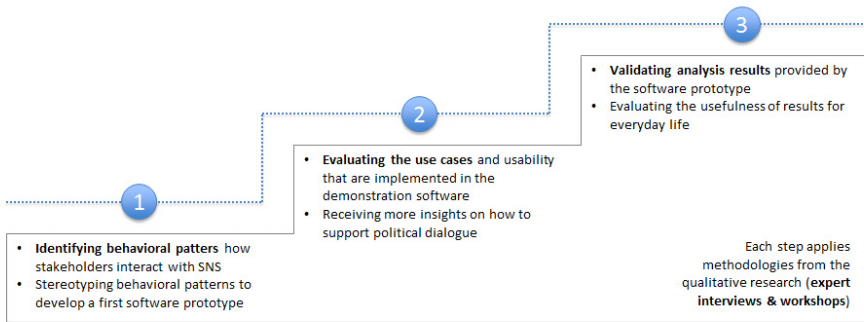


Fig. 3. WeGov methodology steps

Within this process it is necessary to identify and understand the behavioral patterns how the e-Government interact with SNS and what their expectations and problems with this technology are. These patterns will be stereotyped to start the socio-technical process of designing helpful tools to interact with SNS users. Within a further step technical systems are shown to the e-government to evaluate the implemented use cases including the system functionality and its usability. This process is kind of a socio-technical harmonization to better understand behavior and requirements and to improve the everyday life functionality. Within the third step the focus is on the validation of the analysis results that are provided by the software. This step is quite important to receive feedback on the usefulness of the tool results for everyday life and to identify parameter that influences this usefulness for further improvement.

4 Key Research Question

Within the remit of this paper we want to examine two things essentially – the question of 'localization' or 'proximity' of data to make it most relevant and desired from the point of view of the policy makers. And 'the question of 'trust' and provenance – which goes both ways – in terms of the policy maker trusting the data gathered from citizens via SNS and the resulting analysis from our semantic tools, as well as citizen being able to trust in the policy makers. Issues surrounding privacy are particularly relevant in the latter case as there is a fine line of consent between monitoring citizen responses in public debates and discussions and monitoring citizens themselves. Below we examine some key findings along these two thematic axis (localization & trust), emerging from our interview data.

4.1 How Does WeGov Match Evolving Needs?

The process of early stakeholder engagement has identified stereotypes that describe why the engaged stakeholders interact with the social web and how they proceed. These results are mainly from a live demo with stakeholders from the German Bundestag (n=29) that feeds into an open discussion and expert interviews (n=16) that were conducted between June and July 2011. The expert interviews included Members of the European Parliament (n=3), employees that work to a Member of the German Bundestag (n=11), Members from the State Parliament Nordrhein-Westfalen in Germany (n=1) and members with leading positions from parliamentary parties (n=1).

The following table outlines how WeGov matches the policy makers’ evolving needs during their engagement within the WeGov project. Three columns outlines some identified examples (1) why policy makers using SNS, (2.1) what their behavioral patterns are and (2.2) what they expect to solve with Web2.0 technology like WeGov, and (3) how WeGov supports these needs. Main problems with SNS interactions often address issues like, the workflow is applied manually, the tools are black boxes and do not provide transparency, or the tools are changed frequently (e.g. driven by data policy and new features).

Table 1. How WeGov matches the policy makers’ evolving needs

| (1) Identified purposes why policy makers using SNS | (2.1) Identified stereotypes how policy makers using SNS (2.2) Identified stereotypes how policy makers wish using SNS | (3) Supported WeGov functionality that matches policy makers’ needs |
|--|--|---|
| (a) Information fishing – for topics – for persons – within regions – trends of topics | – PM using Twitter to identify information sources (Twitter users), to follow them and being informed on a topic. This requirement is also valid within the constituency. – PMs using tools like Google trends to get “trending” topics. – PMs are responsible for particular topics - so they want to know if topics getting “hotter” or “colder” | – WeGov behavior analysis provides user roles like broadcaster that are interesting to follow. This function is also available local. – WeGov provides for areas that are supported by Twitter its “trending” topics. – WeGov provides long-term analysis on the post frequency to see if a topic gets more impact. |
| (b) Information broadcasting for profiling – Facebook – Twitter profiles | – PMs search for influential Twitter users to disseminate their press releases. | – WeGov behavior analysis provides the user role “influential users”. |

Table 2. (Continued)

| | | |
|--|--|---|
| (c) Gathering citizens' opinions and sentiments <ul style="list-style-type: none"> – on topics – on topics and locations – within locations | – PMs search for citizens' opinions on Twitter and Facebook. It's difficult to identify relevant comments. – PMs monitor Facebook pages (especially within their constituency). But there is much noise on the Facebook feed. | – WeGov provides opinions related to topics. – WeGov provides opinions related to topics and locations. – WeGov will provide sentiments related to topics and locations in the feature. |
| (d) Gathering citizens' feedback <ul style="list-style-type: none"> – on topics – on topics and locations – on the MP's statement | – PMs search Twitter or Facebook pages to gather citizens' feedback on particular topics. – PMs post on Twitter and Facebook and read citizens' comments. | – WeGov provides Twitter queries to identify topics and analyze users' comments – WeGov provides main topics within the discussions. |
| (e) Starting a two-way dialogue with citizens | – PMs often start a dialogue to citizens to address (c) and (d). | – WeGov toolbox will provide a "dialogue" and "reply" button. |

4.2 Our Findings and Evaluation Results

We start our analysis by asking how the WeGov contribution to the policy domain differs from previous non-technological methods of communication between policy makers and citizens. While direct comparison is difficult, it is very common that SNS users react on MP's post with calling the office or writing a letter or email to the MP. Thus we find the more traditional communication approaches are complemented via SNS rather than replaced by them. Furthermore, policy makers use more traditional mediums such as their departmental website or local print media in addition to posts on SNS. This ensures a wider audience base for them. Another key concern reported from our end users refers to the strong lobby of journalists or press who dominate the tracking and response to MP's posts and who dive in the dialogue before citizens have a chance to. Our end users (policy makers) went on to state that the journalist (traditional media) "lobby is strong. They informed us that journalists directly answered on MPs Tweets and blocked citizens interaction out". " This finding highlights an older tension between policy makers and journalists with regard to access to information, which given the more informal setting of SNS, works in the favor of the latter. This was illustrated by our policy makers who said "journalists are fishing for MPs comments" and this greatly impacts on the level of trust then invested in publishing comments and information on SNS.

Our MEPs in Belgium and France also echoed this concern that "the major users of Twitter were the professional journalists and the other policy makers, rather than the citizens".

What are the trends and trials facing the integration of social networks into every day policymaking?

Twitter is very common for publishing information; monitoring information profiles or reacts on posts by replying. In addition it is very helpful to get informed very quickly. Within the German Bundestag it is a common situation that MPs need to speak on side topics that they have not known before. The preparation time for this is only a couple of days. Here Twitter is a useful tool to get informed on the topic very quick and to grab some sentiments what people are thinking on these sub topics. Facebook is more used for public relation. It is hard to find groups or pages (e.g. for topics and constituency) and if they found a group there is less conservation or less 'likes'.

Social Networks – Opening new frontiers or reinforcing old dynamics – What is the inherent value added from data collected off these networks?

MPs believe that web2.0 technology enables open and fast dialogue with citizens. Our end users informed us that "it's difficult to catch citizens opinions", via traditional means, so social networks at once offer a wider reach and immediacy. The proximity is something unparalleled and hence of great attraction to policy makers working at engaging more citizens. Our MEPs found Facebook to be "time consuming with regard to finding influential groups or pages". Twitter was better suited for this kind of work. They also went on to state that "Facebook works as an ice breaker for dialogue" i.e. its inherent value lay in more social connections rather than information load. Our policy makers were interested in "asking citizens questions that typically they would ask other MPs. This would work as an ice breaker for dialogue".

Measuring the impact of their communication online is by far an important challenge for our policy makers. Those interviewed were searching for mechanisms to measure the impact. For instance on Facebook it is possible to gather the counts how often a picture was clicked to view it in the bigger format. But it is not possible to get a count on how many people have recognized one post. And there are viral effects that spread posts to third people. The interviewees expect some tools (in the case of WeGov) to benchmark their activities within the social web. Currently they do not have a solution and need to get direct feedback from their network.

The main rationales behind our policy makers' use of SNS were "press work, public relations, dialogue", " still. They went on to state that the "monitoring of their constituency to measure MP's SNS public relation efforts", was identified as a key target. This links to our previous point of how traditional media relations within policy making are supplemented by SNS rather replaced by it. Our end user experiences show: "citizens want to connect with MPs as friends rather than being fans". The "dialogue on a MP's friend page was felt to have more impact than on a MP's like page".

Local context: the importance of immediacy: What are the implications for trust and what are the demands/ needs for localization? What does this say regarding the 'global' nature of debate on SNS?

The policy makers we spoke to expressed a need for monitoring their local constituency in order to get aware of the problems within it. The French MEP in our user base in particular was very interested in analyses for a geographical area that more or less corresponds with his constituency territory. Meanwhile the city of Ghent asked whether it would be possible to perform analyses according to city quarters that describe the subdivision of the city into 'neighborhoods'. Thus we see an interesting paradox here, where instead of a broadening or 'global' view of policy making, our end users were more interested in applying SNS tools to better understand more fragmented neighborhood level areas within their constituency. We were told the immediacy and proximity this offered allowed them to better address their citizen needs, rather than if they looked thematically at issues affecting wider regions. Most "citizens as friends or fans are from the MP's own constituency" (n=3). Furthermore, "local topics enable more dialogue" (n=4), and "provenance and real user names are important" (n=4). Thus our end users were sending us a clear signal that the local content within their SNS feeds was of more relevance to their everyday work life.

More precisely they were also interested in "finding and observing 'citizen opinion leaders' as this was more effective than finding and reading all citizens' posts" (n=4), where by opinion leaders they referred to those SNS users with the highest impact or influence in public debates. They also stated explicitly "constituency is important towards understanding what makes citizens tick" (n=10).

The question of Trust? Provenance?

Policy makers' still use social media more as an outreach channel (outbound communication) than as an observatory of society. Therefore, the question of relevance and authenticity didn't seem to be a priority by now, within our sample. They were more worried by how the citizens would perceive the authenticity of their communication, how to create impactful posts, and complementary, how to motivate people to like and retweet them. The above was more a priority for our stakeholders that worked in a more institutionalized structure (political party & city administration).

This was illustrated by one of our stakeholders in policy making in Belgium, who stated that the uptake of social media within their local community seemed slower than in a number of other EU countries and hence they felt that "the social media community, mainly the Twitter community was "insufficiently representative for the Walloon public opinion"". This is an interesting observation as the speaker here was directly referring to their local community needs within the context of SNS as a tool of political participation. The same people went on to complain, "A small group of negative people tend to rotten the debate." They were referring to the lack of trust and harm caused by a few social media users who tended to be abusive and not constructive in their comments. This they claimed was "disproportionate to the representativeness of these people and their conversations."

The social media owner of the City of Gent asked what the exact value is of the signals coming from the citizens through the social media? He asked to which extent do these indications arising from SNS have higher emotional load than the more traditional communication media? These are very pertinent communications arising from communities of users who see the value added of greater outreach that SNS promises, but at the same time are wary of expecting too many changes or paradigm shifts in

policy as a result of SNS integration. The presenter of the above doubts also went on to say that the different municipal administrations for which he is coordinating the social media are still analyzing what they exactly want to know from the citizen through the social media. Hence we are still in early days where the needs and expectations of the two groups (policy makers and citizens) are co-evolving.

In conclusion we found from our engagement with stakeholders that many policy makers find social media to be not representative of the total population they seek to address, and that it is still difficult to separate valuable citizen input or feedback from expressions of frustration.

5 Conclusion and Future Directions for Research

In this paper we have offered a close and critical view of the current landscape of social networks within the context of policy making. We have done so through the lens of the WeGov project - whose aim is to engage citizens and policy makers in a meaningful debate. We asked how WeGov matches the evolving, changing landscape of SNS (as well as policy making) and we saw how during the span of the three year project, our users themselves evolved in terms of their needs and sophistication of tools employed. We also saw that in its early stages of evolution there exist several concerns regarding trust, privacy and provenance, which determine update and the impact of such technological interventions within the policy domain. Our policy makers were hesitant in considering the new social media as being truly representative of their constituency and hence were only happy to use it in conjunction with more traditional media at present. With regard to increased accountability, responsiveness, efficient resource mobilization, creation of public policy space, and user (citizen) satisfaction - the sentiment of our policy makers was positive if cautious. It emerged that local needs and priorities trumped all other functionalities and the stakeholders (policy makers) we spoke to unequivocally requested this tool to better address their local constituency needs.

We are aware that the results presented here represent a microcosm of the e-governance world and in particular a very small niche that is fluent with the latest SNS and online tools of engagement. How the insights emerging from this early adaptor point of view, can be generalized to a wider domain, remains an issue of scalability. We believe that with increasing use and integration of SNS within the arena of governance, the lessons we are learning via WeGov will emerge as key best practice guidelines for a broader set of end users.

For future research we propose that the impact of information and data challenges such as inconsistent data structures, semantic issues, abusive, irrelevant posts and incomplete data, on the success of e-government initiatives need to be explored further and presented more thoroughly. In particular we recommend that the need for localization be examined more critically in light of new technologies emerging that enable the detection of provenance of data.

References

1. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13(1) (2007), <http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html>
2. Golbeck, J.: Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. In: Moreau, L., Foster, I. (eds.) *IPAW 2006*. LNCS, vol. 4145, pp. 101–108. Springer, Heidelberg (2006)
3. Zhang, Y., Chen, H., Wu, Z.: A Social Network-Based Trust Model for the Semantic Web. In: Yang, L.T., Jin, H., Ma, J., Ungerer, T. (eds.) *ATC 2006*. LNCS, vol. 4158, pp. 183–192. Springer, Heidelberg (2006)
4. Dawes, Pardo, Connelly, Green, McInerney: Partners in state local information systems: Lessons from the field, from University at Albany /SUNY Center for Technology in Government (1997), http://www.ctg.albany.edu/publications/reports/partners_in_sli/partners_in_sli.pdf (retrieved on October 30, 2002)
5. Hegner: Methoden zur Evaluation von Software, IZ-Arbeitsbericht Nr. 29. Editor: Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V (ASI) (2003), <http://goo.gl/QzxSV> (retrieved June 2012)
6. Kelle, Kluge: Vom Einzelfall zum Typus. Fallvergleich und Fallkontrastierung in der qualitativen Sozialforschung (2010) ISBN: 978-3-531-14704-8
7. Lamnek: Qualitative Sozialforschung. Band 1 Methodologie (1988) ISBN: 3-621-27055-8
8. Mayring: Einführung in die qualitative Sozialforschung (1990) ISBN: 3-621-27095-7
9. Nielsen: Usability-Engineering. Morgan Kaufmann Publishers Inc., San Francisco (1993) ISBN:0125184050
10. Nielsen, Mack: Usability Inspection methods. In: *Proceeding CHI 1994 Conference Companion on Human Factors in Computing Systems*, ACM, New York (1994) ISBN:0-89791-651-4. doi: 10.1145/259963.260531
11. Naveed, Sizov, Staab: ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media. In: *Proceedings Web Science Conference* (2011)
12. Rowe, Angeletou, Alani: Anticipating Discussion Activity on Community Forums. In: *The Third IEEE International Conference on Social Computing*, Boston, USA (2011b)
13. Rowe, M., Angeletou, S., Alani, H.: Predicting Discussions on the Social Semantic Web. In: Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., De Leenheer, P., Pan, J. (eds.) *ESWC 2011, Part II*. LNCS, vol. 6644, pp. 405–420. Springer, Heidelberg (2011)
14. Sizov: GeoFolk: Latent Spatial Semantics in Web2.0 Social Media. In: *Proceedings Web Search and Data Mining* (2010)
15. Wandhöfer, Thamm, Joshi: Politician2.0 on Facebook: Information Behavior and Dissemination on Social Networking Sites – Gaps and Best-Practice. Evaluation Results of a novel eParticipation toolbox to let politicians engage with citizens online. *JeDEM - eJournal of eDemocracy and Open Government* 3(2), S. 207–S. 215 (2011)
16. Ramón Gil-García, J., Pardot, T.: E-government success factors: Mapping practical tools to theoretical foundations. *Government Information Quarterly* 22, 187–216 (2005)

Namelings

Discover Given Name Relatedness Based on Data from the Social Web

Folke Mitzlaff and Gerd Stumme

Knowledge and Data Engineering Group (KDE), University of Kassel
Wilhelmshöher Allee 73, D-34121 Kassel, Germany
{mitzlaff, stumme}@cs.uni-kassel.de

Abstract. During the exhausting search for a given name for the yet unborn baby, the idea of a name recommendation system based on relations mined from the “*social web*” was born. This demonstration paper presents the Nameling¹, a recommendation system, search engine and academic research platform for given names, which attracted more than 30,000 users within four months, underpinning the relevance of the task and associated research questions.

Keywords: Given Names, Network Analysis, Recommendation System.

1 Introduction

Whoever had to chose a given name, knows how challenging it is to find a suitable name which fits to the personal preference as well as the social environment. There is a huge bibliography on books, listing given names in alphabetical order, as well as dozens of respective websites. But finding a suitable name in such a list is an exhausting task. Out of several thousand names, only a small fraction of names is typically relevant to the reader.

In different contexts, recommendation systems are subject to scientific research, as, e. g., finding relevant annotations [5], recommending products [3] or suitable movies [2]. So far, the task of finding relevant given names is not formally investigated, though relevant in practice. Based on data observations from the social web, the Nameling generates such recommendations, enabling users to search for suitable names and browse through a list of more than 35,000 given names.

2 Basic Concepts

The Nameling is designed as a search engine and recommendation system for given names. The basic principle is simple: The user enters a given name and gets a browsable list of “relevant” names, called “*namelings*”. Figure 1a exemplarily shows *namelings* for the classical masculine German given name “Oskar”.

The list of *namelings* in this example (“Rudolf”, “Hermann”, “Egon”, ...) exclusively contains classical German masculine given names as well. Whenever an

¹ <http://nameling.net>

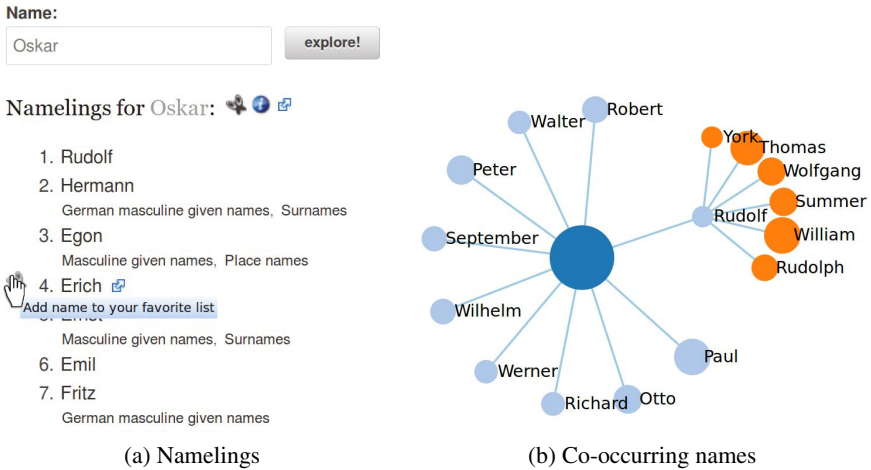


Fig. 1. The user queries for the classical German given name “Oskar”

according article in Wikipedia exists, categories for the respective given name are displayed, as, e. g., “*Masculine given names*” and “*Place names*” for the given name “Egon”. Via hyperlinks, the user can browse for namelings of each listed name or get a list of all names linked to a certain category in Wikipedia. Further background information for the query name is summarized in a corresponding details view, where, among others, popularity of the name in different language editions of Wikipedia as well as in twitter is shown. As depicted in Fig. 1b, the user may also explore the “neighborhood” of a given name, i. e., names which co-occur often with the query name.

From a user’s perspective, the Nameling is a tool for finding a suitable given name. Accordingly, names can easily be added to a personal list of favorite names. The list of favorite names is shown on every page in the Nameling and can be shared with a friend, for collaboratively finding a given name.

3 Background

With the rise of the so called “Web 2.0”, various social applications for different domains emerged, offering a huge source of information and giving insight into social interaction and personal attitudes. The basic idea behind the Nameling was to discover relations among given names, based on such user generated data. In this section, we briefly summarize how data is collected and how relations among given names are established.

The Nameling is based on a comprehensive list of given names, which was initially manually collected, but then populated by user suggestions. It currently covers more than 35,000 names from a broad range of cultural contexts. For different use cases, three different data sources are respectively used, as depicted in Fig. 2:

Wikipedia: As basis for discovering relations among given names, a co-occurrence graph is generated for each language edition of Wikipedia separately. That is, for each

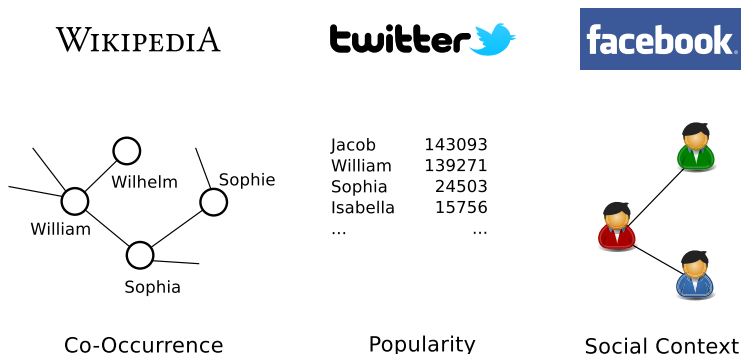


Fig. 2. The Nameling determines similarities among given names based on co-occurrence networks from Wikipedia, popularity of given names via twitter and social context of the querying user via facebook.

language, a corresponding data set is downloaded from the Wikimedia Foundation². Afterwards, for any pair of given names, the number of sentences where they jointly occur is determined. Thus, for every language an undirected graph is obtained, where two names are adjacent, if they occur together at least in one sentence within any of the articles and the edge's weight is given by the number of such sentences.

Relations among given names are established by calculating a vertex similarity score between the corresponding nodes in the co-occurrence graph. Currently, namelings are calculated based on the cosine similarity, which performed best in according experimental evaluations³.

twitter: For assessing up-to-date popularity of given names, a random sample of tweets in twitter³ is constantly processed via the twitter streaming api⁴. For each name, the number of tweets mentioning it is counted.

facebook: Optionally a user may connect the Nameling with facebook⁵. If the user allows the Nameling to access his or her profile information, the given names of all contacts in facebook are collected anonymously. Thus, a “social context” for the user's given name is recorded. Currently, the social context graph too small for implementing features based on it, but it will be a valuable source for discovering and evaluating relations among given names.

4 Emerging Usage Data and Research Questions

Beside being a tool for parents-to-be, the Nameling also serves a research platform. The choice of a given name is influenced by many factors, ranging from cultural background

² <http://dumps.wikimedia.org/backup-index.html>

³ <http://twitter.com>

⁴ <https://dev.twitter.com/docs/api/1/get/statuses/sample>

⁵ <http://www.facebook.com>

over social environment to personal preference. Accordingly, the task of recommending given names is per se subject to interdisciplinary considerations.

Within the Nameling, users are anonymously identified via a cookie, that is, a small identification fragment which uniquely identifies a user's web browser. From now on we talk about users relative to this identification, ignoring the fact that users may use different browsers and/or computers.

Around 37,000 users issued more than 330,000 search queries within the time range of consideration (2012-03-06 until 2012-07-23). For every user, the Nameling tracks the search history, favorite names and geographical location based on the user's ip address and the GeoIP⁶ database. All these footprints together constitute a multi-mode network with multiple edge types. Analyzing this graph (or one of its projections) can reveal communities of users with similar search characteristics or cohesive groups of names, among others.

Most importantly, recommendation systems can be personalized based on the Nameling's usage data via association rule mining [11] or collaborative filtering [6]. But also new approaches can be applied and evaluated, e. g., by considering a users geographical location. Furthermore, the usage data can also be used as a reference for evaluating and improving the process of discovering name relatedness.

5 Conclusion

This demonstration paper introduced the Nameling, a search engine and research platform for given names. From a user's perspective, many more features are desirable - but beforehand, methods for mining relatedness of given names must be evaluated and specialized recommendation systems developed. The analysis of given names and associated social background information is predestined for interdisciplinary considerations, whereby the usage data which accrues at the Nameling may serve as a valuable source of reference data.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB 1994), pp. 478–499. Morgan Kaufmann (September 1994)
2. Golbeck, J., Hendler, J.: Filmtrust: Movie recommendations using trust in web-based social networks. In: Proceedings of the IEEE Consumer Communications and Networking Conference, vol. 96, Citeseer (2006)
3. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7(1), 76–80 (2003)
4. Mitzlaff, F., Stumme, G.: Mining relatedness among given names based on social co-occurrences. Technical report (2012) (submitted for Publication)
5. Pazos-Arias, J.J., Fernández Vilas, A., Díaz Redondo, R.P.: Recommender systems for the social web. Springer, Heidelberg (2012)
6. Terveen, L., Hill, W.: Beyond recommender systems: Helping people help each other (2001)

⁶<http://www.maxmind.com/app/ip-location>

SocialTrends: A Web Application for Monitoring and Visualizing Users in Social Media

Maurizio Tesconi, Davide Gazzé, and Angelica Lo Duca

Institute of Informatics and Telematics
National Research Council (CNR), Pisa, Italy
`name.surname@iit.cnr.it`

Abstract. Nowadays social media trends are becoming very important to describe the variation of popularity, activity and influence of an entity. In this paper we define an abstract model which can be used on different social media to compare metrics with the same meaning. In particular we describe three classes of metrics: *popularity*, *activity* and *influence*. We also present SocialTrends, a web application (<http://www.social-trends.it/>) which collects, elaborates and visualizes social media data. Finally, we describe one experiment we have done to test SocialTrends.

Keywords: Social Media Visualization, Monitoring, Trends, Data Gathering.

1 Introduction

Nowadays the popularity, the activity and the influence of an entity (e.g. politician, singer, journal...) on the web can be measured through Social Media (e.g. Facebook, Twitter, YouTube...) [1], in which each person represented by a profile can express his/her interests. The *popularity*, the *activity* and the *influence* of an entity have been investigated according to two approaches. The first, which is more theoretic, is based on scientific research and describes an abstract model. The second is more practical and is based on web applications, which collect, elaborate and visualize Social Media data. In this paper we merge both the approaches, through the definition of an abstract model and the implementation of a web application.

The purpose of this work is to give a formal model for entity representation in Social Media and implement a web application, which collects, elaborates and visualize real data. Furthermore, we compare the entities by category, in order to understand what is the most popular, active and influential entity for a given category. We describe metrics which can be used to establish the impact of the entity on a channel in their temporal evolution. We propose a web application, called SocialTrends (<http://www.social-trends.it/>), which sorts entities belonging to the same category in order to obtain a ranking and for each entity it shows the distance to the other entities of the same category. SocialTrends can be used by all the categories of end-users, from curious people to expert ones.

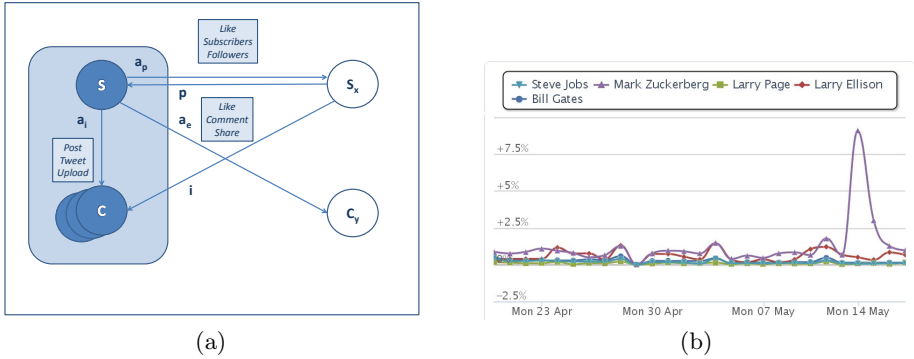


Fig. 1. a) Relationship among two entities. b) Percentage Increase of American CEOs' popularity from Facebook

2 Social Media Model

2.1 Model

We define an abstract model which can be used on different social media. It is composed of four elements: entity, source, channel and metrics. An *entity* is an abstract concept representing something in the real world (e.g. person, journal, political party...). A *channel* is defined as web site where each entity can perform the following operations: a) build its profile, b) share its profile with other individuals, c) communicate with other entities [2]. In practice, once its own profile is built, an entity can establish its relationships with the other entities and interact with them. On a channel an entity may exploit many sources to spread opinions, personal news and thoughts. A *source* is a web place providing a mechanism to express interests, publish news about themselves and establish a relationship with other entities. For example, a Facebook source type can be a *page*, a *group* or an *account*, for Twitter and YouTube it can be an *account*. A source can generate *content*, which includes all the activity the source performs within the channel. A source provides some useful metrics which can be used to establish the impact the entity has on the channel. By *metrics* we denote a set of mechanisms giving statistic information about the entity associated to that source.

2.2 Metrics

We model a channel as directed graph $G(N, A)$, where sources S and contents C represent the nodes N and the actions between a source and a content or between two sources are the arcs A . An arc from a source to a content exists when the source generates that content, while an arc from a source to another one exists when the first expresses an interest in the second. We define the *indegree*

$deg^-(N)$ and the *outdegree* $deg^+(N)$ of a node N as the number of arcs pointing to N and outgoing from N , respectively.

Figure 1a) focuses on a source S and how a generic source S_x can be connected to it. In order to simplify the diagram, only the relationships to S are shown. By C and C_x we denote a content produced by S and S_x . The arc pointing from S_x to S , namely p , represents the interest that S_x has in S . The arc from S to C , namely a_i , represents the action performed by S when it produces its own content, while a_e represents an action performed by S on the content produced by S_x . Finally, the arc from S_x to C , namely i represents an action performed by S_x on the content produced by S .

The *popularity* of a source on a channel is defined as the level of attention it receives from the other sources [3]. More formally, the popularity \mathcal{P}_i of the source S_i is defined as the indegree of S_i :

$$\mathcal{P}_i = deg^-(S_i) = |p| \quad (1)$$

where $|*|$ represents the cardinality.

The *activity* [4] of a source on a channel is defined as the frequency of content publication. More formally, we define the *activity* \mathcal{A}_i of a source S_i as the outdegree of S_i :

$$\mathcal{A}_i = deg^+(S_i) = |a_i| + |a_e| + |a_p| \quad (2)$$

The *influence* [5] of a source on a channel is defined as feedback that it receives on its generated content. More formally, we define the *influence* \mathcal{I}_i of a source S_i as the sum of the indegrees of all the contents C produced by S_i

$$\mathcal{I}_i = \sum_{C \in \{C_i\}} deg^-(C) = \sum_{C \in \{C_i\}} |i_i| \quad (3)$$

where C_i is the set containing all the contents produced by S_i .

Several different models have been defined to calculate the popularity, the influence and the activity of an entity on a Social Media. The drawback of such models is that they do not provide the visualization of raw data, since they show only elaborated data. The purpose of SocialTrends consists in visualizing data as collected from Social Media.

3 SocialTrends

SocialTrends is a web application which collects, elaborates and visualizes data collected from social media. The strength of SocialTrends is that it focuses on matching entities characterized by the same tags (i.e. belonging to the same category). In particular, the match of entities is achieved through visual rankings, grouped according to the metrics described in section 2.2 and to the channel. In practice, SocialTrends sorts entities belonging to the same category in order to obtain a ranking per category and per channel. For each entity it shows the *distance* to the other entities of the same category. By distance between entity i and entity j we denote the difference between the value of the metrics of i and

that of j . To the best of our knowledge, SocialTrends is the first tool providing this feature, since the other web applications emphasize only single entities and not a comparison of them. In fact, for each entity only the position in the ranking of the category it belongs and many statistic details about it are given.

4 Experiment

In this section a description of of one experiment done to test SocialTrends is presented. For other experiments we refer to the poster. In particular, a focus on the popularity of Facebook pages is illustrated.

Figures [11](#) b) shows the percentage increase in popularity of the most popular CEOs on Facebook. The percentage increase of a metrics \mathcal{M} indicates how much the value of the metrics grows in percentage. It is interesting to note that although Mark Zuckerberg is not the most popular CEO on Facebook, his percentage increase had a peak of 9.10% in May 14th, which corresponds to a relatively small absolute increase of 1171 (compared to his number of likes, which is 22373). This peak can be associated to Mark Zuckerberg's birthday.

Acknowledgments. This work has been partially supported by EU FP7 Project CAPER (Grant Agreement no. FP7-261712).

References

1. Cvijikj, I.P., Michahelles, F.: Monitoring trends on facebook. In: Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, DASC 2011, pp. 895–902. IEEE Computer Society, Washington, DC, USA (2011)
2. Boyd, D.M., Ellison, N.B.: Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* (2008)
3. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and Passivity in Social Media. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 18–33. Springer, Heidelberg (2011)
4. Burke, M., Kraut, R., Marlow, C.: Social capital on facebook: differentiating uses and users. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI 2011, pp. 571–580. ACM, New York (2011)
5. Varlamis, I., Eirinaki, M., Louta, M.: A study on social network metrics and their application in trust networks. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2010, pp. 168–175. IEEE Computer Society, Washington, DC, USA (2010)

Demonstration of Dynamic Targeting in an Online Social Medium

Peter Laffin¹, Fiona Ainley¹, Amanda Otley¹,
Alexander V. Mantzaris², and Desmond J. Higham²

¹ Bloom Agency: Green Sand Foundry,
99 Water Lane, Leeds, LS11 5QN, United Kingdom
`plafflin@bloomagency.co.uk`

² Department of Mathematics and Statistics:
University of Strathclyde, 26 Richmond Street,
Glasgow, G1 1XH, United Kingdom
`alexander.mantzaris@strath.ac.uk`

Abstract. A novel way of calculating online influence has been proposed in [2]. Bloom Agency have created new online software capable of collecting social data and calculating these new influence metrics in real time. A demonstration of this software will be given at the conference. Delegates will be encouraged to Tweet using the #socinfo2012 hashtag and the influence of the top ten Tweeters will be shown, along with a visualisation of the evolving conversation.

Keywords: Twitter, Social Media, Online Influence, Dynamic Networks, Katz Centrality

1 Summary

We will demonstrate new software for analysing social media networks, such as Twitter, in real time to calculate influence ratings that identify influential individuals and assess sentiment within a conversation. The software utilises the new dynamic centrality measures [2] developed at the Universities of Reading and Strathclyde and allows a changing social network to be investigated quickly and accurately. The influence score is different to those produced by Klout and Peer Index. The score calculated by the software focuses on a specific conversation rather than a specific individual. When assessed by social media experts, the scores produced by the software have been verified to be more representative and realistic than Klout and Peer Index in relation to specific conversations. The details of the assessment by social media experts, along with the technical detail of the new method of influence scoring is described in a paper to be presented to the conference, entitled Dynamic Targeting in an Online Social Medium [4]. Conference delegates are encouraged to Tweet about the conference, utilising the #socinfo2012 hash tag. In real time, as Tweets are published by delegates, the conversation about the conference can be visualised through a web browser, for delegates to observe. Using dynamic centrality calculations, performed on

the raw Tweet data, a real time list of influential Tweeters will be produced and updated once every 10 minutes for all to see. The conference demonstration will show delegates this list on a screen which we anticipate will be set up in a prominent place for the duration of the conference. We expect this experiment to raise many questions around why am I more influential than x? We can discuss the method we use and justify why we believe, and the social media experts believe, that it is a valid method of measurement which should stand the test of time. Literature written by Bloom, which describes how the software functions, will be available for delegates to review. This demonstration will encourage delegates to Tweet about the conference and thereby increase the volume of Tweets, and the level of exposure, of the conference. Overall, we believe the demonstration will not only create a real talking point at the conference, as delegates discuss how their own influence scores differ, but also offer the opportunity for the conference to get wider attention through Twitter.

2 Background

Bloom Agency are an integrated marketing agency, with 13 years experience of solving marketing related problems for clients. Bloom began investigating how to analyse and visualise what is happening in social networks in early 2011, in response to requests from clients asking for better methods to measure social media marketing campaigns. Contact was made with the Universities of Reading and Strathclyde to ensure the academic rigour and timeliness of the research work undertaken at Bloom. The collaboration is now funded through a UK Technology Strategy Board (TSB) SMART grant project, to investigate how return on investment from social media marketing campaigns can be quantified. The software to be demonstrated is the culmination of one year's research work by Bloom, supported by the TSB and the two universities. The software will be launched commercially in 2013 and this demonstration will be one of the first opportunities for interested parties to view it and discuss possible implications. The software has already been used to successfully target influential individuals in social networks for a range of UK household brands, in order to deliver return on investment for social media and SEO campaigns. The combination of Bloom's big data skills/social media expertise and the new mathematical frameworks developed at Reading and Strathclyde have brought a new dimension to real time social network analysis, with a real focus on delivering the answers to key questions for users of online human interaction data.

3 Methodology

The software that we will demonstrate uses time-stamped Tweets and the underlying follower network. Our approach is to build an appropriate evolving network; that is, an ordered sequence of adjacency matrices that summarize relevant interactions. The details of our methodology are included in more detail in the full paper to be presented to the conference [\[4\]](#). The software provides a

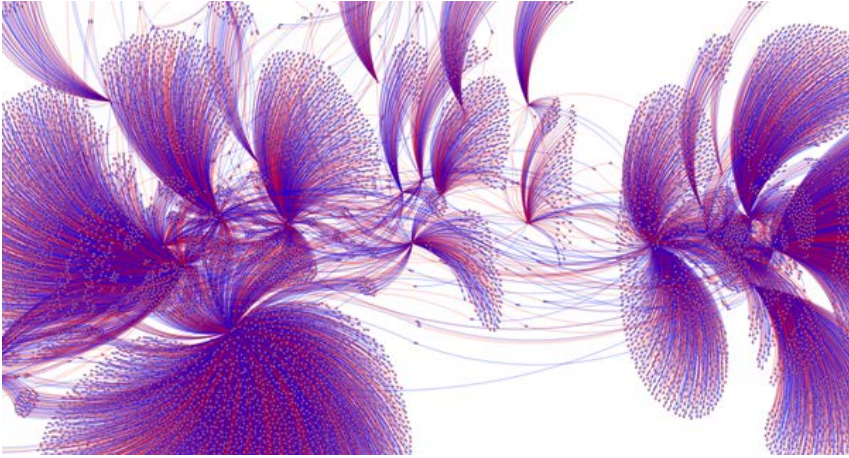


Fig. 1. Simple Network Visualisation

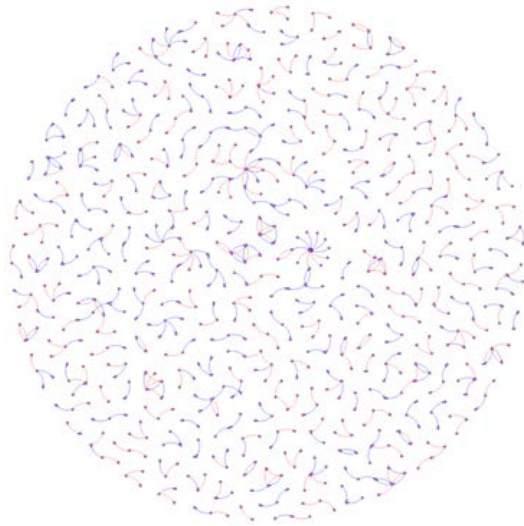


Fig. 2. A network visualisation in greater resolution

convenient, computationally efficient layer to implement the ideas presented in the paper. This shows how an academic project can be developed into a commercial demonstrator and highlights the importance of University / Industry partnerships. Simple visualizations, such as figure [1](#), can give a feel for this type of data.

A graphic similar to this will be produced, in real time, based on the Tweets made about the conference. Delegates will be able to see the conversation evolving in real time and identify those influential individuals who are talking about the conference. Delegates who are interested will be able to speak directly to

Bloom about the nature of this data visualisation. A different form of visualisation, shown at a greater resolution than figure 1, is shown in figure 2.

Our computation is focused on calculating the proposed dynamic measures from [21], which are based on combinatorically counting routes between nodes that respect the time ordering of the interactions. These are generalizations of the classic Katz centrality measures [3] that have proved to be extremely influential in the case of static networks. The dynamic versions maintain the same flavour as the Katz measures in the sense that they involve only basic linear algebra operations; notably the solution of linear matrix-vector systems whose sparsity matches that of the underlying adjacency matrices. This allows us to tackle large networks and opens up the possibility of real-time updates.

References

1. Grindrod, P., Higham, D.J.: A matrix iteration for dynamic network summaries. Tech. Report University of Strathclyde Mathematics Research Report 22 (2011) (to appear in SIAM Review)
2. Grindrod, P., Higham, D.J., Parsons, M.C., Estrada, E.: Communicability across evolving networks. *Physical Review E* 83, 046120 (2011)
3. Katz, L.: A new index derived from sociometric data analysis. *Psychometrika* 18, 39–43 (1953)
4. Laffin, P., Mantzaris, A.V., Ainley, F., Otley, A., Grindrod, P., Higham, D.J.: Dynamic Targeting in an Online Social Medium. In: Aberer, K., Jager, W., Liu, L., Tang, J., Flache, A., Guéret, C. (eds.) *SocInfo 2012*. LNCS, vol. 7710, pp. 82–95. Springer, Heidelberg (2012)

Navigating between Chaos and Bureaucracy: Backgrounding Trust in Open-Content Communities

Paul B. de Laat

Faculty of Philosophy, University of Groningen, Groningen, The Netherlands
p.b.de.laat@rug.nl

Abstract. Many virtual communities that rely on user-generated content (such as social news sites, citizen journals, and encyclopedias in particular) offer unrestricted and immediate ‘write access’ to every contributor. It is argued that these communities do not just assume that the trust granted by that policy is well-placed; they have developed extensive mechanisms that underpin the trust involved (‘backgrounding’). These target contributors (stipulating legal terms of use and developing etiquette, both underscored by sanctions) as well as the contents contributed by them (patrolling for illegal and/or vandalist content, variously performed by humans and bots; voting schemes). Backgrounding trust is argued to be important since it facilitates the avoidance of bureaucratic measures that may easily cause unrest among community members and chase them away.

Keywords: access, bots, bureaucracy, etiquette, open content, trust, vandalism, Wikipedia.

1 Introduction

Online communities that thrive on user-generated content come in various formats. Contents may vary considerably—from text, photographs, videos, designs and logos to source code. Furthermore, cooperation may range from ‘loose’ interaction: uploaded contents are presented as-is—to ‘tight’ interaction: an evolving product is being worked on collectively. This distinction in patterns of cooperation is referred to by Dutton [3] as ‘contributing 2.0’ vs. ‘co-creation 3.0’. Typical examples of the former are Flickr and YouTube, of the latter Wikipedia and open source software.

These communities face the dilemma of determining which contributors are to be accepted as members and how contributions are to be processed and published. Some communities take a cautious approach: only some categories of people are allowed to contribute, and their contributions are critically examined, by filtering before reception or moderating afterwards. A typical example is the Encyclopedia of Earth which only accepts input from acknowledged experts. Moreover, their appointed ‘topic editors’ decide who is to write the entries and who is to participate in reviewing them. In the end they have to approve of entries appearing in a public version. Other communities, though, prefer to hand out a generous invitation to their ‘crowds’ in order to

maximize possible returns. This consists of two parts: (1) Anyone is invited to contribute content without any restrictions on entry; accordingly, access is fully open to anyone who cares to contribute; (2) Contents contributed are subsequently accepted with no questions asked and appear right away in the appropriate place. Publication proceeds without review and without delay. In terms of Goldman [6]: no filtering is applied at the reception stage.

Which communities typically practice this two-fold institutional gesture? Let me mention some of them in so far as they predominantly revolve around soliciting and reworking of *text*. I select these since it seems especially with text that the whole spectrum from contribution (2.0) to co-creation (3.0) unfolds; activities in communities which focus on other kinds of content most often remain at the level of contributing.¹ The first category is ‘*social news*’ sites that focus on creating a collective discussion about topics in the news that are deemed to be relevant. The formula is basically the same for all: users are invited to submit news stories and/or news links that will be put up for public discussion (comments). In this category we find Digg (<http://digg.com>; started in 2004) and Reddit (<http://www.reddit.com>; 2005) which focus on news of all kinds, and Slashdot (<http://slashdot.org>; 1997) and Hacker News (<http://news.ycombinator.com>; 2007) which focus on technology-related issues.²

The second category is user-generated *newspapers* that have been around since 2004. NowPublic (<http://www.nowpublic.com>; 2005), Digital Journal (<http://www.digitaljournal.com>; 2006) and GroundReport (<http://www.groundreport.com>; 2006) invite everybody to become a citizen journalist and contribute their own articles, blog entries and/or images to the site, as well as leave comments on those of others. These contributions essentially remain unaltered. Wikinews (<http://en.wikinews.org>; also in language versions other than English; 2004) goes one step further: in the so-called ‘news room’ articles which have been submitted are polished further by fellow contributors (by means of a wiki). As soon as criticisms have been met, the article can officially appear on the ‘front page’.

The third and final category consists of user-generated *encyclopedias*. Many such communities exist (cf. [8]), but only a few have adopted policies of open access and immediate publication. British h2g2 (<http://h2g2.com>; 2001) invites everybody to compose entries; these are put up on the site for public commentary. Wikipedia (<http://en.wikipedia.org>; also in many languages other than English; 2001) and Citizendium (<http://en.citizendium.org>; 2007) lean more towards co-creation by publishing new entries in an open-access wiki that allows other participants to instantaneously insert their own textual changes.³

¹ These communities will be enumerated with their URL and year of foundation in brackets.

² All websites referred to in text, footnotes, and references were last accessed on 21 September 2012.

³ These communities will serve as cases to be analysed further on in this article. Note that while they practiced unrestricted and immediate access from the outset, some of them recently have been pondering—or actually resorted to—more restrictive editorial policies: filtering before reception (to be commented on below).

2 Trust

This gesture of unrestricted and immediate access to the community platform (to be denoted ‘write access’)⁴ can be interpreted as a form of ‘*institutionalized*’ trust towards prospective participants.⁵ The italics are employed in order to stress two particular points. On the one hand, the gesture is an institutional one: we are dealing here with the ways in which an institution approaches the members it depends on, not with interpersonal trust. On the other hand, the gesture embodies the presumption that prospective participants are willing to contribute content with good intentions and to the best of their capabilities. Their trustworthiness in terms of moral intentions and capabilities is taken for granted. Notice that different capabilities are involved across the various communities. Social news sites aim to stimulate lively debates about current issues. Therefore, the required capabilities concern being able to take a normative stand and provide supporting arguments. Mastering the rules for discussion (rhetoric) is vital. Encyclopedic projects, on the other hand, aim to produce state-of-the-art knowledge about issues that are deemed relevant. Hence, participants should have adequate capabilities to contribute knowledge; their epistemic qualities are sought after. Citizen journals occupy a position in between. They are looking for both kinds of capabilities, since journal articles usually require both reporting of the facts and commenting on them. Good news is a judicious blend of fact and opinion.

That trust is at issue here can easily be seen from the fact that all the communities concerned are exposing their respective repositories of content and *entrusting* them as it were to the whims of the masses. They have decided to rely fully on their volunteers, thereby making themselves vulnerable and taking risks. Discussion sites, published news reports and encyclopedic entries can easily be polluted and spoiled by all kinds of disruptive actions. As Wikipedia defines the matter, ‘cranks’ may insert nonsense, ‘flamers’ and ‘trolls’ may enjoy fomenting trouble, ‘amateurs’ may ruin factual reporting, ‘partisans’ may smuggle in their personal opinion where this is inappropriate, and ‘advertisers’ may just try to promote their products anywhere (<http://en.wikipedia.org/wiki/Wikipedia:RCO>). Repositories polluted in this way undermine the viability of any community, and necessitate laborious cleanups.

Given this gesture of fully trusting potential participants and giving them write access accordingly, what mechanisms of trusting others may be relied on in the process? What processes possibly lie behind it? In the sequel I discuss three

⁴ This term is in use among developers working together on open source software. As a rule, anyone may access the site and inspect the contents (‘read access’). When participants have proven their skills, they may acquire the additional right to directly contribute code to a project’s source code tree: they have obtained ‘write access’.

⁵ Note that such usage of the term ‘trust’ as exercised by institutional actors harks back to Alan Fox, who proposed to interpret organizational regimes (‘work role patterns’) in terms of the amount of trust granted by organizations to their members [4]. In this instance the institution is neither conceived of as a *producer* of trust (‘institutional-based trust’; term coined by Zucker [16]), nor as an *object* of trust (‘system trust’; term coined by Luhmann [10]). Instead, the analysis casts the institution in the role of a trusting party, a ‘trustor’.

well-known mechanisms to handle the trust problem: the assumption, inference, and substitution of trust. Subsequently, I argue for a fourth mechanism that seems to have been neglected in the literature thus far: *backgrounding* trust. In this approach the gesture of full trust is underpinned by developing support mechanisms in the background that render the trust-as-default rule rational in a reductionist way.

First and foremost, the trust involved may be the simple *assumption* that the crowds are trustworthy. Trustworthiness is assumed without any particular evidence to support that assumption. This assumption is not made without reason; its rationale, as observed by Gambetta decades ago, is that precisely by acting *as if* trust is present, one may actually produce it in the process [5]. In Luhmannian terms: the gesture of trust creates a normative pressure to respond likewise. The act of trust can thus be seen as an investment that it is hoped will pay off [10]. Can any good reasons be advanced for the assumption? What mechanism may be argued to underlie the said normative pressure?

In line with Luhmann, Pettit argued that esteem is the driving force [12]. Since people are sensitive to the esteem of others, they will answer an act of trust with trust as it enables them to reap the esteem that is being offered to them. As elaborated before [7: 332], this interpretation of the normative force of trust does not seem wholly convincing in the case of open-content communities. While esteem surely is a driving force, it would seem to be an underlying one, not a paramount one. A more forceful interpretation obtains if we move away from this calculating conception of as-if trust to another conception that is based on a vision of and hope in the capabilities of others. As argued by McGeer [11], showing trust may be rooted in the hope of challenging others to apply their capabilities in return. These others are not manipulated but empowered to show their capacities and further develop them. The trusting party puts his/her bets on a utopian future.⁶ Such reasoning can in a straightforward fashion be applied to our open-content communities since the capabilities that are the cornerstone of this McGeerian vision have quite specific connotations here. By granting unrestricted and immediate access, crowd members are challenged to show their capacities of commenting, reporting news, or contributing reliable knowledge. They are invited to fulfill the promise of a community of exciting, newsworthy, or encyclopedic content.

A second way to handle the tensions that a trusting gesture generates is to *infer* trustworthiness. One looks for indicators that inspire confidence in the other(s) as a trusted partner: perceived individual characteristics like family background, sex, or ethnicity, belonging to a shared culture, connection(s) to respected institutions, or reputation based on performance in the past (this argument can be traced back to Zucker [16]). Moreover, the calculative balance of costs and benefits may seem to preclude a non-cooperative outcome. As argued before [7: 330-31], I do not believe that an open-content community operating in cyberspace has many reliable indicators

⁶ McGeer uses the term ‘substantial’ trust, as opposed to the shallow trust Pettit is supposed to refer to. I prefer to avoid the former term since, in my view, not another type of trust is being defined, but just a different mechanism for generating trust *ex post* that actors may supposedly rely on *ex ante*.

to cling to. Virtual identities are always precarious; the anonymity of contributors only aggravates this problem. Even the common requirement to register and choose a user name (or even disclose one's real name) hardly alleviates the problem (cf. [8]). Moreover, contributors often just enter and leave, precluding any stable identity let alone the formation of a reputation. To sum up: signalling trustworthiness cannot be implemented in a reliable way. While the inference of trust has rightly been regarded a central component of processes of trust formation in *real life*, I do not think it has much value in the virtual surroundings of open-content communities.⁷

A third way to handle the problem of trust may be referred to as the *substitution* of trust. Wherever people interact continuously and some kind of community emerges, rules, regulations, and procedures tend to be introduced. Often these enact restrictions on behavioural possibilities. As a result, reliance on participants' wisdom and judgment in contributing is reduced; their actions become less discretionary. As a corollary, the need to grant them trust is lessened; the problem of trust is partly eliminated. The introduction of a bureaucratic structure of the kind effectively substitutes for the need to estimate—or assume—participants as being trustworthy. Below, evidence is presented on some of our open-content communities recently instituting restrictive rules and regulations: filtering incoming content prior to publication. Write access thus becomes circumscribed and regulated.

However, a fourth mechanism to deal with the tensions of an all-out policy of trust is to be distinguished. It embodies efforts, in the absence of reliable inference, to create a middle road between relying on the normative power of trust on the one hand, and (partly) eliminating the problem by substitution on the other hand. In this approach the default rule of all-out trust is kept intact by underpinning it in the background with corrective mechanisms that contain the possible damage inflicted by malevolent and/or incapable contributors. To my knowledge, this approach, to be referred to as *backgrounding* trust, has been neglected in the literature up to the present. As we will see, the supportive mechanisms themselves are not unknown, but their corrective function for keeping the default rule of trust intact has largely gone unnoticed.

3 Backgrounding Trust

I propose that several types of backgrounding can be distinguished (to be elaborated below in further detail). First, a cultural offensive can be launched to curb potential vandals: legal terms of use and an etiquette of sorts that defines proper behaviour are developed and propagated. Secondly, these standards of behaviour can be underscored by defining sanctions and disciplinary measures. Participants who deviate too much from the ground rules for constructive cooperation may be punished and ultimately expelled from the community. Thirdly, structural schemes can be introduced that aim to guarantee the quality of the community's contents. These range from relatively simple vandalism patrol schemes up to voting and quality enhancement

⁷ To be fair, though, it should be remarked that in many virtual *trading* communities reputation systems have been built that do provide more solid grounds for inferring trust.

programs. The bottom line for all three activities is that they may—at least partly—contribute to sustaining the rationality of the decision to maintain an editorial policy of all-out trust. They serve to keep the default rule of full trust in place.

3.1 Legal Terms and Etiquette

As a consequence of their full-trust write access policy, our open-content communities are quite vulnerable to disruptive behaviour, from posting illegal content to vandalist actions. As a way of defence they are first of all trying to lay down legal guidelines. Plagiarism, libel, defamation, illegal content and the like are strictly forbidden. This is considered the baseline for proper behaviour since deviations from them would land the site in legal trouble.

Interestingly, though, our communities under study also promote ‘good manners’ *beyond* these legal terms of use. An etiquette is formulated for regulating mutual interactions on their sites. Leaving Wikinews and Wikipedia aside for the moment (see below), all of them stress the same kind of exhortations in their ‘community guidelines’, ‘house rules’, ‘netiquette’, or ‘reddiquette’—albeit to varying degrees.⁸ On the positive side, members are urged to always remain respectful, polite, and civil; to stay calm; to be patient, tolerant, and forgiving; to behave responsibly; and/or to stay on topic at all times. On the negative side, the list of interdictions is much longer. One is urged to refrain from calling names, offensive language, harassment, and hate speech. Flaming and trolling are sharply condemned. Commercial spam and advertisements are declared out of bounds. Flooding a site with materials that are offensive, objectionable, misleading, or simply false only amount to an objectionable waste of the site’s resources (nicknamed ‘crapflooding’).

Finally, let us consider Wikinews and Wikipedia. Both under the umbrella of the Wikimedia Foundation, they have adopted virtually the same etiquette (called: Wikiquote). It is in fact the most extended set of rules for polite behaviour in open-content communities to be found anywhere on the Net. Assuming good faith on the part of others—and showing it yourself—is the starting point. Help others in correcting their mistakes and always work towards agreement. Remain civil and polite at all times: discuss and argue, instead of insulting, harassing or personally attacking people. Be open and warm. Give praise, and forgive and forget where necessary. Overall, several pages are devoted to the

⁸ The following observations are based on a range of sources:
digg.com/tos, <http://www.reddit.com/help/reddiquette>,
<http://slashdot.org/faq>,
<http://ycombinator.com/newsguidelines.html>,
http://www.nowpublic.com/newsroom/tips/fine_print/flaming_policy,
http://www.nowpublic.com/newsroom/community/code_of_conduct,
<http://digitaljournal.com/article/179808>,
<http://www.groundreport.com/content.php?section=editorial>,
<http://en.citizendium.org/wiki/CZ:About>,
<http://h2g2.com/dna/h2g2/A901838>, and
<http://h2g2.com/dna/h2g2/A87523211>.

subject (see <http://en.wikinews.org/wiki/Wikinews:Etiquette>; and <http://en.wikipedia.org/wiki/Wikipedia:Wikiquote>).

3.2 Enforcement

Both legal rules and etiquette cannot operate without some mechanism of enforcement. With all the communities above, without exception, sanctioning of deviant users has become the normal state of affairs. Users who (repeatedly) flout the rules of etiquette—let alone the legal rules—can be banned from the community for some period of time, or even forever. As a rule the professional editors employed by the site (‘editorial team’) simply assume these judicial powers themselves. With others, site volunteers are entrusted with the task. At h2g2, these are appointed for the job (as ‘moderators’) by the staff of the company which owns the site (formerly the BBC). The pair of Wikipedia and Wikinews appoints candidates with a procedure that relies on public consultation of the community (‘administrators’). Citizendium does likewise (‘constables’).

The mechanisms of rules and sanctions taken together send the message: respect legal terms of use and be civil and polite—otherwise risk expulsion. Notice how these may impact on the employed policy of unrestricted and immediate access. That policy assumes the trustworthiness of the participants from the outset. Inculcating respect for legal issues and rules of etiquette then may serve to *create* trustworthiness where it is found to be lacking—*afterwards*. Whenever the assumption of trustworthiness appears unwarranted, that defect can (at least partly) be repaired afterwards. As a result, the full write access policy is underpinned and can possibly remain in force after all. ‘Backgrounding’, as I shall call this phenomenon, keeps confidence in full-trust as the default intact.

I would argue, however, that these mechanisms can do just so much. They can only possibly ‘educate’ participants who are staying longer. Newcomers, who are the most likely source of mischief, can hardly be supposed to have read let alone internalized the rules involved upon entry. As a result, the campaign for legal and civil conscience has no effect on them, and the full-trust policy remains vulnerable to their abuse. Therefore we now turn to structural means that may support the full-trust policy. No longer the dispositions of people but the contents they actually contribute come into focus. I shall argue that these tools are ultimately able to do a more powerful job of sustaining that policy.

3.3 Quality Management

The term ‘quality management’ is used with quite a broad meaning: it refers to both *rating* and (for dynamic entries) *raising* the quality of contributed content, throughout the whole quality range, from low to high. At the lower end, the mess of clearly inappropriate content that flouts basic legal terms of use or etiquette has to be cleaned up.

Beyond these tasks of ‘basic cleaning’ (as I shall label them) the quality of the content—as far as it has passed the former test of scrutiny—can be monitored continuously and (in case of dynamic content) raised ever further. Such quality schemes may already be the normal *modus operandi* (cf. the wiki format); they may also be developed as additional mechanisms by the communities involved since they consider their basic mode to be an insufficient guarantee of quality.

Social News Sites and Citizen Journals. Social news sites and citizen journals (apart from Wikinews) are usefully treated together since all operate in the ‘contributing 2.0’ mode. These solicit stories (whether existing—for social news sites, or newly composed—for citizen journals) and comments on them. The tasks of basic cleaning are performed (afterwards) by the editorial teams involved: they scout their sites continuously for illegal and inappropriate content. Usually, site visitors are also solicited to report ‘violations’. Any content of the kind—whether illegal content, flooding, spamming, advertising, hate speech or abusive language—is immediately dealt with and deleted; those who posted them are reprimanded or, after repeated violations, banned from the site.⁹ Such basic cleaning can however just achieve so much: the quality of the contents *above* the baseline of appropriate content remains an issue.

In order to tackle this thornier problem these sites have pioneered a novel approach: stories and comments can be *voted on*, usually as either a plus or a minus. As a rule, all users are entitled to vote. Note, though, that some communities require registration, and in Slashdot the right to vote obtains for a limited amount of time only. Let me elaborate these schemes. Digg has pioneered ‘digging’: if a user ‘likes’ the content, it is digged (+1), if (s)he ‘dislikes’ it, it is buried (-1). GroundReport has adopted the very same scheme. Reddit, Hacker News, and Slashdot use the more neutral wording of voting for the process: a plus if entries are found to be ‘helpful’, ‘interesting’, or ‘constructive’, a minus if they are not. Finally, NowPublic and Digital Journal only allow plus votes, for articles deemed ‘newsworthy’.

The sum total of votes then determines the *prominence* of articles on the site. By default, stories (on the front page) and comments on them (below each story) are displayed in chronological order of submission, with the most recent ones on top. Entries thus have a natural rate of decay. Voting data, fed into one algorithm or another, then force the liked items to remain longer on top of the page (countering natural decay), while at the same time forcing the disliked items—at least as far as ‘dislikes’ are part of the scheme—to plunge down the page quicker (accelerating natural decay).¹⁰ Slashdot uses a slight variation: with vote totals for items being limited to the range -1 to + 5, readers can choose their own personal threshold level to determine whether items become *visible* to them or not when they enter the site. Thus articles of bad repute are no longer punished by being pushed down the page, but by being ‘deleted’ for all practical purposes.

⁹ In Reddit, those who start a ‘subreddit’ usually are awarded the same powers for their particular subreddit.

¹⁰ Some basics of these algorithms are elaborated in <http://www.seomoz.org/blog/reddit-stumbleupon-delicious-and-hacker-news-algorithms-exposed>.

Encyclopedias and Wikinews. The remaining communities in my sample operate in proper ‘co-creation 3.0’ mode (Wikinews and encyclopedias). They also resort to basic cleaning concerning illegal or inappropriate content; in addition they have introduced elaborate quality schemes that go beyond simple voting. Let me start with h2g2 that does not use the wiki format, but just old-fashioned commenting. Tasks of basic cleaning are executed by the aforementioned volunteer ‘moderators’ (as appointed by the owner). As they phrase it, someone has to ‘clean the flotsam’. In addition, they decide on banning users who are found to be in violation. Higher up the quality scale, authors may strive for their article to appear in the ‘edited guide’. To that end, it has to be put up for public review, be recommended by a ‘scout’, and edited by ‘subeditors’. Notice that these two roles (volunteer roles one has to apply for) are intended to support authors, as opposed to control them. They are urged to operate as ‘first among equals’.

Citizendium, Wikipedia, and Wikinews have the wiki mode of production in common. This wiki is *the* place to carry out basic cleaning of illegal and inappropriate contents. Users are always on the alert regarding the content, allowed to immediately correct new edits in the wiki, and invited to ‘report’ any transgressor to the authorities concerned (constables and administrators respectively). The three communities have quite similar procedures as well for identifying and promoting high quality content (apart from normal ‘wikiing’). In Citizendium an entry may gain the status of ‘approved’. To that end, an appointed moderator (denoted ‘editor’) has to give his/her approval. This role incumbent is also to exercise ‘gentle oversight’ concerning matters of evolving content. So here again, as in h2g2, a non-authoritarian role, a ‘*primus inter pares*’. Wikinews and Wikipedia, on their part, elaborated wholly public procedures for entries to gain the status of ‘good’ article, or even ‘featured’ article: an article that meets ‘professional standards of writing, presentation and sourcing’ (http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria). As a preliminary step towards acquiring such a status an entry may be put up for public ‘peer review’ first.

Wikipedia in particular, though, over time has come to develop *additional* efforts of quality management that supplement the basic wiki mode of production. The most extended quality-watch program anywhere in our communities is to be found here. It revolves around a kind of permanent mobilization of Wikipedians who are invited to focus their energies on quality enhancement. In their fight against ‘vandalism’ basic cleaning is high on the agenda. Users can maintain personal ‘watch lists’: listed entries are kept under surveillance for new edits coming in. ‘New Pages Patrol’ is a system for users to scan newly created entries for potential problems right after they are submitted. Furthermore hundreds of software bots have been developed for the purpose. After severe testing and public discussion within the Wikipedian community, these may be ‘let loose’ on a 24 hours basis. A famous example is ClueBot, which is instructed to intervene whenever suspicious words are inserted (‘black lists’) or whole pages are deleted (<http://www.acm.uiuc.edu/~carter11/ClueBot.pdf>). The ‘new generation’ ClueBotNG operates along quite different lines: as a neural network. The bot has to be fed with both constructive and vandalist edits. By

interpreting those data it is hoped that it will learn in the long run to correctly diagnose instances of vandalism (http://en.wikipedia.org/wiki/User:ClueBot_NG).

Close watch also extends beyond the issue of vandalism. Wikipedian pages and articles are under constant surveillance whether they should be kept, deleted, merged, redirected, or ‘transwikied’ (meaning: transferred to another Wikimedia project). More importantly, in order to raise the quality of entries further, ‘WikiProjects’ (with subordinate ‘taskforces’) are formed in which people focus on specific themes (such as classical music or Australia). Each project takes the relevant entries under its wing and promotes improvement. In particular they are entrusted with the task of grading the articles in their purview by quality (7 degrees, the highest being featured and good, cf. above) and importance (4 degrees) (http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Council/Guide).

Last but not least, tools are made available to users which allow judging the credibility of entries from their revision histories. The WikiDashboard displays the edit trends of an article, and the editing activities of the most active contributors to it [14]. Furthermore, the WikiTrust extension colours words in an entry according to their ‘age’, as proxy for their credibility [1]. The colour chosen is orange: the ‘younger’ the text, the darker the orange it is rendered in. Contributors to Wikipedia may use these indicators for focused quality enhancement of entries.

Intensity of Quality Control. Before embarking on a discussion of the relationship between measures of quality control and trust, let me first put them in a comparative perspective across the whole range of the open-content communities under study. Legal rules and etiquette (3.1 and 3.2) seem to be emphasized throughout, in about equal measure. This stands to reason, since these revolve around behavioural norms of trust and respect which are universally applicable to all communities of open textual content. Not so however for quality management efforts: these are clearly intensifying if we move towards the encyclopedic end of the range. For one thing, patrolling for improper content is increasing. For another, voting schemes make way for a variety of teams that focus on quality within the wiki mode. Why this more intense mobilization?

I want to argue that this is mainly due to the different types of content involved. Social news sites aim to foster discussions; an exciting exchange of *opinions* is what they are after. These discussions, moreover, have a kind of *topicality*—in the long run their importance simply fades away. To that end, a ‘contributing 2.0’ mode is sufficient. In order to guarantee quality in this mode, scouting for inappropriate content combined with voting schemes is good enough: good discussions will remain in view (longer), while bad discussions will disappear out of sight (quicker). The natural tendency for time to produce ‘decay’ is intensified. To citizen journals, furthermore, similar arguments apply.

Encyclopedias, however, aim to render the ‘*facts*’ about particular matters. Such entries cannot be produced in one go, but have to evolve over time. Moreover, such entries are to remain *permanently* visible, ready to be consulted. For the purpose, ‘co-creation 3.0’ is the preferred mode: Wikipedia, Wikinews, and Citizendium have

chosen the interactive wiki format as their mode of production (which does not necessarily have to be so: h2g2 prefers a ‘contributing’ approach). Obviously, such a dynamic mode is susceptible to disruptions. Watching over quality therefore becomes a more urgent and permanent task. For that purpose, the wiki is turned into a space of intense patrolling and quality enhancement efforts.

Backgrounding Trust by Quality Control. After this assessment of quality management efforts across our sample of open-content communities, their connection with the default rule of full trust concerning write access finally remains to be specified. To what extent may this institutionalized trust be said to be ‘backgrounded’ by quality control? As far as this control is concerned with basic cleaning tasks, there obviously *is* a connection. Scouting for inappropriate or outright vandalist contributions—whether inside a wiki or not, whether by special volunteer patrol teams or the editorial team only, whether by humans or bots—combined with appropriate corrective action and disciplining of transgressors, is a contribution to keep the policy of full write access viable. Since disruptive contributions can always be sifted out afterwards, the gates may remain open to all. The same conclusion may apply to voting schemes which are devised to push high quality articles to a prominent and/or visible position (social news and citizen journals). To the extent that the communities involved consider it a basic aspect of quality that contributions on their sites display a minimum amount of decency and relevance, such schemes do contribute to keep their practice of full write access intact.

‘Backgrounding’ of the kind may effectively allow unrestricted and immediate write access to remain the default. Note in this respect, that whereas the campaigns for observing rules of etiquette and legal terms of use (treated above) directly impinge on the (presumed) trustworthiness of contributors, these schemes that focus on amending lower level content do not. Rather, they yield more reasons to rely on the mechanism (à la Luhmann) of *assuming* the crowds to be trustworthy and take the risk of full write access, since disruptive content can swiftly be remedied *ex post*.¹¹

The remaining efforts under the rubric of quality control which aim to promote really high quality, however, are *not* likewise connected to trust: the efforts to promote articles to the ‘edited guide’ (h2g2), to develop ‘approved’ articles (Citizendium), or to produce ‘good’ or ‘featured’ articles (Wikinews, Wikipedia). These on-going initiatives cannot be considered to support the institutional trust exhibited. Instead,

¹¹ Backgrounding trust, by a civilization campaign and/or by quality control, has an analogue in the epistemology of testimony. A default rule of accepting speakers’ utterances as true (under normal conditions) may be adhered to for non-reductionist, a priori reasons (cf. the acceptance principle). Reductionist reasoning though may also support the default rule: background evidence from our testimonial practice (like truthfulness as the norm, or reputations and sanctions) is considered to provide sufficient reasons for acceptance (for all this cf. [2]). Note though that in the classic epistemological case, backgrounding has to do with the perception of mechanisms that operate within the community of speakers who *send* the messages. In our case, it has to do with the active creation of an etiquette and/or filtering and grading mechanisms within the community of readers who *receive* the messages (and who try to incorporate the senders into their midst).

rather the reverse applies: they profit from and thrive on this policy of full write access for everybody, since it solicits a maximum inflow of contributions.

4 Discussion

My treatment of quality management (3.3) is unorthodox and bound to be controversial. In particular, critics may object that the relevant rules, regulations, and procedures cannot neatly be sorted into those that either substitute or background trust (or, in reverse fashion, profit from it); they are just variations on the same theme of concern for quality that only differ in the temporality of their application. I would argue, however, that the distinction is sound and important. My argument proceeds along the following lines.

On the one hand, schemes for quality control can aim directly at the discretion of participants and reduce it (e.g., filtering). This reduction of discretion by definition leaves less-than-full-trust to participants. As a corollary, hierarchical distinctions among participants need to be defined (such as determining who is entitled to carry out filtering, and who is to be subjected to it).¹² If so, some amount of bureaucracy proper has been introduced into the community. Note finally, that the substitution of trust as effectuated is precisely the intention of such schemes. On the other hand, measures of quality control can also buttress policies of write access for all (e.g., scouting and patrolling for vandalism, whether by humans or bots; voting schemes). Institutionalized full trust remains a viable option because of the ‘damage repair options’ that are unfolding. As long as these schemes take care to mobilize the whole community, they can avoid introducing hierarchical distinctions. Furthermore, the supporting effect on institutionalized trust towards participants is more properly a side effect; the main focus of such campaigns is quality overall. Obviously, besides these two categories, quality management initiatives can be discerned that do *not* likewise touch upon our issue of institutional trust. The above mentioned quality rating schemes are cases in point: they more properly thrive on the full-trust-policy.

The contrast can best be captured in terms of the trust assumptions embodied in the various write access policies involved. In the case of patrolling new inputs and new contributors and of voting schemes (as well as quality watch schemes more generally), the assumption of full trust of potential participants is left intact and untouched. The default remains: ‘we trust your inputs, unless proved otherwise.’ In the case of filtering which reduces the trust offered, this default is exchanged for quite another one: ‘we can no longer afford to trust your inputs, and accordingly first have to check them carefully.’

The arguments just developed can be used to show that backgrounding trust in open-content communities is very important for their functioning. The mechanism allows the full-trust write access policy to remain in force. By the same token, other available mechanisms to manage the trust problem do *not* have to be resorted to. In particular, the substitution of trust by installing bureaucratic measures can be

¹² Cf. by way of analogy the common distinction between developers and observers in open source software projects.

avoided. Before elaborating this point let me first provide some examples of steps towards bureaucracy as considered or actually taken by our communities. The Slashdot editorial team routinely scans incoming stories and only accepts the ‘most interesting, timely, and relevant’ ones for posting to the homepage (<http://slashdot.org/faq>). Furthermore, since 2009, Now Public and GroundReport filter incoming news before publication (<http://www.pbs.org/idealab/2009/06/citizen-journalism-networks-stepping-up-editorial-standards158.html>). With the former, first articles from aspiring journalists are thoroughly checked by the editorial team; subsequent ones may go live immediately and are only checked afterwards (<http://www.nowpublic.com/newsroom/community/faq>). With the latter, the site’s editors have to give their approval to all proposed articles prior to publication. Only reporters with a ‘strong track record’ are ‘upgraded’ to ‘Preferred Reporters’ who obtain full write access (<http://www.groundreport.com/info.php?action=faq&questionID=1>). In the Wikimedia circuit, finally, proposals for checking incoming edits for vandalism before publication have been circulating for several years; only after approval are edits to become publicly visible. Such review is to be carried out by experienced users. In this fashion, evidently, trust in newcomers gets restricted. The proposal is actually in force in a number of their projects from 2008 onwards: Wikipedia and Wiktionary (German versions), as well as Wikinews and Wikibooks (English versions).^{13,14}

Why then would it be important to avoid bureaucracy? The answer is that such measures may meet a chilly reception and cause unrest and trouble among community members. A conspicuous example of such unrest is the heavy contestation of the system of reviewing edits prior to publication (called Flagged Revisions) in English Wikipedia: the proposal encountered fierce resistance and finally had to be abandoned [9]. Community members may simply detest bureaucratic rules and threaten to withdraw their commitment accordingly. That is why backgrounding trust is such an important mechanism.¹⁵ Note also in this context the conspicuous role of software bots in Wikipedia. These have been and still are very active in detecting vandalism—often ahead of flesh and blood patrollers. The home page of ClueBot is full of ‘barn stars’ from co-Wikipedians, awarded since the bot had detected vandalist edits before them,

¹³ The proposal is also in force in several smaller language versions other than English, German, or French (http://meta.wikimedia.org/wiki/Flagged_Revisions).

¹⁴ In our sample it is editorial teams (social news sites, citizen journals), moderators (h2g2), constables (Citizendium) and administrators (Wikipedia, Wikinews) who hold the powers to clean up messy content and/or to discipline members. Obviously, these power holders also represent bureaucracy—the difference from the filtering measures mentioned being, that no community members seem to be opposed to such a baseline of bureaucracy.

¹⁵ Note in this respect how some of our communities try to bolster the quality process by introducing specific supportive roles that are intended as ‘prime among equals’ (cf. ‘editors’ in Citizendium, and ‘subeditors’ in h2g2). Their intention is clearly to *avoid* introducing hierarchical relations in this fashion. But trying to operate as such a ‘primus’ is walking a tight rope: in his/her performance, the role occupant may easily come to be perceived as an ordinary boss.

in just a few seconds. Reportedly it identifies, overall, about one vandalist edit per minute (over a thousand per day). Due to ClueBot and its likes, Flagged Revisions were not inevitable and the plans could be shelved.

Recently both Simon [13] and Tollefsen [15] asked themselves the question: can users rely on Wikipedia? In their affirmative answers they pointed to editorial mechanisms in place that may ensure *high quality*: the wiki format with associated talk pages [13: 348], and the procedure for acquiring ‘good’ or ‘featured’ status [15: 22]. My question has been a slightly different one: can Wikipedia trust their users and grant them unrestricted and immediate write access? No wonder my answer—though equally affirmative—turns out to be slightly different. Contributors can fully be trusted since swift procedures to filter *low quality* submissions afterwards are in place. In complementary fashion, a continuous campaign among participants promotes respect for etiquette and basic rules of law.

References

1. Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., Raman, V.: Assigning trust to Wikipedia content. In: Proceedings of the 4th International Symposium on Wikis (WikiSym 2008), Porto, Portugal, September 8-10 (2008). Obtained from <http://dx.doi.org/10.1145/1822258.1822293>
2. Adler, J.: Epistemological problems of testimony. The Stanford Encyclopedia of Philosophy (2006). Obtained from <http://plato.stanford.edu/entries/testimony-episprob/>
3. Dutton, W.H.: The wisdom of collaborative network organizations: Capturing the value of networked individuals. *Prometheus* 26(3), 211–230 (2008)
4. Fox, A.: *Beyond Contract: Work, Power and Trust Relations*. Faber and Faber, London (1974)
5. Gambetta, D.: Can we trust trust? In: Gambetta, D. (ed.) *Trust: Making and Breaking Co-operative Relations*, pp. 213–237. Blackwell, Oxford (1988)
6. Goldman, A.I.: The social epistemology of blogging. In: van den Hoven, J., Weckert, J. (eds.) *Information Technology and Moral Philosophy*, pp. 111–122. Cambridge University Press, Cambridge (2008)
7. de Laat, P.B.: How can contributors to open-source communities be trusted? On the assumption, inference, and substitution of trust. *Ethics and Information Technology* 12(4), 327–341 (2010)
8. de Laat, P.B.: Open source production of encyclopedias: Editorial policies at the intersection of organizational and epistemological trust. *Social Epistemology* 26(1), 71–103 (2012)
9. de Laat, P.B.: Coercion or empowerment? Moderation of content in Wikipedia as ‘essentially contested’ bureaucratic rules. *Ethics and Information Technology* 14(2), 123–135 (2012)
10. Luhmann, N.: *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*, 4th edn. Lucius & Lucius, Stuttgart (2000; originally 1968). English translation published in Luhmann, N. *Trust and Power*. John Wiley, Chichester (1979)
11. McGeer, V.: Trust, hope and empowerment. *Australasian Journal of Philosophy* 86(2), 237–254 (2008)
12. Pettit, P.: The cunning of trust. *Philosophy and Public Affairs* 24(3), 202–225 (1995)

13. Simon, J.: The entanglement of trust and knowledge on the Web. *Ethics and Information Technology* 12(4), 343–355 (2010)
14. Suh, B., Chi, E.H., Kittur, A., Pendleton, B.A.: Lifting the veil: Improving accountability and social transparency in Wikipedia with WikiDashboard. In: *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy, April 5-10 (2008). Obtained from <http://dx.doi.org/10.1145/13570541357214>
15. Tollefsen, D.P.: Wikipedia and the epistemology of testimony. *Episteme* 6(1), 8–24 (2009)
16. Zucker, L.G.: Production of trust: Institutional sources of economic structure, 1840-1920. *Research in Organizational Behaviour* 8, 53–111 (1986)

Author Index

- Abbassi, Zeinab 365
Abrishami, Soheila 393
Ahluwalia, Ansuya 15
Ainley, Fiona 82, 539
Allen, Beccy 517
Altamirano, Luis 162
AnaLoui, Morteza 448
Anwar, Misita 216
Aperjis, Christina 365
Azman, Norhidayah 489
- Bakhshi, Rena 406
Bandari, Roja 15
Bandyopadhyay, Somprakash 43
Banerjee, Shrabastee 43
Bavaud, François 68
Bénel, Aurélien 202
Berangi, Reza 448
Bernstein, Abraham 124
Birkholz, Julie M. 406
Bródka, Piotr 54
- Cahier, Jean-Pierre 202, 309
Carminati, Barbara 323
Cetnarowicz, Krzysztof 475
Chechev, Milen 434
Chen, Lu 379
Ciuberek, Sylwia 503
Conti, Marco 174
Crespi, Noël 1
- Deichmann, Dirk 96
De Ita, Guillermo 162
de Laat, Paul B. 543
Duell, Rob 292
- Ebrahimi, Touradj 448
El Mawas, Nour 202
- Felin, Teppo 351
Ferrari, Elena 323
Fichman, Pnina 260
Fröch, Christopher 152
- Gazzé, Davide 535
Georgiev, Petko 434
- Ghorbel, Hatem 138
Giraud-Carrier, Christophe 351
Gliwa, Bogdan 475
Grindrod, Peter 82
Groenewegen, Maartje 96
Groenewegen, Peter 406
Guex, Guillaume 68
- Harige, Ravindra 406
Higham, Desmond J. 82, 539
Hoang, Tuan-Anh 337
Huang, Allen 15
Huang, Xiaodi 1
Huberman, Bernardo A. 365
- Iamnitshi, Adriana 29
Ivanov, Ivan 448
- Jalali, Mehrdad 393
Jankowski, Jarosław 462, 503
Johanson, Graeme 216
Joshi, Somya 517
- Kayes, Imrul 29
Kazienko, Przemysław 54, 462
Kołoszczuk, Bartosz 54
Koulolias, Vasilis 517
Kozłak, Jarosław 475
Kywe, Su Mon 337, 420
- Laffin, Peter 82, 539
Lim, Ee-Peng 337, 420
Lo Duca, Angelica 535
López-López, Aurelio 162
- Ma, Xiaoyue 309
Mantzaris, Alexander V. 82, 539
Michalski, Radosław 462, 503
Millard, David E. 489
Minder, Patrick 124
Mitzlaff, Folke 531
Moyao, Yolanda 162
Mukherjee, Apratim 43
- Naghibzadeh, Mahmoud 393

- O'Donovan, John 232
 Otley, Amanda 82, 539
 Passarella, Andrea 174
 Paul, Thomas 188
 Pezzoni, Fabio 174
 Puscher, Daniel 188
 Qian, Xiaoning 29
 Roychowdhury, Vwani 15
 Schaal, Markus 232
 Schumann, Martin 152
 Sheth, Amit P. 379
 Skulimowski, Andrzej M.J. 246
 Skvoretz, John 29
 Smyth, Barry 232
 Stirling, Wynn 351
 Stopczynski, Martin 188
 Stoyanov, Dimo 96
 Strufe, Thorsten 188
 Stumme, Gerd 531
 Taylor, Steve 517
 Tesconi, Maurizio 535
 Treur, Jan 275, 292
 van Breda, Ward 275
 Van Eeckhaute, Catherine 517
 van Halteren, Aart 96
 van Steen, Maarten 406
 van Wissen, Arlette 275
 Viviani, Marco 323
 Volkamer, Melanie 188
 Wandhöfer, Timo 517
 Wang, Wenbo 379
 Wang, Zhu 110
 Weal, Mark J. 489
 Yang, Dingqi 110
 Yazdani, Sasan 448
 Yu, Zhiyong 110
 Zbieg, Anita 503
 Zhang, Daqing 110
 Zhao, Zhenzhen 1
 Zhou, Xingshe 110
 Zhu, Feida 337, 420
 Zygmunt, Anna 475