Michael Chau G. Alan Wang
Xiaolong Zheng Hsinchun Chen
Daniel Zeng Wenji Mao (Eds.)

LNCS 6749

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2011
Beijing, China, July 2011
Proceedings

Springer

Lecture Notes in Computer Science 6749

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

Michael Chau G. Alan Wang
Xiaolong Zheng Hsinchun Chen
Daniel Zeng Wenji Mao (Eds.)

Intelligence and Security Informatics

Pacific Asia Workshop, PAISI 2011
Beijing, China, July 9, 2011
Proceedings

Springer

Volume Editors

Michael Chau
The University of Hong Kong, 7/F Meng Wah Complex, Pokfulam, Hong Kong
E-mail: mchau@business.hku.hk

G. Alan Wang
Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, USA
E-mail: alanwang@vt.edu

Xiaolong Zheng
Chinese Academy of Sciences, Beijing, China
E-mail: xiaolong.zheng@ia.ac.cn

Hsinchun Chen
The University of Arizona, Tucson, AZ, USA
E-mail: hchen@eller.arizona.edu

Daniel Zeng
The University of Arizona, Tucson, AZ, USA
and Chinese Academy of Sciences, Beijing, China
E-mail: dajun.zeng@ia.ac.cn

Wenji Mao
Chinese Academy of Sciences, Beijing, China
E-mail: wenji.mao@ia.ac.cn

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-22038-8 e-ISBN 978-3-642-22039-5
DOI 10.1007/978-3-642-22039-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011929876

CR Subject Classification (1998): H.4, H.3, C.2, H.2, D.4.6, K.4.1, K.5, K.6

LNCS Sublibrary: SL 4 – Security and Cryptology

© Springer-Verlag Berlin Heidelberg 2011

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

Intelligence and security informatics (ISI) is an interdisciplinary research area concerned with the study of the development and use of advanced information technologies and systems for national, international, and societal security-related applications. In the past few years, we have witnessed that ISI experienced tremendous growth and attracted significant interest involving academic researchers in related fields as well as practitioners from both government agencies and industry.

In 2006, the Workshop on ISI was started in Singapore in conjunction with PAKDD, with most contributors and participants from the Asia-Pacific region. The second Pacific Asia Workshop on ISI, PAISI 2007, was held in Chengdu. Following that, the annual PAISI workshop was held in Taipei (2008), Bangkok, Thailand (2009), and Hyderabad, India (2010).

Building on the momentum of these ISI meetings, we held PAISI 2011 together with IEEE ISI 2011 in Beijing, China, in July 2011. PAISI 2011 brought together technical and policy researchers from a variety of fields and provided a stimulating forum for ISI researchers in Pacific Asia and other regions of the world to exchange ideas and report research progress. This volume of Springer's *Lecture Notes in Computer Science* contains 13 research papers presented at PAISI 2011. It presents a significant view on regional data sets and case studies, including Asian language processing, infectious informatics, emergence response, and cultural computing.

PAISI 2011 was jointly hosted by the Chinese Academy of Sciences, the University of Arizona, and the University of Hong Kong.

We wish to express our gratitude to all members of the Workshop Program Committee and additional reviewers who provided high-quality, constructive review comments within a tight schedule. Our special thanks go to the members of the Workshop Organizing Committee, as well as Guanpi Lai and Yanqing Gao for their help. We would like to acknowledge the excellent cooperation with Springer in the preparation of this volume. Last but not least, we thank all researchers in the ISI community for their strong and continuous support of the PAISI series and other related intelligence and security informatics research.

July 2011

Michael Chau
G. Alan Wang
Xiaolong Zheng
Hsinchun Chen
Daniel Zeng
Wenji Mao

Organization

Workshop Co-chairs

Hsinchun Chen	The University of Arizona, USA
Michael Chau	The University of Hong Kong
Daniel Zeng	Chinese Academy of Sciences and The University of Arizona
Wenji Mao	Chinese Academy of Sciences

Program Co-chairs

G. Alan Wang	Virginia Tech, USA
Xiaolong Zheng	Chinese Academy of Sciences

Program Committee

Ahmed Abbasi	University of Wisconsin-Milwaukee
Indranil Bose	The University of Hong Kong
Zhidong Cao	Chinese Academy of Sciences
Weiping Chang	Central Police University
Kuo-Tay Chen	National Taiwan University
Reynold Cheng	The University of Hong Kong
Uwe Glaesser	Simon Fraser University
Eul Gyu Im	Hanyang University
Hai Jin	Huazhong University of Science and Technology
Da-Yu Kao	Central Police University
Siddharth Kaza	Towson University
Paul Kwan	University of New England
Kai Lam	Chinese University of Hong Kong
Mark Last	Ben-Gurion University
Ickjai Lee	James Cook University
You-Lu Liao	Central Police University
Hongyan Liu	Tsinghua University
Hsin-Min Lu	The University of Arizona
Xin Luo	The University of New Mexico
Anirban Majumdar	SAP Research, SAP AG
Byron Marshall	Oregon State University
Robert Moskovitch	Ben-Gurion University
Dorbin Ng	The Chinese University of Hong Kong
Shaojie Qiao	Southwest Jiaotong University
Jialun Qin	University of Massachusetts Lowell

Shrisha Rao	International Institute of Information Technology, Bangalore
Srinath Srinivasa	International Institute of Information Technology, Bangalore
Aixin Sun	Nanyang Technological University
Paul Thompson	Dartmouth College
Jau-Hwang Wang	Central Police University
Shiuh-Jeng Wang	Central Police University
Jennifer Xu	Bentley University
Wei Zhang	Tianjin University and Tianjin University of Finance and Economics
Yilu Zhou	George Washington University
William Zhu	University of Electronic Science and Technology of China

Additional Reviewers

Liwen Sun	The University of Hong Kong
Wai Kit Wong	The University of Hong Kong
Peng Xu	Huazhong University of Science and Technology

Table of Contents

Spatio-temporal Similarity of Web User Session Trajectories and Applications in Dark Web Research

Sajimon Abraham[1] and P. Sojan Lal[2]

[1] School of Management & Business Studies, Mahatma Gandhi University, Kerala, India
[2] School of Computer Sciences, Mahatma Gandhi University, Kerala, India
sajimabraham@rediffmail.com, padikkakudy@gmail.com

Abstract. Trajectory similarity of moving objects resembles path similarity of user click-streams in web usage mining. By analyzing the URL path of each user, we are able to determine paths that are very similar and therefore effective caching strategies can be applied. In recent years, World Wide Web has been increasingly used by terrorists to spread their ideologies and web mining techniques have been used in cyber crime and terrorism research. Analysis of space and time of click stream data to establish web session similarity from historical web access log of dark web will give insights into access pattern of terrorism sites. This paper deals with the variations in applying spatio-temporal similarity measure of moving objects proposed by the authors in PAISI 2010, to web user session trajectories treating spatial similarity as a combination of structural and sequence similarity of web pages. A similarity set formation tool is proposed for web user session trajectories which has applications in mining click stream data for security related matters in dark web environment. The validity of the findings is illustrated by experimental evaluation using a web access log publically available.

Keywords: Spatio-temporal Similarity, Web based Intelligence Monitoring, Web Usage Mining, Web User session Trajectory, Dark Web.

1 Introduction

Modern monitoring systems such as GPS positioning and mobile phone networks have made available massive repositories of spatio-temporal data by recording human mobile activities, call for suitable analytical methods, capable of enabling the development of innovative, location-aware applications. Moving objects which carries location-aware devices, produce trajectory data in the form (*Oid, t, x, y*)-tuples, that contain object identifier and time-space information. In moving object applications, tracking of objects in identifying objects that moved in a similar way or followed a similar movement pattern is to find their common spatio-temporal properties.

With the rapidly increasing popularity of WWW, websites are playing a crucial role to convey knowledge to the end users. In recent years, World Wide Web has been increasingly used by terrorists to spread their ideologies[5] and Web mining techniques have been used in cyber crime and terrorism research [11][12]. Terrorist websites are so dynamic in nature as usually it suddenly emerge, the content and

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 1–14, 2011.
© Springer-Verlag Berlin Heidelberg 2011

hyperlinks are frequently modified, and they may also swiftly disappear [3]. The click stream generated by various users often follows distinct patterns and mining of the access pattern will provide knowledge. This knowledge will help in recommender system of finding browsing pattern of users, finding group of visitors with similar interest, providing customized content in site manager, categorizing visitors of dark web etc.

This paper has made an attempt to establish the resemblance between spatio-temporal similarity of network constrained moving object trajectories[7] and web user session trajectories, which will have wider applications in web based intelligence monitoring and analysis. This approach is unique in the literature illustrating the applications of network constrained moving object trajectories in web usage mining. The proposed spatial similarity includes method for measuring similarities between web pages that has taken into account not only the common URL's visited but also the structural as well as sequence alignment. Based on this a spatio-temporal similarity measure algorithm (WEBTRASIM) is proposed for similarity set of web user session trajectories. We are going to discuss two major topics in detail (i) Trajectory Similarity of Network Constrained Moving Objects which was discussed in PAISI 2010 by the authors of this article[7] and (ii) Similarity of web user session trajectories. Both concepts have wider applications in the domain of Security Informatics. The first one has applications in studying the massive flow of traffic data to monitor the traffic flow and discover traffic related patterns where as the second has applications in tracking terrorist web site visits in the emerging security informatics area of Dark Web Research[12].

The rest of the paper is organized as follows. Related work and literature are surveyed in section 2. We propose a spatio-temporal trajectory similarity measure for web user session trajectory and introducing the algorithm WEBTRASIM in section 3. The implementation and experimental evaluation of this algorithm is made in section 4 and section 5 concludes the paper with directions for future research.

2 Related Work

Most of the works on trajectory similarity of moving objects are inappropriate for similarity calculation on road networks since they use the Euclidian distance as a basis rather than the real distance on the road network [6]. This point has motivated to propose a similarity measure based on the spatio-temporal distance between two trajectories [4] using the network distance. They have proposed an algorithm of similar trajectory search which consists of two steps: a filtering phase based on the spatial similarity on the road network, and a refinement phase for discovering similar trajectories based on temporal distance. In the above work the authors propose a similarity measure based on Points Of Interest (POI) and Time Of Interest(TOI) to retrieve similar trajectories on road network spaces and not in Euclidean space. One of the recent works in trajectory similarity problem for network constrained moving objects [9] introduces new similarity measures on two trajectories that do not necessarily share any common sub-path. They define new similarity measures based on spatial and temporal characteristics of trajectories, such that the notion of similarity in space and time is well expressed, and more over they satisfy the metric properties. The above

work has identified some of the drawbacks of spatio-temporal similarity measure proposed by Hwang [4] and this issue is addressed by the authors of this paper in [7] by finding similarity percentage based on number of points a trajectory is visited, with applications in security informatics.

The first and foremost question needed to be considered in clustering web sessions is how to measure the similarity between two web sessions. Most of the previous related works apply either Euclidean distance for vector or set similarity measures, Cosine or Jaccard Coefficient. For example, a novel path clustering method was introduced [8] based on the similarity of the history of user navigation. A similar method [1] which first uses the longest common sub-sequences between two sessions through dynamic programming, and then the similarity between two sessions is defined through their relative time spent on the longest common sub-sequence. Methods have been developed [10] considering each session as a sequence and borrow the idea of sequence alignment in bioinformatics to measure similarity between sequences of page accesses. One of the recent work in web session clustering [2] uses dynamic programming technique in finding sequence similarity after finding structural similarity. In none of the papers the concept of spatio-temporal measure has been discussed in finding session similarity, which add the time component to the spatial similarity measure. There are resemblance of spatio-temporal similarity measure of network constrained trajectories and trajectories of user sessions in web usage mining [9] but there is no work in the literature exploring it in detail. This paper explore this idea proposing an algorithm for spatio-temporal similarity set formation using web access log data.

3 Trajectory Similarity of Moving Objects and Web User Sessions

Trajectory database management has emerged due to the profusion of mobile devices and positioning technologies like GPS or recently the RFID (Radio Frequency Identification). Studying people and vehicle movements within some road network is both interesting and useful especially if we could understand, manage and predict the traffic phenomenon. In essence, by studying the massive flow of traffic data we can monitor and discover traffic related patterns. There are several advantages in representing trajectory data in road network distance rather than Euclidian Distance [6]. The dimensionality reduction is one advantage where a trajectory is represented as a set of (loc, t) points instead of (x,y,t) points in Euclidian space.

3.1 Trajectory Similarity in Spatial Networks

Let T be a trajectory in a spatial network, represented as

$$T = ((b_1,t_1),(b_2,t_2),(b_3,t_3),\ldots\ldots\ldots\ldots,(b_n,t_n))$$

where n is the trajectory description length, bi denotes a location in binary string [6] and t_i is the time instance (expressed in time units, eg. seconds) that the moving object reached node bi, and $t_1 < t_i < t_m$, for each $1 < i < m$. It is assumed that moving from a node to another comes at a non-zero cost, since at least a small amount of time will be required for the transition.

The similarity types related to road network environment are

(i) Finding objects moving through certain Points of Interest
(ii) Finding Objects moving through Certain Times of Interest
(iii) Finding Objects moving through certain Points of Interest and Time of Interest.

A detailed discussion of these methods and evaluation can be found in [4] and modified algorithms in [7]. These methods and algorithms are the base line content of the section 3.2 where similarity of web user session trajectories and proposed spatio-temporal similarity algorithms are being discussed.

3.2 Similarity in Web User Session Trajectory

There are a number of resemblance between trajectories of moving objects on road network and trajectories of web click-streams. A web link structure can also assumed as network constrained since the user clicks can traverse only through predefined paths which is stored as web structure links in web storage space. By analyzing the URL path of each user, we are able to determine paths that are very similar, and therefore effective caching strategies can be applied.

3.2.1 Web User Session

The usage data is one that describes the pattern of usage of web pages like IP address, Page references, date and time of accesses. A web server log records the browsing behavior of site visitors and hence it is important source for usage information like Cookies and query data in separate logs. Cookies are tokens generated by the web server for individual client browsers in-order to automatically track the site visitors. Due to the stateless connection of HTTP protocol, tracking of individual users is not an easy task. A user is defined as the single individual accessing files from one or more web servers through a browser. A user stream session is the click-streams of page view for a single user across the entire web. In this work we consider a user session as a sequence of web clicks of individual user on a particular website.

3.2.2 Comparison of Network Constrained Moving Object Trajectory and Web User Session Trajectory

A moving object Trajectory is the path of a moving object along a road during a given period of time. A web user session trajectory is a set of user clicks during a period of time. Both are made up of individual nodes which denotes either location or url and time instant as shown in Fig 1.

Some of the variations in these two representations are

(i) A node in Moving Object Trajectory(MOT) is the name of a location on road where as a node in Web user session trajectory is a url name or web page name.

(ii)The spatial distance between two nodes in MOT is the network distance between the locations where as in the case of web user session trajectory it is the structural distance of the web page in the tree made by the link structure.

(iii) In the case of MOT the direction of object movement is bidirectional but in web user session trajectories it is unidirectional.

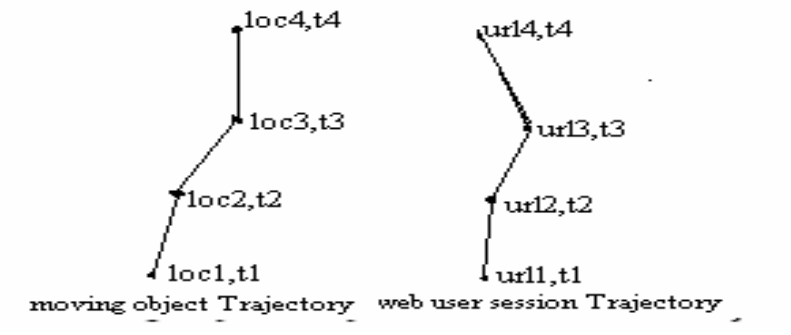


Fig. 1. Comparison of Moving Object Trajectory and Web User session Trajectory

(iv) The time between two locations in MOT is the travel time required to reach a location from the previous location which also depends on speed of movement. In the case of Web user session trajectory the time between two url's is time lag between receipt of request of previous url and the next url page. It consists of page load time and page view time as shown in Fig 2.

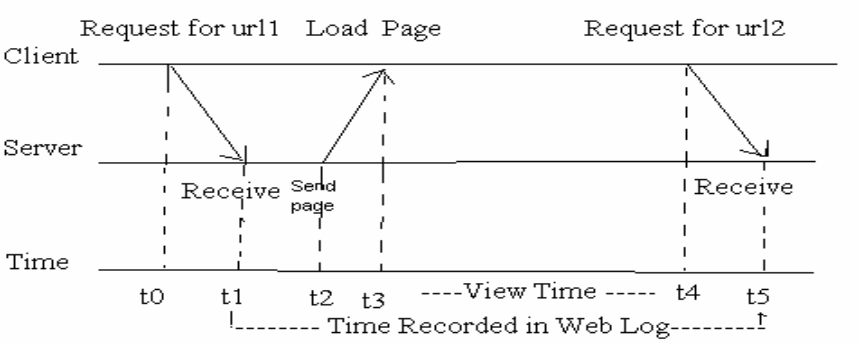


Fig. 2. Time of request between two url's

Though the web server records the time between two url's, as (t5 - t1), the actual page view time[2] of url1 will be (t4 - t3). In this paper, as we are establishing resemblance of web user session with moving object trajectory as shown in Fig 1, we consider the time recoded in web server log, instead of viewing time in finding the temporal similarity.

3.2.3 Finding Similar User Session in Web Access Logs

As discussed in the previous section, the path similarity of a user click-stream in web access log resembles the trajectory similarity problem of Moving object on Road Networks [7]. In this context Point of Interest (POI) is the interested Web URL's (We call URL's of Interest or UOI) chosen by the user as a query set. These are the URL's of specific interest to the user. For example in the case of Dark Web, which is a concept of web organization under terrorism, UOI's will be set of noted terrorism web

sites/pages in which the security people wants to find out how many people are visiting and the duration of such visits. Similarly Time of Interest (TOI) is the specific Time in which the user requesting the specified UOI's also has importance in practical situation. For example in web based security informatics one needs to find out similar viewing sessions based on a set of chosen pages visited and also based on viewing time in each page.

3.2.4 Definition of Web User Session Trajectory

The paper suggest a two step process of filtering trajectories based on spatial similarity and then refining similar trajectories based on temporal distance. For this the following definition is given, in comparison with definitions given in section 3.1.

Definition 1

Let S be a set of trajectories in a set of web user sessions, in which each trajectory is represented as

$$T = ((URL_1,t_1),(URL_2,t_2),(URL_3,t_3),\ldots\ldots\ldots\ldots,(URL_n,t_n))$$

where n is the trajectory description length, URL_i denotes a web page and ti is the time instance (expressed in time units, e.g. seconds) that the user requested for web page URL_i, and $t_1 < t_i < t_m$, for each $1 < i < m$. It is assumed that moving from a URL node to another comes at a non-zero cost, since at least a small amount of time will be required for loading the requested page by the server and viewing by the client.

3.2.5 Creation of User Session Trajectories from Web Access Logs

The goal of session identification is to divide the page accesses of each user into individual sessions. The following is the rules in identifying and distinguishing a user session from others from a web access log which we used in our experiment:

i) If there is a new user, there is a new session

ii) In one user session, if the referred page is null, there is a new session

iii) If the time between page requests exceeds a certain limit(30 minutes), it is assumed that the user is starting a new session, on the assumption that a user will generally may not spend an average of 30 minutes in a particular web page.

3.3 Finding Similar Web Usage Session Trajectories

We introduce the spatio-temporal similarity measures between two user-sessions, assuming resemblance to moving object trajectories in a constrained network. For moving object trajectories, finding spatio-temporal similarity [7] is a two step process of filtering trajectories based on spatial similarity and then refining similar trajectories based on temporal similarity. But in the case of web user sessions the spatial similarity is a combined measure of how many common URL's in both the sessions, structural similarity of web pages and the sequences of appearance of URL's in the trajectories. Therefore in this paper we are proposing a three stage process for spatio-temporal similarity of web user session trajectories as given in Fig 3.

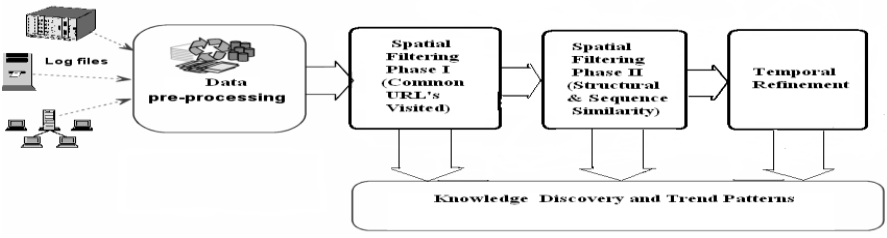


Fig. 3. Spatio-Temporal Similarity Framework

The practical importance of spatio-temporal similarity measure which includes the structural and sequence similarity is illustrated in the following case.

In Security Informatics concerning dark web environment the security officials wants to know when, how many and what sequence insurgents or suspicious people browsing through terrorist sites and based on which their common tactics can be judged.

3.3.1 Spatial Similarity Matrix Based on UOI

We introduce the similarity measures between two user-sessions, assuming both are two moving object trajectories in a constrained network [7] incorporating concepts discussed in [15]. Here the spatial similarity is measured based on three concepts.

a) Common URL's visited by two trajectories
b) Structural similarity of web pages in the trajectories
c) Sequence Similarity of user sessions.

a) Common URL's visited by two trajectories

Let T_i and T_j be two web user session trajectories. We introduce a Spatial Similarity measure $SSim_C(T_i,T_j)$ which attempt to incorporates number of URL's involved in trajectories. This similarity matrix measures the number of common URL's accessed during the two sessions relative to the total number of URL's accessed in both sessions.

Definition 2

Let T_i and T_j be two user session trajectories. The spatial similarity measure between these two trajectories is defined based on the concept that the trajectories passes through how many URL's in common.

$SSim_C(T_i,T_j)$ = Number of URL's common to Ti and Tj/Total Number of URL's in Ti and Tj.

Note that the above Similarity measure satisfies the following properties.

i. $SSim_C(T_i,T_j)$)is lies in the range [0,1]
ii. $SSim_C(T_i,T_j) = SSim_C(T_j,T_i)$

b) Structural Similarity

In order to measure Structural similarity[2], the content of pages is not considered but it is baed on the paths leading to a web page (or script). For example Suppose "/faculty/pub/journal/ieee.htm" and "faculty/pub/conference.htm" are two requested pages in web log. By assuming the page connectivity as a tree structure with root coming as home page , each level of a URL can be represented by a token from left to right in the order 0,1,2,3 etc. Thus, the token string of the full path of a URL is the concatenation of all the representative tokens of each level. We get different token strings of the URL correspond with the different requested pages traversing the tree structure of the web site. Assuming that the two token strings of the above two web pages are "0222" and "021" respectively, we compare each corresponding token of the two token strings one by one from the beginning until the first pair of tokens is different.

Marking Llonger as the larger value of (l_1, l_2) where l_1 and l_2 are lengths of the two token strings, we give weight to each level of the longer token: the last level is given weight 1, the second to the last level is given weight 2, the third to the last level is given weight 3, and so on, until the first level which is given weight Llonger. The similarity between two token strings is defined as the sum of the weights of those matching tokens divided by the sum of the total weights.

Definition 3

The Structural similarity of two web pages is defined as

$WSim_S (P_i,P_j)$ = Sum of the weights of matching token strings/Sum of the total weights.

For our example

P1= "/faculty/pub/journal/ieee.htm" and P2= "/faculty/pub/conference.htm"
Llonger = 4 ,
Token string in P1: 0 2 2 2
Token string in P2: 0 2 1
Weight of each token: 4 3 2 1

Since the first and second digit matches in P1 and P2, we take the weights of the matching locations 4 and 3 respectively.

The similarity of the two requested web pages is

$WSim_S (P1,P2)= (4 + 3)/(4+3+2+1) = 0.7.$

Note that the above Similarity measure satisfies the following properties.

i. $WSim_S (P_i,P_j) = WSim_S (P_j ,P_i)$
ii. $WSim_S (P_i,P_j)$ lies in the range [0,1]
iii. $WSim_S (P_i,P_j) = 0$ when there is no structural similarity between pages Pi and Pj
iv. $WSim_S (P_i,P_j) = 1$ when two pages Pi and Pj are exactly same.

c) Sequence Similarity of web sessions

We use a scoring system which helps to find the optimal matching between two session sequences. An optimal matching is an alignment with the highest score. The

score for the optimal matching is then used to calculate the similarity between two sessions. We find the optimal matching using dynamic programming techniques[2] to compute the similarity between web sessions as given in the following algorithm

---

Input: Pair of session sequences which constitutes each trajectories T_1,T_2
Output: Optimal Similarity score value OSSV

---

1. Matrix DPM is created with K+1 columns and N+1 rows where k and N corresponds to the sizes of T_1 and T_2 sequences respectively.
2. Align the sequences by providing gap between the sequence so that two sequences can be matched as much as possible.
3. Place sequences T1 in top of Matrix and sequences T2 on left side of the matrix.
4. Assign top left corner of matrix =0
5. For each pair of web pages /*Find optimal path */
6. Find $WSim_S(P_i,P_j)$
7. Find score = -10+30* $WSim_S(P_i,P_j)$
8. DPM cell entry=max(left entry+score,topentry+score,above left entry+score)
   /* Find path for every two sequence pair of web pages as follows */
9. If both the pages are matching
10. Diagonal move
11. Else if gap on top horizontal sequence
12. Right move
13. Else if gap on left vertical sequence
14. Down move
15. End if
16. Optimal Similarity score value(OSSV) = value at the lower right corner of the matrix DPM.
17. Return OSSV

After finding the final score for the optimal session alignment, the final similarity between sessions is computed by considering the final optimal score and the length of the two sessions. In our method, we first get the length of the shorter session lshorter, then the similarity between the two sessions is achieved through dividing the optimal matching score by 20* lshorter because the optimal score cannot be more than (20* lshorter) in our scoring system. Thus the spatial similarity measure between two web session trajectories is defined based on the concept of Sequence similarity as $Ssim_s(T_i,T_j)$ = Optimal score / (20* length of the shorter trajectory)

Note that the value of $Ssim_s(T_i,T_j)$ lies between 0 and 1.

3.3.2 Combined Spatial Similarity Matrix

The combined spatial similarity measure is obtained by

$Ssim_{cs}(T_i,T_j) = (SSim_C(T_i,T_j) + Ssim_s(T_i,T_j))/2$

$Ssim_{cs}(T_i,T_j)$ satisfies the following properties.

i. $Ssim_{cs}(T_i,T_j)$ is in [0,1]
ii. $Ssim_{cs}(T_i,T_j) = Ssim_{cs}(T_j,T_i)$

The combined spatial similarity measure is used as the distance measure in locating the objects within the similarity set.

3.3.3 Temporal Distance Measure Based on UOI and TOI

The similarity measure defined in the previous section takes into consideration only the spatial concept, which consists of structural similarity and sequence similarity. In real applications, the time information associated with each trajectory is also very important. So to measure the similarity we have to consider the concept of space and time together. Here we are considering the temporal distance by taking the difference in web page request time as shown in Fig 4.

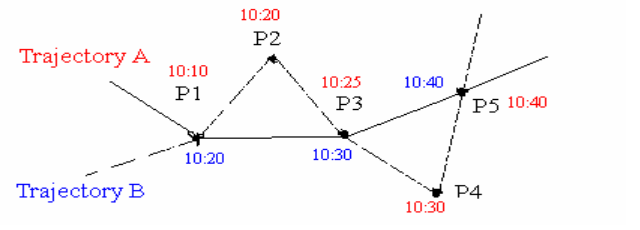


Fig. 4. Temporal Distance of Two Trajectories

Temporal distance between Trajectory A (T_A) and Trajectory B (T_B) will be $dist_T(T_A,T_B)$ = Differences in time at common URL's visited

$$P1,P3,P5 = 10+5+0 = 15$$

Definition 4

The temporal distance between two user session trajectories is defined as

$$dist_T(Ti,Tj) = \frac{1}{m*z} \sum_{m} |t(Ti.URL_k) - t(Tj.URL_k)|$$

where summation will go for only common URL's visited by both the trajectories(m) and z is a constant chosen by the user to denotes the time span in which the similarity is being measured. For example z is fixed to 24, when similarity is measuring within 24 hours in a day. Note that this distance measure also satisfies the following properties

i. $dist_T(Ti,Tj)$ is in [0,1]
ii. $dist_T(Ti,Tj) = dist_T(Tj,Ti)$

The difference between the above time similarity measure with the time similarity measure in [2] is that the later considering the viewing time similarity of a web page.

3.3.4 WEBTRSIM – Similarity Tool for Web User Session Trajectory

We propose a web session similarity set formation algorithm WEBTRASIM(Web user session Trajectory Similarity Algorithm). The objective of this algorithm is to form spatio-temporal set of trajectories, which are similar to the query trajectory formed from url's of interest(UOI) and Time of Interest (TOI).

Algorithm: To Create Spatio-Temporal Similarity Set based on UOI and TOI

Input: Input user session trajectories TR_{IN}, spatial threshold ρ, δ, temporal threshold-t σ , query trajectory tr_Q which is made up of UOI set and TOI set in required sequence; Output: Spatio-Temporal cluster set CTR containing trajectories similar to query trajectory tr_Q

---

1. Set $TR_{Candidate}$, Trout, CTR as Empty set
2. n1= number of URL's in tr_Q , , n2= number of URL's in TR_{IN},
3 For each t_r in TR_{IN},
4. n2= number of URL's in t_r , t_r.s=0
   /* Spatial filtering(first phase) based on common url's visited*/
5. For each u in tr_Q
6. If u is on tr then $t_r.s = t_r.s+1$
7. End For
8. $t_r.s= t_r.s/(n1+n2)$
9. If $t_r.s > \rho$ then $TR_{Candidate} = TR_{Candidate} \cup \{t_r\}$
10. End For
    /* Spatial filtering(second phase) based on structural and sequence similarity measure */
11. For each $t_r \in TR_{Candidate}$
12. n2=no of URL's in t_r
13. if n1<n2 then lshort=n1 else lshort=n2
14. Opscore= SSCORE-DPROG(t_r,tr_q)
15. Ssims = opscore/ (20*lshort) , tr.s=(tr,s+ Ssims)/2
16. If Ssims > δ then $TR_{OUT} = TR_{OUT} \cup \{ t_r \}$
17. End For
    /* Temporal refinement based on temporal distance */
18. For each tr in TR_{OUT},
19. n2= number of URL's in t_r , t_r.t=0, m=0
20. For each u in tr_Q
21. If u is on t_r then
22 $t_r.t = t_r.t + |t_r.u.t - tr_Q.u.t|$
23 m=m+1
24. End For
25. $dist_T(t_r, tr_Q) = 1/(m*24*60)* t_r.t$ /* one day- 24hours as the time span*/
26. $tr._{temp\text{-}dist} = dist_T(t_r, tr_Q)$
27 if $dist_T(t_r, tr_Q) < \sigma$ then $CTR = CTR \cup \{ t_r \}$
28 .End For
29 .return CTR

The algorithm will make a similarity set based on the spatio-temporal measure discussed above. One advantage of the method is that along with each trajectory in the set, spatial and temporal similarity measures(t_r.s, $tr._{temp\text{-}dist}$) are also stored which will be used later to visualize the scattering of similar user sessions within the similarity

set, useful for finding nearest neighbor trajectories based on a given UOI and TOI. A major difference of WEBTRASIM algorithm with the algorithm – WSCBIS proposed in [2] is that the later will take temporal similarity measure only when two trajectories are spatially similar. But our approach take a combined similarity measure since spatio-temporal features have to be considered in finding more accurate similarity.

4 Experimental Evaluation

To validate the efficiency of the clustering algorithm discussed above, we have made an experiment with the web server log publically available with statistics about raw data is shown in Table 1. The initial data source of our experiment is for one day on 26th Aug 2009, in which the size is 5372MB. The only cleaning step performed on this data was the removal of references to auxiliary files (image files)

Table 1. Summary of access log characteristics(Raw Data)

Item	Count
Total Requests(No of entries)	47,685
Average Request/Hour	1986.875
Total Bytes Transferred MB	304198.707
Average Bytes transferred per hour	12167.94
No of requests after pre-processing	33,864
No of Users	1,907
No of user sessions	1,986

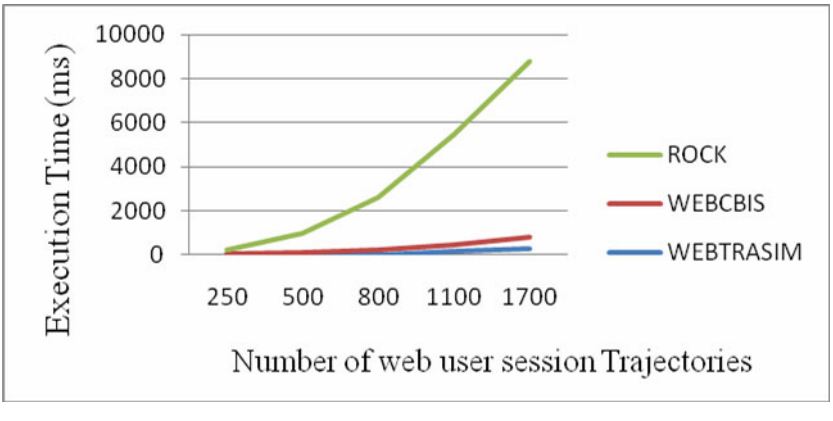


Fig. 5. Comparison of Execution Time

Our experiments were performed on a 2.8GHz Core-2 Duo Processor, 1GB of main memory, Windows XP professional using the visual Programming Language Visual Basic. As shown in above Table, after data cleaning, the number of requests declined from 47,685 to 33,864. We have applied the algorithms WSCBIS, ROCK and WEBTRASIM on the number of user sessions obtained as shown in Table 1, by giving a UOI consists of 4 urls with ρ=0.95, δ=0.95 and σ=0.05,.

Fig 5 shows that the execution time linearly increase when the number of web user session increases and the performance of WEBTRASIM is almost similar to WSCBIS even when it extends temporal similarity measure in addition to spatial similarity. Thus we claim that WEBTRASIM could become a better tool for spatio-temporal similarity of web user session trajectories.

5 Conclusion

The WWW, an effective information presentation and dissemination tool, has been widely used by terrorist groups as a communication medium. The Web presence of these terrorist groups reflects their different characteristics and may provide information about planned terrorist activities. Thus, monitoring and studying the content, structural characteristics and usage pattern of terrorist websites may help us to analyze and even to predict the activities of terrorist groups. In this paper we have proposed a method to measure the spatio-temporal similarity of web user sessions, by introducing the algorithm WEBTRASIM. Our experiments with a web access log shows that the algorithm performs equally well with similar algorithms even with the extension of temporal similarity measure. We are planning to use this similarity measure for web user session trajectory clustering to extract various browsing patterns to test with different dark web access logs.

Acknowledgements

This research is being supported by University Grants Commission(UGC) under Ministry of Human Resources, Government of India under the scheme of Grant to Academic Research for Faculty members in Universities vide order No. 37-633/2009 (SR).

References

1. Banerjee, G.A.: Clickstream clustering using weighted longest common subsequences. In: Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, Chicago, pp. 158–172 (2009)
2. Chaofeng, L.: Research on Web Session Clustering. Journal of Software 4(5) (2009)
3. Weimann, G.: How Modern Terrorism Uses the Internet, United States Institute of Peace, Special Report 116 (2004), `http://www.terror.net`
4. Hwang, J.-R., Kang, H.-Y., Li, K.-J.: Searching for Similar Trajectories on Road Networks Using Spatio-temporal Similarity. In: Manolopoulos, Y., Pokorný, J., Sellis, T.K. (eds.) ADBIS 2006. LNCS, vol. 4152, pp. 282–295. Springer, Heidelberg (2006)
5. Lee, E., Leets, L.: Persuasive storytelling by hate groups online - Examining its effects on adolescents. American Behavioral Scientist 45, 927–957 (2002)

6. Abraham, S., Lal, P.S.: Trigger based security alarming scheme for moving objects on road networks. In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 92–101. Springer, Heidelberg (2008)
7. Abraham, S., Lal, P.S.: Trajectory Similarity of Network Constrained Moving Objects and Applications to Traffic Security. In: Chen, H., Chau, M., Li, S.-h., Urs, S., Srinivasa, S., Wang, G.A. (eds.) PAISI 2010. LNCS, vol. 6122, pp. 31–43. Springer, Heidelberg (2010)
8. Shahabi, C., Zarkesh, A., Adibi, J.: Knowledge discovery from users' web-page navigation. In: Proceedings of the 7th International Workshop on Research Issues in Data Engineering (RIDE 1997) High Performance Database Management for Large-Scale Applications, pp. 20–31. IEEE Computer Society, Washington, DC, USA (1997)
9. Tiakas, E.: Searching for similar trajectories in spatial Networks. J. System Are (2009), doi:10.1016/j.jss.2008.11.832Y
10. Wang, W., Zaane, O.R.: Clustering Web sessions by sequence alignment. In: Proceedings of the 13th International Workshop on Database and Expert Systems Applications, pp. 394–398. IEEE Computer Society, Washington, DC (2002)
11. Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M.: Crime data mining: A general framework and some examples. Computer 37, 50–54 (2004)
12. Xu, J., Chen, H., Zhou, Y., Qin, J.: On the Topology of the Dark Web of Terrorist Groups. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 367–376. Springer-Verlag, Heidelberg (2006)

Specific Similarity Measure for Terrorist Networks: How Much Similar Are Terrorist Networks of Turkey?

Fatih Ozgul[1], Ahmet Celik[2], Claus Atzenbeck[3], and Zeki Erdem[4]

[1] Faculty of Computing, Science & Technology, University of Sunderland, SR6 0DD, UK
[2] Institute of Information Systems, Hof University, Germany
[3] Diyarbakir A.Gaffar Okkan Vocational School, National Police, Diyarbakir, Turkey
[4] TUBITAK- UEKAE, Information Technologies Institute,41470 Gebze, Kocaeli, Turkey
fatih.ozgul@istanbul.com, acelik@rutgers.edu,
claus.atzenbeck@iisys.de, zeki.erdem@bte.tubitak.gov.tr

Abstract. Some countries suffer from terrorism much more than others, Turkey as one of the most suffering countries who owns about a hundred terrorist groups; most of these organizations cooperate, and interchange knowledge, skills, materials used for terrorist attacks. From criminological perspective terrorist networks of Turkey are categorized into three main groups: extreme left (i.e. Marxist) networks, extreme right (i.e. Fundamentalist, Radical Islamist) networks, and separatist (i.e. ethic, racist) networks. By using their criminal history including the selection of crimes, attacking methods and modus operandi, a crime ontology is created, terrorist networks are attached to this ontology via their attacks and a similarity measure (COSM) is developed according to this ontology. Results of this similarity measure performed better than two common similarity measures; cosine and Jaccard. Results are also presented to domain experts in hierarchical clustering and they also commented as positive. Based on attributes of crimes, COSM similarity can also be applied to other types of social networks.

Keywords: Terrorist networks, similarity, crime ontology, cosine, Jaccard.

1 Introduction

Terrorist networks are everywhere. Terrorism used to be a national problem, but international community has long realized that it is a global problem. It is a well known fact that many terrorist networks get in touch, cooperate, and coordinate attacks. Here the question arise into minds that, how they choose their counterparts for cooperation? Are they similar terrorist networks? Are they coordinated from the same power centers? Do they use the same type of weapons and materials? Do they operate by using the same modus operandi patterns? Based on these questions, is it possible to find a similarity metric that can hierarchically cluster all terrorist networks in one spectrum. A similarity measure is performing better than well known other similarity metrics as well as based on domain knowledge of terrorist groups.

Foreseeing a possible cooperation of terrorist material, weapons, skills and using similar modus operandi can give an idea for possible hazard and thereby providing the

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 15–26, 2011.
© Springer-Verlag Berlin Heidelberg 2011

chance of preventing an upcoming terrorist attack. The police and intelligence officers use prospective techniques and intelligence to decide terrorist networks and their activities [9]. So, this paper investigates whether it is possible to find a similarity metric that can measure similarity between terrorist groups which can also hierarchically cluster all terrorist networks in a database. To our best of knowledge, there is no similarity measure that developed for the similarity of terrorist or criminal networks. However, there are some examples of similarity measures developed for social networks [10,12]. Some of social similarity measures use graph similarity measures such as graph isomorphism, some use Euclidian distance, vector space based cosine distance, node-edge similarity measurement score [12]. But, for similarity of terrorist networks we should be looking into similarity over their modus operandi and previous crimes of targeted terrorist networks and identifying patterns of their past criminal behavior. Comparison of past criminal behaviors can yield good results for terrorist networks similarity. Representing crime domain knowledge and criminal past behavior might be as in vector space, in relational categorical tables, or in crime ontology including behavior of selected terrorist networks. A similarity measure based on crime ontology representation is more sophisticated, and considered better than other options to export domain knowledge for training of computer about crime domain.

2 Crime Ontology for Terrorist Networks

Ontology is defined as a "formal, explicit specification of a shared conceptualization" [10]. A crime ontology which represents all terrorist attacks and all terrorist networks in a geographical area within a time segment should be developed. Theoretically, such ontology can help modeling crimes, attacks and related terrorist networks. Such crime ontology can be constructed based on classification of crimes, their attack types, and modus operandi. In such ontology, a terrorist network, with all of its attacks and modus operandi skills, can be represented as a node, and overall ontology can be represented as a graph. It is then possible to use such crime ontology to guess the extent of similarity and relationship between terrorist networks. So, two things need to be modeled here; first terrorist networks, second overall crimes and attacks which are committed by these terrorist networks. Overall, terrorist attacks and network names can be merged in a graph-based-ontology.

3 Existing Ontology Examples for Crime

There are existing examples of using ontology for analysis of crime. As one of the existing systems that use ontology, Terrorist Modus Operandi System, TMODS [8] used terrorism threat cases as input graphs and matched them against crime ontology (Fig.1).

Not as entirely crime ontology but as legal core ontology, there are some examples are available. Such a legal ontology has been built up for several purposes: information retrieval, statute retrieval, normative linking, knowledge management, or legal reasoning. As one example for them, Kingston, Schafer, and Vandenberghe [4] presents a legal modeling approach, which looks at a three-tier approach that describes

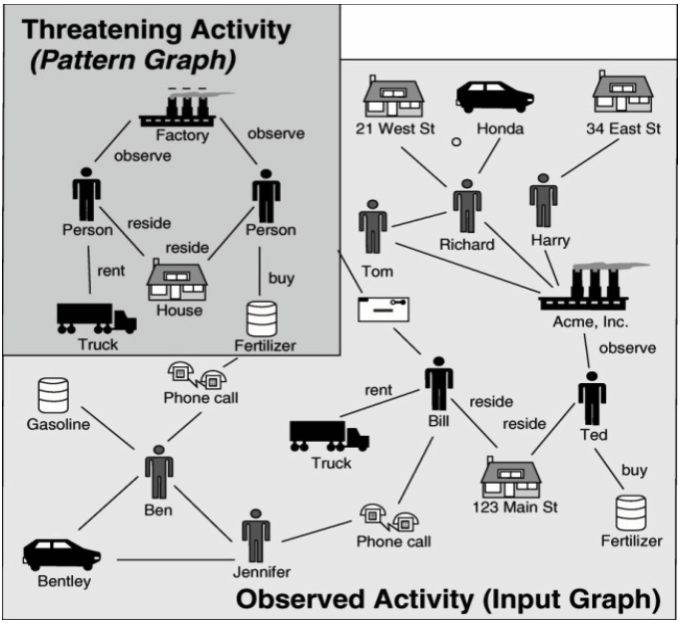


Fig. 1. TMODS ontology matching showcase. A threat graph is matched against available ontology in TMODS. Source: Marcus et al. [8].

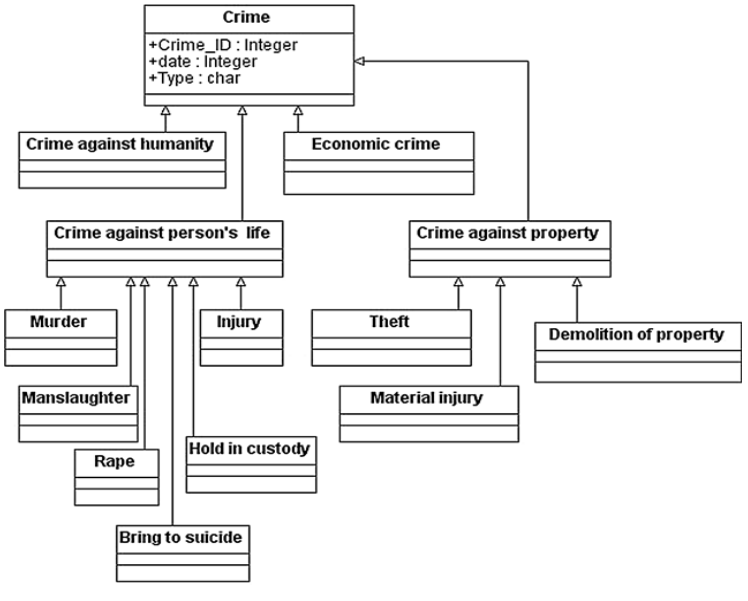


Fig. 2. Lithuanian Crime Ontology prepared by Dzemydiene and Kazemikaitiene [3]

the investigative process as the investigative process as hypothesis (tier one), law (tier two), and evidence (tier three). The emphasis of this approach is to model a specific crime. The primary aim of such legal ontology is to provide a means of translating financial security fraud across the laws of multiple nations in the European Union. As another example for legal ontology, Asaro [1] describes a concept of mapping evidence to the elements of a crime so that judges can visualize whether a guilty verdict is supported. In his doctoral research, Lazarevich [6] has also studied ontologies and probabilistic relational models, which are counterpart of Bayesian networks in statistical relational learning, to aid in cyber crime investigation and decision support. Probabilistic were better than ontologies to provide decision support in determining probable clause. Among existing examples of ontology, Dzemydiene and Kazemikaitiene [3], in a more recent work, prepared Lithuanian crime ontology to help prepare a framework and ensure collection, accumulation, storage, treatment, and transmission of important investigation information, in a proper form, which establishes conditions to make optimal decisions in the investigation of crimes (Fig.2).

Based on the existing examples of crime ontology, our crime ontology model is offered in the following section.

4 The Model

In our model, we prepared crime ontology in Direct Acyclic Graph (DAG) form. Then we developed link weights for every two nodes in this ontology, finally distances from every node to every node in this ontology are calculated, the distance for every selected two terrorist networks are found. Less distant terrorist networks are accepted are more similar networks, more distant terrorist networks are accepted as less similar networks.

4.1 Crime Ontology

Crime ontology is developed as DAG. In this DAG, similar to Dzemydiene and Kazemikaitiene's work [3], on the top level the crime node is created (Fig. 3). Below this node there is crime class node, which is terrorism in our case, but crime classes might be other internationally recognized types of crimes [7], are linked to top crime node. Following this level, we created modus operandi nodes which are linked to crime class or terrorism node. Finally as the last level, we put defined crimes nodes, which are specially defined in accordance with national (i.e. Turkish) criminal codes, and they are linked to modus operandi node. After creation of such ontology model, terrorist networks, represented as nodes, are linked to defined crime nodes as the bottom level nodes in ontology. Based on committed crimes (i.e. defined crimes nodes) and used modus operandi methods (i.e. modus operandi nodes), terrorist networks are linked to related nodes in crime ontology. Two terrorist networks (i.e. Terrorist Network#1 and Terrorist Network#2) which are similar by committing the same crime (i.e. defined crime) are represented on the left (Fig.3). Another similarity case is two networks didn't commit the same crime but used the same modus operandi as presented on the right (Fig.3). In both cases we need link weights for every two nodes in this DAG, in order to calculate the distances between every pair of nodes, including selected two terrorist networks' nodes.

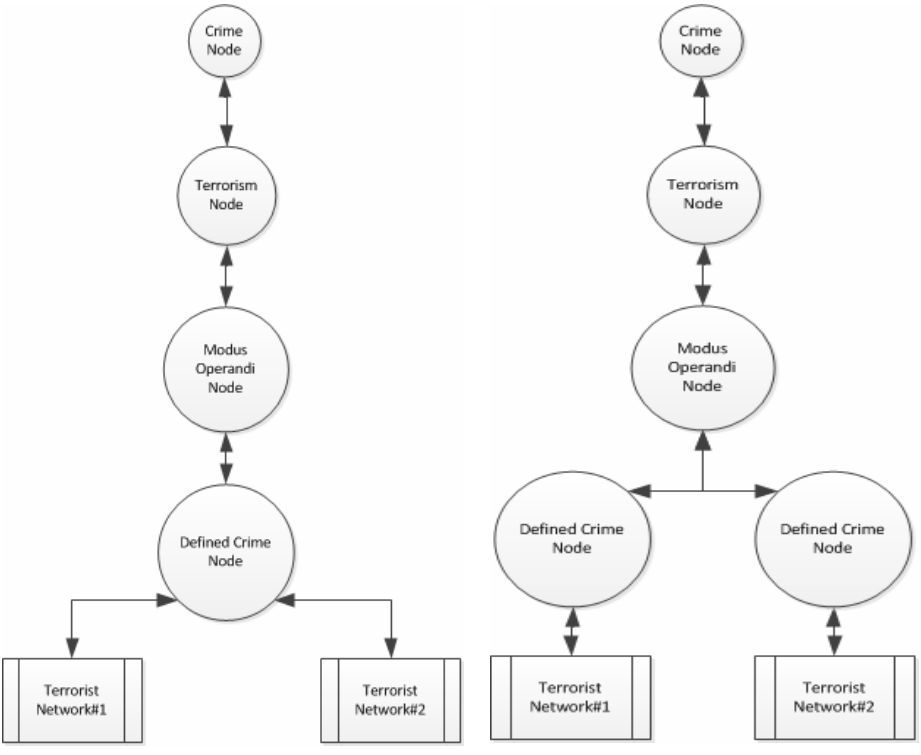


Fig. 3. Crime ontology model representing two terrorist networks committed the same crime (left). Model representing two terrorist networks committed different crimes with the same modus operandi (right).

4.2 Link Weighting in Crime Ontology

Link weights between nodes Crime to Terrorism, Terrorism to Modus Operandi, and Modus Operandi to Crime nodes are attained as 1. Conceptually, they are all in equal distance, so they are linked as in identical distances. Link weights from terrorist network nodes to defined crime nodes vary. They depend on whether a terrorist network committed more than one crime, whether they committed many types of crimes, whether they used many types of modus operandi while they conducting their attacks. Link weights for all links in ontology are presented below (in Fig.4).

For instance, for selected two networks in Fig.4, weights between networks and crimes are calculated as follows:

$$W(\text{Terrorist Network\#1, Crime\#1}) = 1 - \log\left(1/\sum \text{Terrorist Network\#1, Crime\#1}\right) \quad (1)$$

$$W(\text{Terrorist Network\#1, Crime\#2}) = 1 - \log\left(1/\sum \text{Terrorist Network\#1, Crime\#2}\right) \quad (2)$$

$$W(\text{Terrorist Network\#2, Crime\#2}) = 1 - \log\left(1/\sum \text{Terrorist Network\#2, Crime\#2}\right) \quad (3)$$

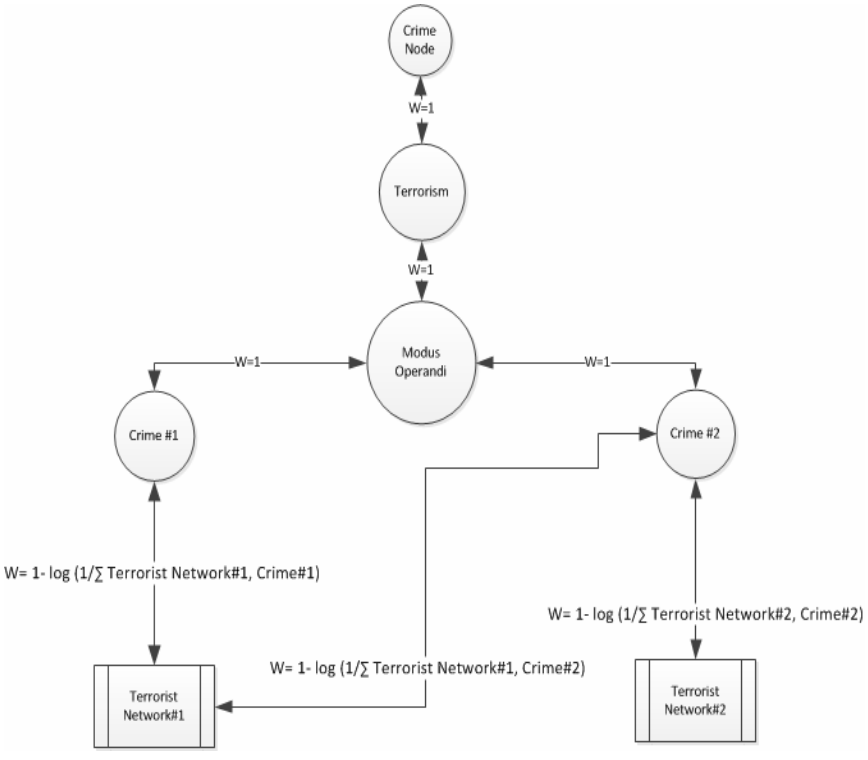


Fig. 4. Link weights between terrorist network and crime nodes. Terrorist network#1 committed two different crimes; Crime #1 and Crime#2. Terrorist network#2 committed only Crime#2 crime. Three link weights; two for Terrorist network#1, one for Terrorist Network#2 are found separately.

4.3 Similarity Calculation between Terrorist Networks

The last step is finding the similarity between pair of terrorist networks. This is realized by measuring distances between all pairs of nodes using Johnson's all pairs shortest path algorithm [2]. Given the example networks and ontology in Fig.4, if there is crime similarity between networks and W (Terrorist Network#1, Crime#2) exists, and then similarity is calculated as;

$$\textit{Similarity (Terrorist Network\#1, Terrorist Network\#2)} = \sum \textit{W (Terrorist Network\#1, Crime\#2)} \tag{4}$$

If there is only modus operandi similarity between networks and W (Terrorist Network#1, Crime#2) doesn't exist, then similarity is calculated as;

$$\textit{Similarity (Terrorist Network\#1, Terrorist Network\#2)} = \sum \textit{W (Terrorist Network\#1, Crime\#1), W (Crime\#1, Modus Operandi), W (Modus Operandi, Crime\#2), W (Crime\#2, Terrorist Network\#2)} \tag{5}$$

For all pairs of terrorist networks, we obtain a similarity score; as a result we obtain a similarity matrix. A similarity matrix is used for hierarchical clustering of networks, and finally we get a dendrogram of all networks. Since our model is based on crime ontology it is called as Crime Ontology Similarity Model (COSM) and experiments made for all terrorist networks in Turkey, based on their attacks and crime histories. Experimental results using terrorist networks data of Turkey is exhibited in the following section.

5 Experiments

Turkey data set includes all terrorist attacks happened between 1975 and 2005. This data set in obtained as public domain from START database [5]. There are eighty six terrorist networks in Turkey dataset. From criminological perspective, terrorist networks of Turkey are categorized into three main groups: extreme left (i.e. Marxist) networks, extreme right (i.e. Fundamentalist, Radical Islamist) networks, and separatist (i.e. ethic, racist) networks.

Some of the terrorist networks are of ethnic origin such as Armenian Secret Army (ASALA), Kurdish Hawks, Kurdish Workers Party (PKK), PUK/PKK (fraction of PKK), HPG/PKK (a fraction of PKK). Some of them are left extremists such as Revolution Youth, DHKP/C, MLKP-FESK, MLKP, TKP/ML, and TKP/ML TIKKO.

Table 1. Distance matrix for major terrorist networks in Turkey

	Al qaida	Armenian secret army	Revolution Youth	DHKPC	Kurdish Hawks	Great eastern Islamic raiders	Hezballah	Islamic Jihad	PKK	MLKP-FESK	MLKP	PUK PKK	HPG PKK	TKP-ML	TKP-M	Turkish Hezballah	TKP-ML TIKKO
Al qaida	0	3,447	2,301	3	2,301	3,663	4,301	2,602	4,3	2,602	2,301	4,78	2,3	2,301	4,3	4,301	2,301
Armenian secret army	3,447	0	3,146	2,954	3,146	2,778	4,477	2,477	4,78	3,447	3,146	4,48	3,15	3,146	4,48	4,477	3,146
Revolution Youth	2,301	3,146	0	2,699	2	3,362	4	2,301	4	2,301	2	4,48	2	2	4	4	2
DHKPC	3	2,954	2,699	0	2,699	2,301	2,301	2,477	3,56	2,602	2,699	4	2,3	2,699	2,3	2,301	2
Kurdish Hawks	2,301	3,146	2	2,699	0	3,362	4	2,301	4	2,301	2	4,48	2	2	4	4	2
Great eastern Islamic raiders	3,663	2,778	3,362	2,301	3,362	0	2	2,301	3,26	3,663	3,362	4	2	3,362	2	2	2,301
Hezballah	4,301	4,477	4	2,301	4	2	0	3,204	3,26	3,204	4	4	2	4	2	2	2,301
Islamic Jihad	2,602	2,477	2,301	2,477	2,301	2,301	3,204	0	4,3	2,602	2,301	4	2,3	2,301	4	4	2,301
PKK	4,301	4,778	4	3,556	4	3,255	3,255	4,301	0	4,301	4	3,53	3,26	4	3,26	3,255	2,602
MLKP-FESK	2,602	3,447	2,301	2,602	2,301	3,663	3,204	2,602	4,3	0	2,301	4,6	2,3	2,301	4,3	4,301	2,301
MLKP	2,301	3,146	2	2,699	2	3,362	4	2,301	4	2,301	0	4,48	2	2	4	4	2
PUK-PKK	4,778	4,477	4,477	4	4,477	4	4	4	3,53	4,602	4,477	0	4	4,477	4	4	4
HPG-PKK	2,301	3,146	2	2,301	2	2	2	2,301	3,26	2,301	2	4	0	2	2	2	2
TKP-ML	4,301	4,477	4	2,301	4	2	2	4	3,26	4,301	4	4	2	4	0	2	2,301
TKP-M	3,146	3,176	2,845	2,602	2,845	3	3,204	2,699	3,83	3,146	2,845	2,3	2,85	2,845	4,3	4,301	2,602
Turkish Heballah	4,301	4,477	4	2,301	4	2	2	4	3,26	4,301	4	4	2	4	2	0	2,301
TKP-ML TIKKO	2,301	3,146	2	2	2	2,301	2,301	2,301	2,6	2,301	2	4	2	2	2,3	2,301	0

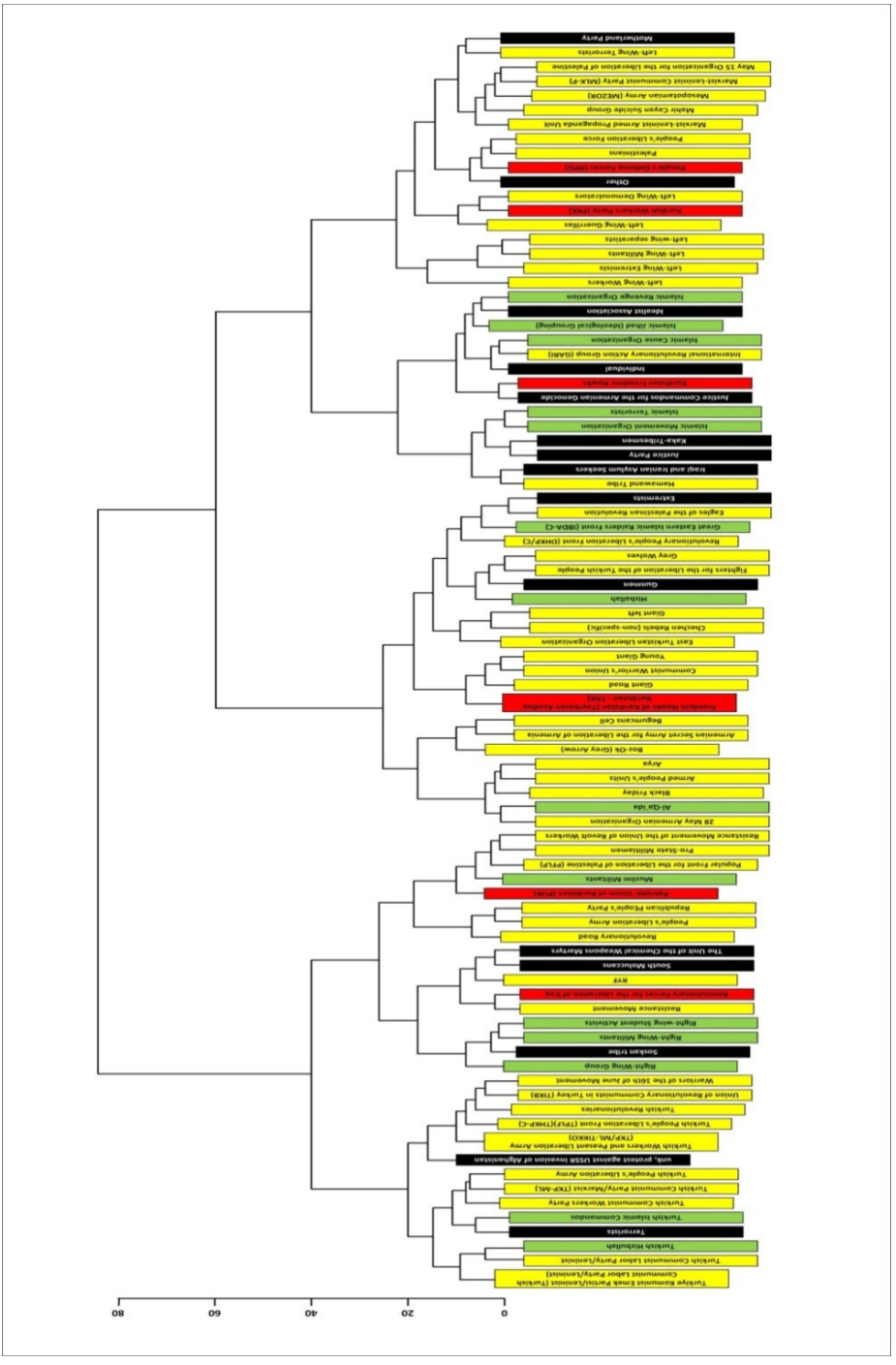


Fig. 5. Turkey terrorist networks similarity presented in hierarchical clustering dendrogram

Some of them are religious fundamentalists and racists groups such as Al Qaida, Great Eastern Islamic Raiders, Hezballah, Islamic Jihad, and Turkish Hezbollah. Major terrorist networks' COSM scores as distance matrix is presented in the table 1 below. The highest distance score is between Armenian Secret Army (ASALA) and Kurdish Workers Party (PKK) as 4.78. Following highest distance scores are between MLKP-FESK and PUK-PKK as 4.602. Another one is between Turkish Hezballah and PUK-PKK as 4.78 and also Turkish Hezballah and Al Qaida as 4.301. Similarity representation is also given in hierarchical clustering dendrogram in figure 5.

As presented in Figure 5, terrorist networks are displayed in different colors. Separatist movements of ethnic origin such as Armenian Secret Army (ASALA), Kurdish Hawks, Kurdish Workers Party (PKK), PUK/PKK (fraction of PKK), HPG/PKK (a fraction of PKK) are represented in red. Left extremists movements such as Revolution Youth, DHKP/C, MLKP-FESK, MLKP, TKP/ML, and TKP/ML TIKKO are represented in yellow. Extreme right, religious, fundamentalists and racists groups such as Al Qaida, Great Eastern Islamic Raiders, Hezballah, Islamic Jihad, and Turkish Hezbollah are represented in green. Terrorist movements and networks represented in black are the ones which couldn't be categorized into three categories.

According to criminal network similarity results in dendrogram, extreme left terrorist networks are too many in members and fractions, so they are closely represented in hierarchical tree. All of these terrorist networks are left-extremist groups and their ideology is very similar. On the other hand, Extreme religious-right wing terrorist networks are in the right-side of dendrogram.

6 Evaluation

MDS plots of three similarity scores are presented. COSM result for Turkey terrorist networks are exhibited in figure 6. There are seven main clusters in the MDS plot. They are fairly clustered and the clusters on the top and two clusters on the right are mainly extreme left and ethnically originated separatist groups. The cluster on the top-left including right wing networks, the cluster in central left also contains extreme right wing networks. Two clusters on the bottom-left contain various networks not specific to any particular type.

Cosine similarity result for Turkey terrorist networks are exhibited in figure 7. There are six main clusters in the MDS plot. They are mixed clustered and not giving particular type of similar networks. Based on this MDS plot, COSM outperforms better than cosine similarity. Jaccard similarity result for Turkey terrorist networks are exhibited in figure 8. There are six main clusters in the MDS plot. Just like cosine similarity MDS plot, they are mixed clustered and not giving particular type of similar networks. Based on this MDS plot, COSM outperforms better than Jaccard similarity.

For Turkey terrorist networks dataset results of COSM, good feedback is provided from domain experts in Turkey. Similar networks are approximated whereas different networks represented as distant. Although accepted as intuitive, terrorism experts contacted said that most of the similarities provided by COSM are depending on few similarities based on crimes and modus operandi. They said COSM can be further developed by adding more attributes, such as target/attack site selection, location selection, date/time selection of terrorists; COSM shouldn't be only based on crime and modus operandi similarity.

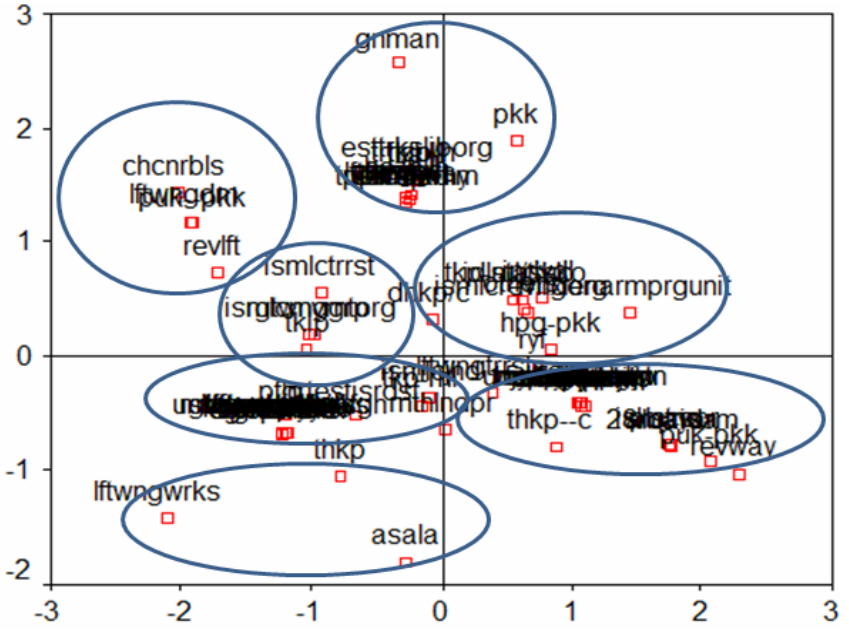


Fig. 6. MDS plot for COSM

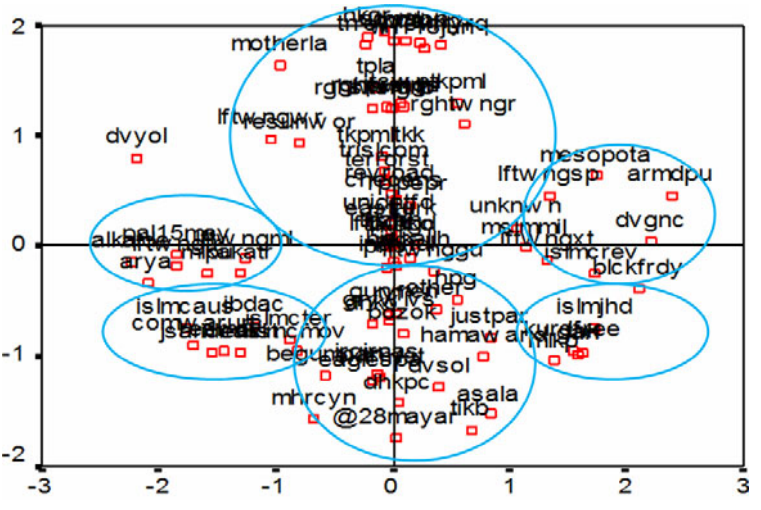


Fig. 7. MDS plot for cosine

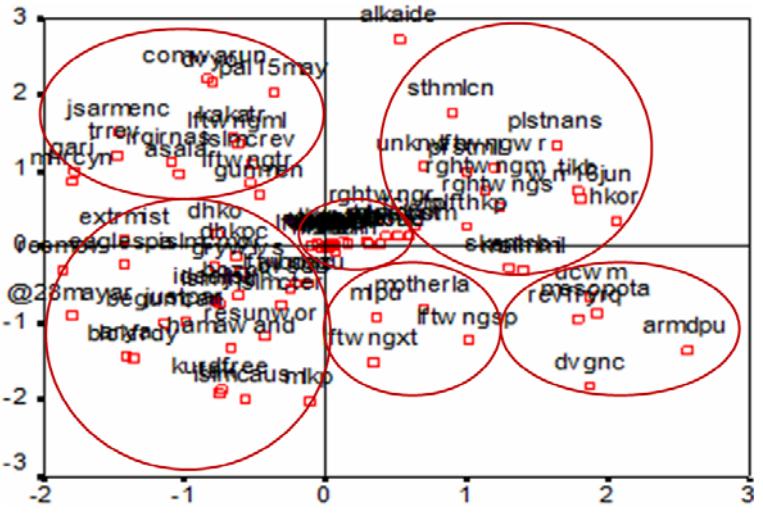


Fig. 8. MDS plot for Jaccard

7 Conclusion

Many terrorist networks get in touch, cooperate, and coordinate attacks, so most of similar terrorist networks use the same type of weapons and modus operandi patterns. So**,** is it possible to find a similarity metric that can hierarchically cluster all terrorist networks in one spectrum? And is it possible to find a similarity metric that can measure similarity between terrorist groups which can also hierarchically cluster all terrorist networks in a database? The answer is affirmative. COSM similarity measure is performing better than well known cosine and Jaccard similarity metrics as well as based on domain knowledge of terrorist groups. COSM tried to find similar terrorist networks based on crime and modus operandi similarity. Based on attributes of social networks data, a similarity metric can be developed using attribute similarity, like COSM does for terrorist networks.

References

1. Asaro, C., Biasiotti, M.A., Guidotti, P., Papini, Sagri, M., Tiscornia, D.: A Domain Ontology: Italian Crime Ontology. In: ICAIL 2003 Workshop on Legal Ontologies & Web Based Legal Information Management (2003)
2. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. (2001)
3. Dzemydeiene, D., Kazemikaitiene, E.: Ontology-based decision supporting system for crime investigation processes. Information Systems Development: Advances in Theory, Practice and Education, 427–438 (2005)
4. Kingston, J., Schafer, B., Vandenberghe, W.: Towards a Financial Fraud Ontology: A Legal Modelling Approach. Journal of Artificial Intelligence and Law 12(4), 419–446 (2004) ISSN 0924-8463

5. LaFree, G., Dugan, L.: Global Terrorism Database 1.1 and II, 1970-1997, and 1998-2004 ICPSR22541-v1, ICPSR22600- v2. Inter University Consortium for Political and Social Research, Ann Arbor (2008)
6. Lazarevich, A.: Use of ontologies and probabilistic relational models to aid in cyber crime investigation decision support. George Mason University, Washington D.C., US, Ph.D. thesis (2005)
7. Macrimeanalysts, Massachusetts Association of Crime Analysts Home Page (2010), http://www.macrimeanalysts.com/articles/classification.pdf (accessed at February 12, 2010)
8. Marcus, S.M., Moy, M., Coffman, T.: Social Network Analysis in Mining Graph Data, Cook, D.J., Holder, L.B. (eds.). John Wiley &Sons, Inc., Chichester (2007)
9. Skillicorn, D.: Knowledge Discovery for Counterterrorism and Law Enforcement. CRC Press, Boca Raton (2009)
10. Spertus, E., Sahami, M., Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD 2005), pp. 678–684. ACM, New York (2005)
11. Protégé, Standford University (2010), http://protege.stanford.edu/overview/ (accessed at February 12, 2010)
12. Wasserman, S., Faust, K.: Social Network Analysis Methods and Applications. Cambridge University Press, Cambridge (1994)

Social Network Analysis Based on Authorship Identification for Cybercrime Investigation

Jianbin Ma[1], Guifa Teng[1], Shuhui Chang[1], Xiaoru Zhang[2], and Ke Xiao[1]

[1] College of Information Science and Technology, Agricultural University of Hebei, Baoding 071001, China
[2] Audio and Computer Center, The Central Institute for Correctional Police, Baoding 071000, China
majianbin@hebau.edu.cn

Abstract. With the rapid development of Internet, cybercrime by means of Internet become serious. Mining their communication can be used to discover some latent criminal activities. Social network analysis is used for understanding their social communication status. But promulgating information on Internet is free. Criminals can hide in any corner by anonymity or forging their personal information. Authorship identification methods based on stylometry is necessary to identify criminal's real authorship. So in this paper, social network analysis methods based on e-mail and blog was provided. Authorship identification using to judge authorship's authenticity was proposed. Experiments on e-mail and blog dataset's social network analysis were demonstrated.

Keywords: social network analysis, authorship analysis, cybercrime investigation.

1 Introduction

With rapid development of Internet, people can communicate each other without geographical boundary. E-mail facilitates digital messages' exchanging from one author to one or more recipients. People can perform functions such as uploading and downloading software and data, reading news and bulletins, and exchanging messages with other users by means of BBS. A web forum is a virtual platform for expressing personal and communal opinions, comments, experience, thoughts, and sentiments [1]. Other Internet services such as blog, microblog etc are used for exchanging information each other.

Every thing has two sides. Internet provides convenience, at the same time negative impact does harm to people or society seriously. Antisocial mail, fraud mail, racketeering mail, terroristic threatening mail, pornographic mail appear in everyone's mailbox frequently. Some terrorists use forums to recruit members, broadcast antisocial information, and upload obscene pictures. Mining their communication can be used to discover some latent criminal activities. Building social network by analyzing their communication information is used for detecting latent cybercrime members, which can assist court to collect criminal evidence. Internet is a free and open place. Its anonymous nature

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 27–35, 2011.
© Springer-Verlag Berlin Heidelberg 2011

facilitates criminal hiding in any corner by anonymity or forging their personal information. Social networks built by analyzing e-mail's mailbox address and BBS users when registering are unbelievable. Authorship identification methods based on stylometry is necessary to identify criminals' real authorship. We had done a lot of research on authorship identification formerly. The techniques could be used for identifying authorship. In this paper, social network analysis methods based on e-mail and blog was provided. Authorship identification using to judge authorship's authenticity was proposed. Experiments on e-mail and blog dataset's social network were demonstrated.

The remainder of the paper is organized as follows. Section 2 presents a general review of social network analysis and authorship identification. Section 3 describes our research design. Section 4 is social network analysis methods based on e-mail and blog. Section 5 provides our experiments. Section 6 is the conclusions of the paper.

2 Related Works

2.1 Social Network Analysis

Social network is defined as social structure of individuals in terms of network theory based on a common relation of interest, friendship, trust, etc. Social network analysis is the study of social networks to understand their structure and behavior. A social network is a graph, G=(V,E), where V is a set of nodes representing persons and E is a set of edges (V*V) representing the relationship between the corresponding persons. Social network analysis views social relationship in terms of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors. Social network are often visualized in a social network graph, where nodes are the points and ties are the lines.

Social network analysis has emerged as a key technique in modern sociology. The research of social network can be traced back to the late 1800s. A summary of the progress of social network and social network analysis has been written by Linton Freeman [2]. The earlier researches focused on the theory of social structure[3][4]. Originally social network analysis has been applied to the studies of social movement, adolescent behavior, and disease transmission[5][6]. In recent years several researchers began to focus on social network analysis for discovering terrorist organization by analyzing web information[7]-[9]. For the particularity of e-mail, e-mail data could be transformed into social network easily by analyzing the *header* information such as *from* field and *to* field[10][11]. But the *header* information of e-mails is unauthentic and is always ignored or forged. Moreover, the phenomena of several persons using the same e-mail address is common.The user's information of BBS or Blog when registrating is always foged to avoid investigation. So authorship identification methods based on stylometry is necessary to identify criminals' real authorship.

2.2 Social Network Community

Social network is complex network. It has complex network's characteristic such as small world and scale-free. A lot of studies indicated that some networks were

isomerous. That is to say, the connection is more among same type of nodes than different type of nodes. Community is sub-graph composed of same type of nodes and ties between nodes. Connection is very tight within community and sparse between communities.

The research of network community has close relation with graph partitioning problem in computer sciences and hierarchical Clustering in sociology. To find out community in complex network, some algorithms were proposed. The classical algorithms were Kernighan-Lin algorithm[12] based on computer science, agglomerative method[13] and divisive method[14] based on sociology.

2.3 Authorship Identification

Based on the theory of stylometry, authorship identification is a process of matching unidentified writings to an author based on the similarity of writing styles between the known works of the authors and unidentified pieces. Abbasi and Chen presented a comprehensive analysis on the stylistics features, namely, lexical, syntactical, structural, content-specific and idiosyncratic features[15][16]. Zheng (2006,2003) analyzed the authorship of web-forum, using a comprehensive set of lexical, syntactical, structural features, and content-specific features[17][18]. Teng and Ma [19]-[22] researched on authorship identification methods on e-mail and web information. Various writing-style features including linguistic features, structural features, and format features were analyzed. Support vector machine algorithm was adopted to learn the author's writing features.

3 Research Design

3.1 The Framework of Social Network Analysis Methods

Figure 1 presents the framework of our social networks analysis methods. There are four major components, namely, information extraction, authorship analysis, social network analysis, and social network visualization.

(1) Information extraction: To build social network, information extraction from e-mail, blog, BBS, academic paper etc is the first step to do. By mining the information, social relationship can be obtained.
(2) Authorship identification: Internet user's personal information is always inauthentic or spurious, our authorship identification methods that analyzed web information author's writing features and adopted machine learning algorithm to identify web information's real authorship were used to judge authorship's authenticity.
(3) Social network analysis: By analyzing web information, mining author's social intercourse, social network can be built.
(4) Network visualization: Using the technique of graph, social relationship is demonstrated in a social network graph.

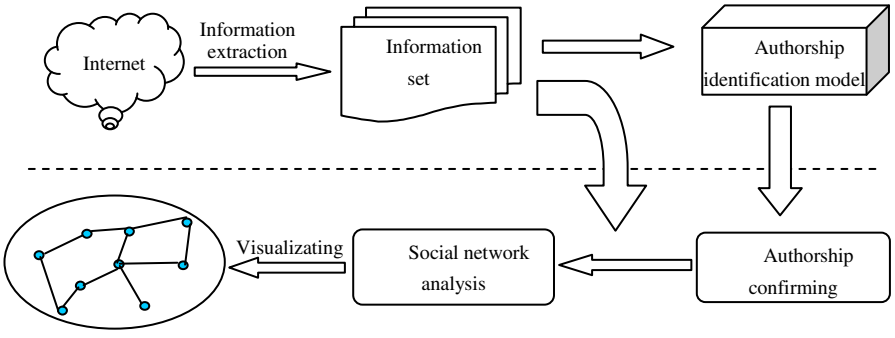


Fig. 1. The framework of social network analysis

3.2 Authorship Identification Methods

Everyone has own handwriting, which can be used to identify authorship. Authors of web information have their inherent writing habits. The writing habits can not be changed easily, which embody writing features such as usage of certain words, the length of sentences and paragraphs, and the format of the text etc.

In our study, three types of feature including linguistic features, structural features, and format features were extracted. Linguistic features were based on the frequency of certain words. tf-idf techniques were used to calculate the weight of linguistic features. Information gain(IG) methods were used to select more effective linguistic features. Structural features reflect the structure of article, syntax, and morphology. 10 structural features, 30 punctuations features including Chinese and English punctuations, and 12 common used part of speech features were extracted. As a special type of web information, e-mail has format features besides linguistic and structural features. Same as written letters, e-mail should obey the format of letters. Appellation, honorific, name, and date are the writing format of e-mail. Format features were treated as pattern. For example, the honorific "best regards", "best wishes", and "yours sincerely" are different pattern.

Machine learning techniques including decision trees, neural network, and support vector machine are the most common analytical approaches used for authorship identification in recent years. The distinctive advantage of the support vector machine is its ability to process many high-dimensional applications. In our study, support vector machines were used for learning the authors' writing features. Authorship identification model was gained.

3.3 Cybercrime Investigation Methods Based on Social Network

Cyber criminals are assembled into one virtual community. Social network analysis provides one tool to search social member's social relationship. If one criminal promulgate illegal web information, the criminal's social relationship can be investigated. Exhaustive investigation on one criminal's social relationship can detect lots of latent cybercrime group, which provide one technical means to restraint cybercrime. The social network investigation methods were showed in figure 2.

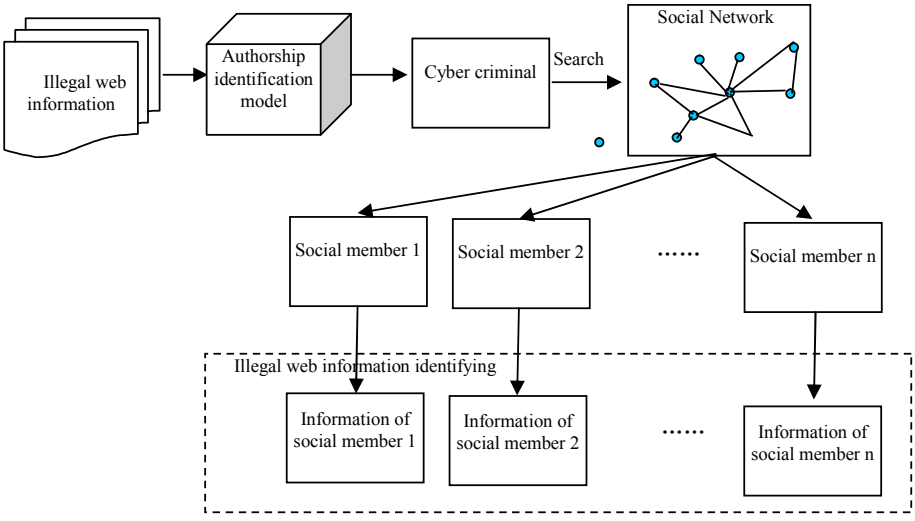


Fig. 2. The method of social networks investigation

4 Social Network Analysis Methods

4.1 E-mail's Social Network Analysis

With rapid development of Internet, e-mail has replaced traditional communication means such as letters or telegraphs gradually and became an expedient and economical means of communication. E-mail brings together participants by their communication activities each other, which suggest their social relationships.

An e-mail message consists of two components, the message *header*, and the message *body*. The message *header* containing control information, including, minimally, an originator's e-mail address and one or more recipient addresses was used for discovering the social relationship pattern between participants. The message *body* was used for extracting the authors' writing style, which was mined and formed authorship identification model.

The message *header* should include at least the following fields:

- *From*: The sender's e-mail address, and optionally the name of the author. For example: *From: admin@internet.com*. Like the envelope *from* address, e-mail *header* can themselves be forged. The e-mail is supposedly sent "From" *admin@internet.com*, but maybe in reality, that's an address forged by the *Sobig.F worm*, stolen for the purpose of masking the real authorship.
- *To*: The e-mail address of the ultimate destination, and optionally name of the message's recipient.
- *Subject*: A brief summary of the topic of the message.
- *Date*: The local time and date when the message was written.

By analyzing *from* field and *to* field, the communication relationship can be obtained. In figure 3, we can see that the sender f ties to t_1, t_2, t_3, t_4, and t_5, because sender f often sends e-mails to the five receivers. The social network is a directed graph. So the linkage between two nodes is an arrowhead line. The line $\vec{ft}_1$ denote the sender f communicate with t_1 actively; while the line $\overleftarrow{ft}_1$ denote the sender t_1 communicate with f actively.

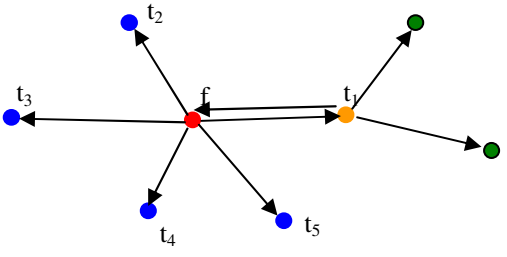


Fig. 3. The map of social network

The sender f is related to five receivers. But which one links to f by closer relationship. The degree of relationship should be evaluated. We computed the weight of link between two nodes as formula 1.

$$W\left(\vec{ft}_1\right) = \frac{N_{\vec{ft}_1}}{\max(N_{\vec{ft}_1}, N_{\vec{ft}_2}, \cdots, N_{\vec{ft}_i})} \tag{1}$$

Where $W(\vec{ft}_1)$ denotes the weight from node f to node t_1. $N_{\vec{ft}_1}$ denotes the number of the sender f sending e-mails to the receiver t_1. The purpose of function *max* is to get the maximum from all the number of the sender f sending e-mails to the receivers. The value of $W(\vec{ft}_1)$ is a decimal fraction at the closed interval between 0 and 1.

4.2 Blog's Social Network Analysis

Blog is one new Internet communication mean in recent years. Blog is a type of website maintained by an individual with regular entries of commentary, descriptions of events, or other material such as graphics or video. Most blogs are interactive, allowing visitors to comment and reprint each other. Members of Blog are gathered as a virtual community. Social relationship can be extracted from blog by analyzing their comment and reprinting each other.

The social network is demonstrated by a directed and weighted graph. Member commenting or reprinting somebody's blog are taken for having relation to the one. The degree of closeness between nodes is showed by weight. The weight is determined by the number of comment and reprinting each other. Here comment and reprinting was evaluated based on different coefficient of weight. The weight was computed by formula 2 and 3.

$$W\left(\vec{ft}_1\right) = \frac{N_{\vec{ft}_1}}{\max(N_{\vec{ft}_1}, N_{\vec{ft}_2}, \cdots, N_{\vec{ft}_i})} \tag{2}$$

$$N_{\vec{ft}_1} = m \times CN_{\vec{ft}_1} + n \times RN_{\vec{ft}_1} \tag{3}$$

Where $W(\vec{ft}_1)$ denotes the weight from node f to node t_1. $N_{\vec{ft}_1}$ denotes the number of user f comments or reprints the blog of user t_1. The purpose of function *max* is to get the maximum from all the number of the user f comments or reprints everybody's blogs that user f relates to. $CN_{\vec{ft}_1}$ denotes the number of user f comments the blogs of user t_1. $RN_{\vec{ft}_1}$ denotes the number of user f reprints the blogs of user t_1. Comments and reprinting are evaluated based on different coefficient of weight. m and n are the coefficient of weight in formula 3.

5 Experiments

Due to involve privacy, e-mail service organization doesn't provide e-mail dataset publicly. E-mail dataset concerning cybercrime was difficult to gain. Furthermore blog dataset with reagard to cybercrime was hard to obtain. So here datasets used for test were experimented to demonstrate our methods of social network analysis. We constrcted a social network based on the message of e-mails including 23 participants and 362 e-mails collected from our research laboratory. Link weighting between nodes was lined out. Figure 4 was the example of e-mail dataset's social network.

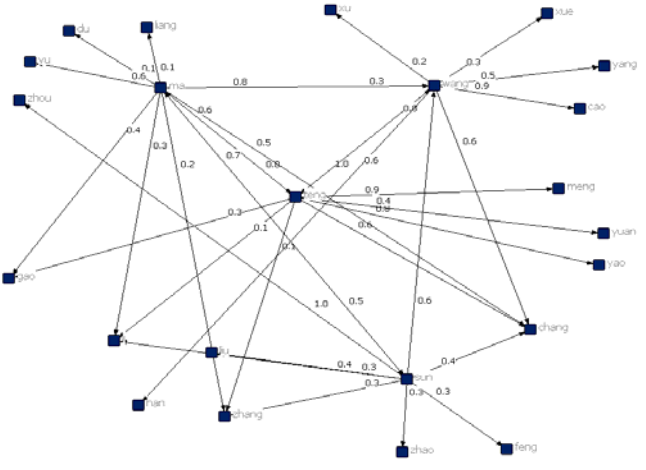


Fig. 4. One example of e-mail dataset's social network

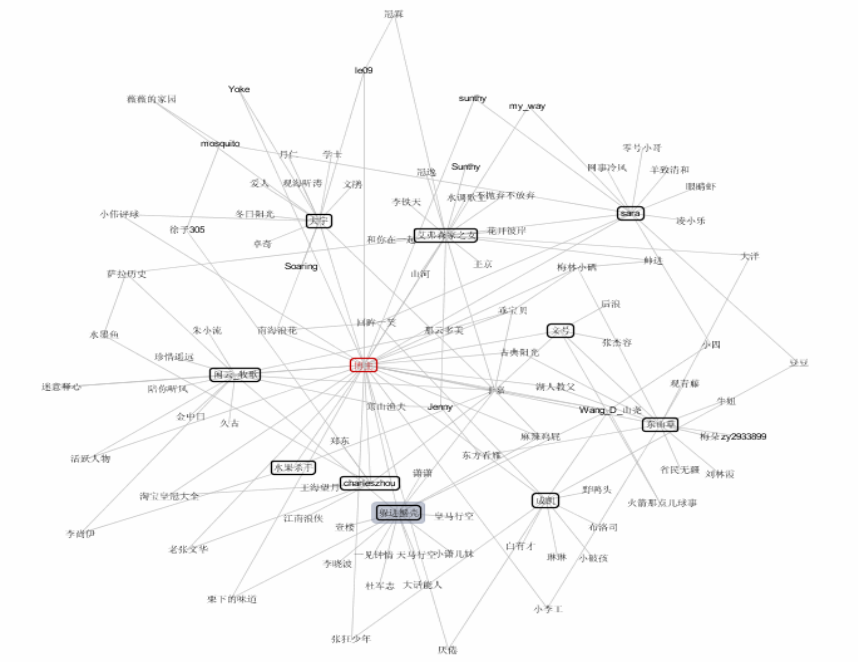


Fig. 5. One example of blog dataset's social network

Blog dataset was collected from *http://blog.sina.com.cn/*. Figure 5 was the example of blog dataset's social network.

6 Conclusions

In this paper, social network analysis method for cybercrime investigation was described. The *from* field and *to* field of e-mail's *header* information were analyzed to obtain the participants' communication pattern. Blog's comment and reprinting were extracted to analyze blog user's social relationship. E-mail's *header* information or blog's registration information were always ignored or forged. Therefore authorship identification methods to judge authorship's authenticity were introduced. Experiments on e-mail and blog dataset were provided to illuminate the method.

References

1. Yang, C.V., Tobun, D.N.: Analyzing Content Development and Visualizing Social Interaction in Web Form. In: Proceedings of 2008 International Conference on Intelligence and Security Informatics,Taipei, pp. 25–30 (2008)
2. Freeman, L.: The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press, Vancouver (2004)
3. Nadel, S.F.: The Theory of Social Structure. Cohen and West, London (1957)
4. Watts, D.: Six Degrees: The Science of a Connected Age. Norton & Company, America (1957)

5. Klovdahl, A.S., Potterat, J.J., Woodhouse, D.E., Muth, J.B., Muth, S.Q., Darrow, W.W.: Social Network and Infections Disease: the Colorado Springs Study. Social Science & Medicine 38, 79–88 (1994)
6. Rosenthal, N., Fingrutd, M., Ethier, M., Karant, R., McDonald, D.: Social Movements and Network Analysis: A Case Study of Nineteenth-Century Women's Reform in New York State. The American Journal of Sociology 90, 1022–1054 (1985)
7. Basu, A.: Social network analysis of terrorist organization in india. In: 2006 Conference of the North American Association for Computational Social and Organizational Science, Notre Dame, USA (2006)
8. Memon, N., Larsen, H., Hicks, D., Harkiolakis, N.: Detecting hidden hierarchy in terrorist networks: some case studies. In: Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO International Workshops on Intelligence and Security Informatics,Taipei, pp. 477–489 (2008)
9. Fu, T., Chen, H.: Analysis of cyberactivism: a case study of online free tibet activities. In: Proceedings of 2008 IEEE International Conference on Intelligence and Security Informatics,Taipei, pp. 1–6 (2008)
10. Frantz, T., Carley, K.: Transforming raw-email data into social-network information. In: Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO International Workshops on Intelligence and Security Informatics, Taipei, pp. 413–420 (2008)
11. Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A.: Mining email social networks. In: MSR 2006: Proceedings of the International Workshop on Mining Software Repositories, Shang hai,China, pp. 137–143 (2006)
12. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for patitioning graphs. Bell System Technical Journal 49(2), 291–307 (1970)
13. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69(6), 066133 (2004)
14. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99(12), 7821–7826 (2001)
15. Abbasi, A., Chen, H.: Visualizing authorship for identification. In: Mehrotra, S., Zeng, D.D., Chen, H., Thuraisingham, B., Wang, F.-Y. (eds.) ISI 2006. LNCS, vol. 3975, pp. 60–71. Springer, Heidelberg (2006)
16. Abbasi, A., Chen, H.: Writeprints: A Stylemetric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. ACM Transactions on Information Systems 26(2) (2008)
17. Zheng, R., Li, J., Huang, Z., Chen, H.: A framework for authorship analysis of online messages: Writing-style features and techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)
18. Zheng, R., Qin, Y., Huang, Z., Chen, H.: Authorship analysis in cybercrime investigation. In: Proceedings of the First International Symposium on Intelligence and Security Informatics, Tucson AZ, USA, pp. 59–73 (2003)
19. Teng, G.F., Lai, M.S., Ma, J.B., Li, Y.: E-mail Authorship Mining Based on SVM for Computer Forensic. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics, Shanghai, China, pp. 1204–1207 (2004)
20. Teng, G.F., Lai, M.S., Ma, J.B.: Selection and Extraction of Chinese E-mail Feature for Authorship Mining. Information 8(3), 437–442 (2005)
21. Ma, J.B., Li, Y., Teng, G.F.: Identifying chinese E-mail documents' authorship for the purpose of computer forensic. In: Proceeding of 2008 IEEE Intelligence and Security Informatics Workshops, Taipei, pp. 251–259 (2008)
22. Ma, J.B., Teng, G.F., Zhang, Y.X., Li, Y.L., Li, Y.L.: A cybercrime forensic method for chinese web information authorship analysis. In: Chen, H., Yang, C.C., Chau, M., Li, S.-H. (eds.) PAISI 2009. LNCS, vol. 5477, pp. 14–24. Springer, Heidelberg (2009)

Topic-Oriented Information Detection and Scoring

Saike He, Xiaolong Zheng, Changli Zhang, and Lei Wang

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Abstract. This paper introduces a new approach for topic-oriented information detection and scoring (TOIDS) based on a hybrid design: integrating characteristic word combination and self learning. Using the characteristic word combination approach, both related and unrelated words are involved to judge a webpage's relevance. To address the domain adaptation problem, our self learning technique utilizes historical information from characteristic word lexicon to facilitate detection. Empirical results indicate that the proposed approach outperforms benchmark systems, achieving higher precision. We also demonstrate that our approach can be easily adapted in different domains.

Keywords: characteristic word combination, Z-Score, self learning.

1 Introduction

TOIDS is a critical task in Intelligence and Security Informatics (ISI)[1,2]. Briefly, there are two problems demanding prompt solution. First, due to free wording and phrasing in web pages, most existing systems pale to differentiate topic related information from the others [3][4]. Secondly, traditional domain-oriented design blocks system's transferring ability, making domain adaption nontrivial [5][6]. In this paper, we propose a hybrid approach that utilizes characteristic word combination to improve domain-specific performance, while addressing the domain adaption problem by self learning.

The rest of the paper is structured as follows. Section 2 reviews related work in topic oriented information detection. In section 3, we present the architectural design and detailed technical information of our TOIDS system. Section 4 reports the results of our evaluation study. Section 5 concludes this paper with a summary.

2 Related Work

TOIDS has been attracting attention increasingly since the frequent occurrence of social security incidents [7]. Topics concerned in this paper include: pornographic violence, criminal offence, terrorism, public health and so on. Previous studies on topic-oriented information detection can be roughly categorized into two groups: heuristic dictionary-based methods and machine learning methods. In the following, we will review concrete research in each category.

2.1 Dictionary-Based Methods

Dictionary-based methods mainly employ a predefined dictionary and some hand generated rules for detecting topic-related webpage. Matching techniques, such as:

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 36–42, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Forward Maximum Matching (FMM), Reverse Directional Maximum Matching (RMM) and Bi-directional Maximum Matching (BMM) are employed to make relevance judgment based on word matching. Primal disadvantages of such systems lie in three aspects. First, the performance greatly depends on the coverage of the lexicon, which unfortunately may never be complete because new terms appear constantly. Secondly, without context taken into consideration, misunderstanding may be incurred in judgment procedure, especially when solid matching used over context sensitive words. Finally, judgment based on single characteristic word is highly unreliable, especially when several common words are used together to express topic related information[8].

2.2 Statistical and Machine Learning Methods

Some researchers cast topic-oriented information detection as a classification task. Li et al. [8] uses kernel based method to filter sensitive information. Greevy and Smeaton [3] classify racist texts with Support Vector Machine. In paper [7], Zhou et al. employ MDS to analyze hyperlink structures, thus uncovering hidden extremist group Web sites. Tsai and Chan [9] use probabilistic latent semantic analysis to detect keywords from cyber security weblogs. However, main drawback of such methods is their paleness in domain adaption, a key problem inherent in most statistical methods.

In this paper, we focus on improving detection precision by using characteristic word combination, akin to previous work in [8]. Apart from related words, we also involve unrelated words to make the final judgment. To facilitate domain adaptation, we integrate self learning technique by deducting based on historical prediction results. The following session will present all the technical details in our TOIDS system.

3 A Hybrid Approach for TOIDS

Figure 1 illustrated our proposed hybrid approach. We first extract characteristic words from training data, and then utilize topic related and unrelated characteristic words to generate word combination features for statistical model. Finally, we construct characteristic word lexicon with self learning based on prediction results. We believe this design could improve performance from two ways: word combination can help filter topic related web pages with higher accuracy. At the same, self learning will improve recall rate by enlarging its lexicon iteratively. This can also be considered as an effective way for domain adaption by utilizing information learned from new domain.

3.1 Z-Score Algorithm

Characteristic words originally refer to those words representative for topic related information. In order to extract such words, we employ Z-Score algorithm, similar to [10]. To meet our objective for TOIDS, we redefined the contingency tale, as given in Table 1.

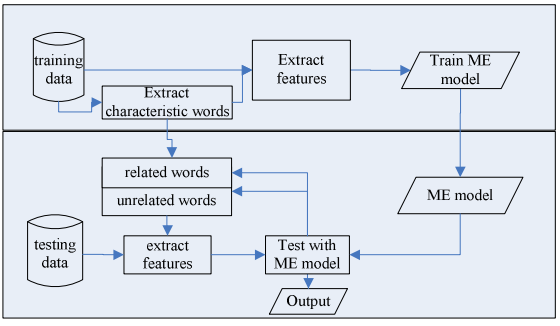


Fig. 1. Flow chart of TOIDS system

Table 1. Contingency table for characteristic word extraction

	Topic related	**Rest**	
ω	a	b	a + b
notω	c	d	c + d
	a + c	b + d	n = a + b + c + d

In table 1, the letter *a* represents the number of occurrences (tokens) of the wordω in topic related documents. The letter *b* denotes the number of tokens of the same wordωin the rest of the whole corpora while $a + b$ is the total number of occurrences in the entire corpora. Similarly, $a + c$ indicates the total number of tokens in topic related documents. The entire corpora corresponds to the union of 'Topic related' and 'Rest' that contains n tokens (n = a + b + c + d).

Here, we define a variable $Zscore(\omega)$ to measure the representative ability of word ω according to Muller's method [10].

$$Zscore(\omega) = \frac{\mathrm{a} - \mathrm{n!\,Pr}\,(\omega)}{\sqrt{\mathrm{n!\,Pr}\,(\omega) \cdot \big(1 - \mathrm{Pr}(\omega)\big)}} \tag{1}$$

where:

$$\mathrm{Pr}\,(\omega) = (\mathrm{a} + \mathrm{b})/\mathrm{n} \tag{2}$$

$$\mathrm{n!} = \mathrm{a} + \mathrm{c} \tag{3}$$

As a rule, we consider words whose Z-Score values are above 0.8 as related while those below -0.8 as unrelated. The thresholds used here are chosen according to our best configuration in experiments. In word combination, we incorporate related and unrelated words into characteristic word set, for we believe both kinds of words are useful. With stop words pruned, related and unrelated words are added into the lexicon with their respective Z-Score value.

3.2 Statistical Model

As each webpage document is composed of sentences, thus, we intend to use the relevance of each sentence to derive a document's relevance, of which the former one comes down to a binary classification task. Here, we adopt Maximum Entropy Model (MEM) [11] to classify the relevance of each sentence. Features used here are mainly extracted based on characteristic words within a sentence, and grouped into four categories: n-gram word feature, n-gram POS feature, word number feature and finally, major POS feature. Table 2 lists these features with a detailed description.

Table 2. Features used in Maximum Entropy model

Feature	Type	Description
n-gram word	Related	n-gram for related words
	Unrelated	n-gram for unrelated words
n-gram POS	Related	n-gram for related POS tags
	Unrelated	n-gram for unrelated POS tags
word number	Related	related word number in current sentence
	Unrelated	unrelated word number in current sentence
major POS	Related	POS tag correspond to highest Z-Score value
	Unrelated	POS tag correspond to lowest Z-Score value

3.3 Topic Oriented Information Detection and Scoring Algorithm

To derive the relevance of document *d*, we introduce a variable *Rel_score(d)*:

$$Rel\_score(d) = \#\text{Rel\_Sentence}/\#\text{Sentence} \tag{4}$$

in which #Sentence indicates the total number of sentence in the current web page while #Rel_Sentence indicates the number of related ones predicated by machine. If *Rel_score(d)* exceeds 0.5, then it is considered as related.

For scoring with more concrete measurement, we also calculate the relevance degree for each web page. This is quantified by calculating the relevance probability in average for all sentences in a webpage, for we believe that prediction probability for each class could depict the relative importance of sentence in fine–grained level. In our TOIDS system, the original relevance degree is scaled into 5 levels[1].

3.4 Self Learning

For the problem of low coverage rate as well as domain adaption, we design a self learning technique by utilizing historical information. Specifically, we augment characteristic word lexicon based on prediction results: words that occur in related sentence yet not found in unrelated word lexicon are added to related word lexicon; words that occur in unrelated sentence yet not found in related word lexicon are added

[1] $Rel\_degree(d)$ values fall into span within [0, 0.5) are cast to 0, those within [0.5, 0.6) are cast to 1, those within [0.6, 0.7) are cast to 2, those within [0.7, 0.8) are cast to 3, and all above 0.8 are cast to 1.

to unrelated word lexicon. Thus, information from historical prediction can be reused in later classification phase.

4 Experiments and Results

We evaluate our approach from three aspects: the effectiveness of characteristic word combination, the influence of self learning and the ability of domain adaptation. Scoring results from actual system are also provided.

Corpora used in our experiments are comprised of about 5000 webpage documents from 10 websites, where 5 are non-governmental websites, the other are portals websites. One point worth mentioning here is that documents are crawled from websites directly rather than retrieval with specified key words, thus, the vocabulary set is open. This aspect guarantees that no priori knowledge of the test corpora is provided, which is highly similar to the circumstance in industrial applications. All collected web pages are manually labeled by professional knowledge engineers. Incompletely statistics show that there are about 20 percent of topic related documents in the whole corpora. The performance is measured by F-score, $F = (\beta + 1)RP/(\beta R + P)$, where R and P are recall and precision rate respectively. For high recall rate preference, we prioritize R by setting β to 2.

Table 3. System comparison under different configurations

Round	Precision	Recall	F-Score
ME + SingleRelWordMatch	62.38	84.36	75.49
ME + RelatedWordCom	79.84	82.75	81.76
ME + CharacteristicWordCom	82.30	82.81	82.64
ME + CharacteristicWordCom + Self-Learning	83.49	83.02	83.18

To test the effectiveness of characteristic word combination and self learning, we evaluate the performance of MEM under different configurations. In our experiment setting, 4-fold cross validations are conducted. Final results are reported in Table 3.

In table 3, 'ME + SingleRelWordMatch' serves as a base line system, which implements same combination strategy in [8]. In this experiment group, features for ME model are extracted only based on uni-gram related words. In comparison, round 'ME + RelatedWordCom' indicates that word combination could bring performance improvement. When unrelated words added to group 'ME + CharacteristicWordCom', performance was further boosted, from 81.76 to 82.64. This is primarily attributed to the improved detection precision, where unrelated words may play a key role. The effectiveness of self learning is justified from the last row in Table 3. For quantitative analysis, we have calculated ROOV value for our data set, which is defined as the ratio of word number in the training data compared to that in the testing data. The average value is 57.48, indicating a relatively low coverage rate. After classification procedure finished, statistics show that entries in our characteristic word lexicon increase averagely by 30 percent. We guess this portion of information remedies low coverage rate and ultimately attributed to overall performance improvement.

In domain adaptation experiment group, we subtract one sub-collection related to transportation (500 documents) to test another one concerning criminal incidents (400 documents). In this experiment, whenever one hundred new documents were classified, F-score is recalculated over all testing documents processed till the current time. Experiment results are given in Figure 2.

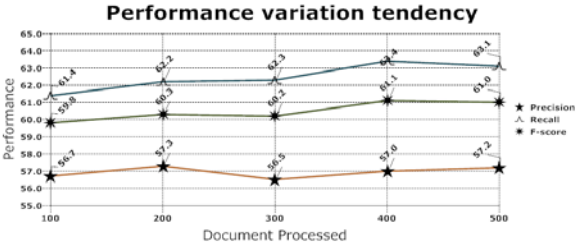


Fig. 2. Performance variation tendency

Table 4. Scoring results from TOIDS system

Title	Rel_score
一颗子弹 马英九连战矛盾面临 … One Bullet The contradiction between Ma Ying-jeou and Lien Chan faces …	5
男子为讨 68.6 万贷款持刀 … For debt collection of 686,000, men armed with knives …	4
河南交通厅长董永安落马… Transport Minister in Henan province Dong Yongan collapses …	5
被囚俄罗斯寡头能把 2 亿… Jailed Russian oligarch uses 200,000,000 …	3

Inspection into performance variation tendency in Figure 2 reveals that: historical prediction results have guided current classification phase. This consists with our original objective in design.

Apart from topic related information detection, results of our scoring strategy from a real TOIDS system are provided in Table 4, which only presents part of the filtered web pages with relevance degree level higher than 3.

5 Conclusions

In this paper, we propose a new method for TOIDS based on characteristic word combination and self learning. By integrating unrelated words into word combination, precision rate is improved while keeping recall rate at a high level. We solve domain adaptation problem by accumulate characteristic words from historical information.

Our future work includes the following:

- Use quantified Z-score of characteristic words to judge a sentence's relevance.
- Distinguish the importance of sentences occurring at different positions.

- Implement mutual enhancement mechanism between sentence and document for their relevance are prohibitively interdependent.
- Enable mistakenly extracted characteristic words eliminated automatically.

Acknowledgments

This work is supported in part by NNSFC #91024030, #60921061, #90924302, #71025001, #60875049 and #70890084, MOST #2009ZX10004-315 and #2008ZX10005-013, and CAS #2F070C01.

References

1. Wang, F.-y.: Social Computing: Fundamentals and Applications. In: Proceedings of IEEE International Conference on Intelligence and Security Informatics, ISI 2008, pp. 17–20 (2008)
2. Zeng, D., Wang, F.-y., Carley, K.M.: Guest Editor's Introduction: Social Computing. IEEE Intelligent Systems 22(5), 20–22 (2007)
3. Greevy, E., Smeaton, A.F.: Classifying racist texts using a support vector machine. In: Proceedings of the 27th Annual International ACM SIGIR Conference, New York, NY, USA, pp. 468–469 (2004)
4. Basilico, J., Hofmann, T.: Unifying collaborative and content-based filtering. In: Proceedings of International Conference of Machine Learning, New York, NY, USA (2004)
5. He, J., Liu, H.-y., Gong, Y.-z.: Techniques of improving filtering profiles in content filtering. Journal on Communications 25(3), 112–118 (2004)
6. Ma, L., Chen, Q.-x., Cai, L.-h.: An improved model for adaptive text information filtering. Journal of Computer Research and Development 42(1), 79–84 (2005)
7. Zhou, Y.-l., Reid, E., Qin, J.-l., Chen, H., Lai, G.: US domestic extremist groups on the Web: link and content analysis. IEEE Intelligent Systems 20(5), 41–51 (2005)
8. Li, W.-b., Sun, L., Nuo, M.-h., Wu, J.: Sensitive information filtering based on kernel method. Journal on Communications 29(4) (2008)
9. Tsai, F.S., Chan, K.L.: Detecting Cyber Security Threats in Weblogs Using Probabilistic Models, vol. 4430, pp. 46–57. Springer, Heidelberg (2007)
10. Zubaryeva, O., Savoy, J.: Opinion and Polarity Detection within Far-East Languages in NTCIR-7. In: Proceedings of NTCIR-7 Workshop Meeting, Tokyo, Japan, pp. 318–323 (2008)
11. Berger, A., Pietra, S.D., Pietra, V.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Language 22(1) (2001)

Agent-Based Modeling of Netizen Groups in Chinese Internet Events

Zhangwen Tan, Xiaochen Li, and Wenji Mao

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{zhangwen.tan,xiaochen.li,wenji.mao}@ia.ac.cn

Abstract. Internet events are public events with the participation of netizens to express their opinions or comments. As an emerging phenomenon, Internet events often draw nationwide attention and eventually influence offline events. Netizen groups who participate in the Internet events play a central role in such events. In this paper, we focus on the study of netizen groups and propose an agent-based model to capture their dynamics and evolvement in Internet events. Our experiment is based on two case studies of Chinese Internet events. We test the proposed model by running simulations and comparing experimental results with real social media data to show the effectiveness of our model.

Keywords: agent-based modeling, netizen group, Internet event.

1 Introduction

Internet events refer to public events which draw the attention and participation of large numbers of netizens. In an Internet event, unorganized netizens autonomously express their sentiments and opinions online which often influence the offline behavior associated with the event. Netizen groups who participate in the Internet events play a central role in such events. In recent years, with the development of Internet and mobile technologies, Internet events happen frequently and have become an emerging social phenomenon that significantly impact and reshape public and social lives.

Internet events are often involved with the participation and interaction of netizen groups——the biggest difference between Internet events and normal events. In order to understand this new phenomenon, it is important to model and capture the mechanism of netizen-centered Internet events from a computational perspective. This is a challenging issue, though. Some research attempts to investigate the issue through statistical analysis and data mining, but their focus is on the specific aspects such as netizen groups' past topics, interests or opinions. Due to the complexity of the interactions in an Internet event and the involvement of multiple parties, it is difficult to model netizen groups and capture the mechanism of Internet events using traditional methods such as statistical analysis or mathematical methods.

To tackle this issue, we ground our work on agent-based modeling of netizen groups and Internet events. Agent-based modeling is a tractable method for capturing netizen group's opinion and actions [1-3]. It is capable of modeling the emergence of social phenomenon through the decomposition and specification of individually

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 43–53, 2011.
© Springer-Verlag Berlin Heidelberg 2011

involved parties (i.e., agents) and their interactions. Recent studies show that the causes of Chinese Internet events can be located in public concerns, sentiments, and demands [4]. Thus netizen groups' sentiments and opinions, which largely reflect public concerns, are the key factors in promoting the evolution of Internet events. Therefore, one particular focus of our agent-based model is on the sentiments and opinions of the netizen groups in Internet events. In addition, to capture the interactions among multiple parties, our model considers all the parties involved in Internet events. We evaluate our model using two case studies of Chinese Internet events [5, 6]. The preliminary modeling results verify the effectiveness of our proposed model.

The rest of this paper is structured as follows. Section 2 briefly reviews the related work. Section 3 presents our agent-based model in detail. Then in Section 4, we test our model based on social media data of two Chinese Internet events occurred in 2010, Synutra event [5] and "360 versus QQ" event [6]. Section 5 concludes this paper.

2 Related Work

Netizen groups have been studied in crisis management domain. Hughes *et al.* [7] outlined several types of online social convergence behavior during times of crisis: *help*, *be anxious*, *return*, *support*, *mourn*, *exploit*, and *be curious*. Sutton *et al.* [8] studied information sharing and dissemination practice by the public during the October 2007 Southern California Wildfires. The authors suggested that community information resources and other unofficial communicative activities enabled by social media are gaining significant momentum in practice, despite concern by officials about the legitimacy of information shared through such means. They argued that these emergent uses of social media are precursors of broader future changes to the institutional and organizational arrangements of disaster response. Vieweg *et al.* [9] analyzed a selected set of online interactions that occurred in the aftermath of the 2007 shooting rampage at Virginia Tech, which represented a new and highly distributed form of participation by the public. These findings suggest strong self-organization which includes the development and evolution of roles and norms wildly exists among netizen groups.

Netizen groups are a special type of online community. Research on general online communities can offer insights and research methods that could benefit investigating netizen-centered Internet events. Computational studies of online communities mainly fall into two categories: communities mining [10, 11] and community characteristics computing [12-15]. Community mining aims to identify online communities that are implicit, inconspicuous, or even hidden. A range of computing techniques such as Web crawling, social network analysis, content analysis, and link analysis, have been applied in community mining research. Community characteristics computing involves analyzing community structure [14], identifying leaders and experts from communities [15], searching for information, and finding friends with similar hobbies, among others.

Previous related work mainly focused on mining or modeling specific aspects of netizen groups based on online social media information. We aim at modeling the

dynamics and evolvement of netizen-centered Internet events. In next section, we shall propose an agent-based model of netizen group and other parties involved in Internet event.

3 Proposed Model

The growth of social media websites in China enables more and more netizens to engage in online discussions about hot topics and thus gives rise to the increasing occurrence of Internet events. Through investigating a number of Chinese Internet events occurred in recent years, we find there are five main parties involved in an Internet event, i.e., *main party*, *opposite party*, *netizen group*, *media*, and *government* (See Figure 1). Main party refers to people or group who initiates a hot event and the opposite party is the one having conflict interest against the main party. Both main party and opposite party are directly involved in the event. They are main parts of the event and their actions significantly influence the evolution of Internet events. Netizen group is the collection of netizens who are associated with the Internet event via online participation. Media includes traditional media (e.g., newspaper and TV) and online media (e.g., news websites and social network sites). Media reports latest news about Internet events. It is the medium of information diffusion. Government plays intermediate roles in some events. The response and policies made by government always impact netizens' opinion toward the event and government.

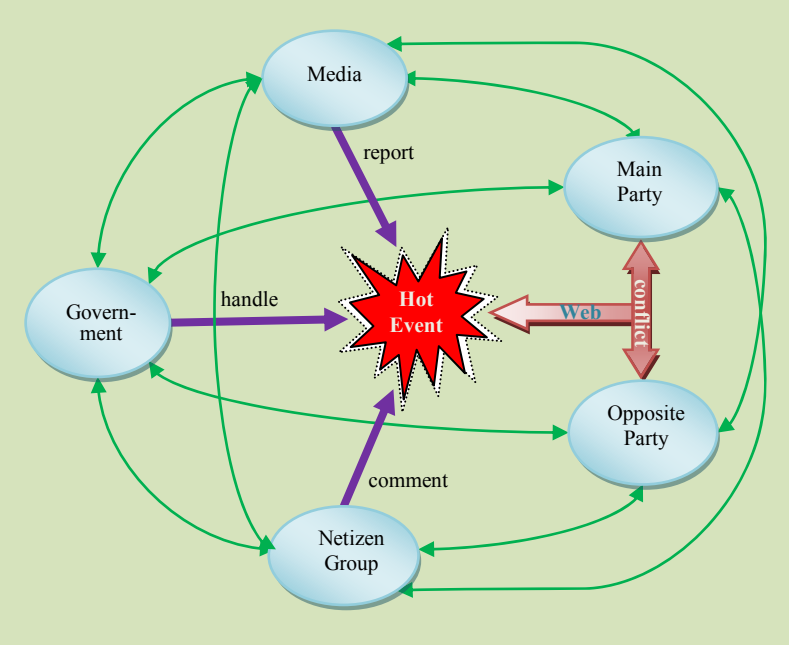


Fig. 1. Agent-based model for netizen-centered Internet events

Netizen groups play a key role in Internet events. In our model, we define the states and actions of the five parties (i.e., agents) as well as their interaction rules. Below we shall introduce the agent design.

States. States are inherent attributes of agents which are closely related to their actions. All the five parties have belief states denoting their belief about an Internet event. For netizen group, media and government, they are not directly involved in an Internet event. They have *concern* as a belief state. Concern denotes the attention of a party to the event. It is a numerical variable ranging from 0 to 1. The main party and the opposite party are directly involved in an Internet event and therefore have *benefit* as a belief state instead of concern. Benefit denotes the benefit the main party/opposite party gets from the event. Benefit is a nominal variable with value *positive*, *negative* or *null*. In addition, we also use *opinion* to capture the positive/negative attitude one agent holds about another agent. For example, the *netizen group* has *opinion toward the main party, the opposite party*, *media*, and *government*, respectively. Each opinion is a numerical variable which ranges from -1 to 1, and positive/negative value represents positive/negative opinion polarity.

Actions. While the states of the five parties are a bit similar, their actions are rather diverse. Among all the possible actions of one party, we are particularly interested in those actions which have impact on the attitude/opinion of one agent toward others because our model aims to capture the dynamics aspect of a netizen-centered Internet event. An action has corresponding *target* and *intensity*. Target is the object of the action and can be one of the five agents defined. If an action has no target, the event itself is the default target. Intensity degree could be low, medium or high.

Actions of the main party and opposite party are the same. They consist of a number of actions, which are classified into positive actions and negative actions. Positive actions include those actions like *donation*, *compensation*, *surrender*, *praise*, etc. Negative actions include actions such as *prosecution*, *accusation*, *criticism*, *blackmail*. Actions of netizen group are also divided into positive actions and negative actions. Positive actions can be online or offline actions such as *help*, *support* or *digging* target's positive news. Negative actions are actions like *threat*, *accusation*, and *digging negative news*. Each action has its target and intensity degree. Media's actions include *praise*, *criticism* and *report-in-view*. Praise/criticism means media writes columns or comment articles to praise/criticize target. Praise and criticism are media's subjective opinions on the event. Report-in-view denotes media reports positive or negative news about target. Report-in-view is media's objective opinions on the event.

Government has five actions: *get-involved*, *judge-winner*, *appeasement*, *award*, *punishment* and *no-response*. Each action has intensity and target. Get-involved indicates government starts to get involved in the event. The intensity can be low, medium or high. The intensity is determined by the type of government, for example, the intensity is low if local government is involved. Judge-winner means government judges the target as winner after investigation, and the target is null when no party wins. Appeasement represents that government gives compensation to the target. Appeasement and award are positive actions, and no-response and punishment are negative actions.

Interaction rules. Interaction rules specifies how one party's states will change in response to another party's actions. For the main party, for example, the opinion/attitude toward the opposite party would decrease if the benefit of the main party is negative or the opposite party takes negative actions. It is also affected by government's judgement. The opinion toward the media is affected by the media's action. If the media reports positive news of the main party or negative news of the opposite party, the main party's opinion toward the media will increase. The opinion toward the government is affected by the government's action. If government takes negative actions against the main party, the main party's opinion toward the government will decrease.

With the increase of the conflicts between main party and opposite party, the concerns of media and netizen group will increase. Actions from other parties will increase concerns as well, such as criticism from netizen group to government. Negative actions often have greater influence than positive actions. This is probably due to the negativity bias according to social psychology [16]. If no action occurs, the concern will decrease with time following the power law, as found in social computing research.

Opinion of a netizen group can be influenced by other parties' states, for example, opinion toward the main party/opposite party is affected by the party's benefit. Actions from other parties can also change the opinion of the netizen group. For example, the opinion of the netizen group toward the main party will increase if the government judges the main party as winner. The netizen's opinion toward the government would decrease if government does not respond to the event in certain time limit or the investigation results/the actions of the government are contrary to the netizen's expectation. During the event, the netizen group's opinion will decay over time [17, 18]. Rules for the opinion of the media to the other parties are similar to those of the netizen group.

Below we choose two typical Internet events as case studies to test our proposed model.

4 Case Studies

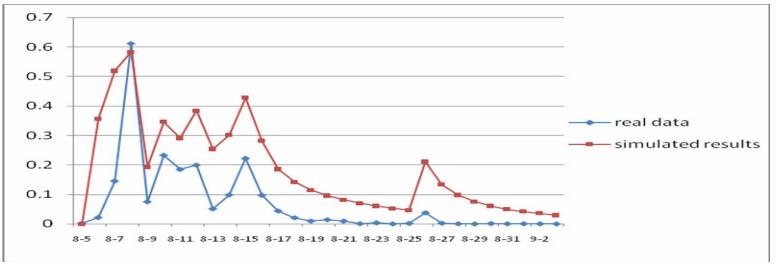
To illustrate the usefulness of our model, we test our model by simulating two typical Chinese Internet events occurred in 2010, i.e., *Synutra event* and "*360 vs. QQ*" event. We select Repast [19] as the agent-based modeling tool. In each scenario, we first initialize the model by setting event type and initial states of all the parties according to the related online news. The actions are directly transferred from real actions of all the parties in the event. The default action of every agent is no action and the default value of benefit is null. These parameters are extracted from text data by hand, and each time step corresponds to a day in the real world event. After running the model step by step, we get the netizen group's concern and opinions as the output of the simulation. To test the model's performance, the simulation results are then compared with results mined from real-time social media data using our sentiment analysis tool [20].

4.1 Synutra Event

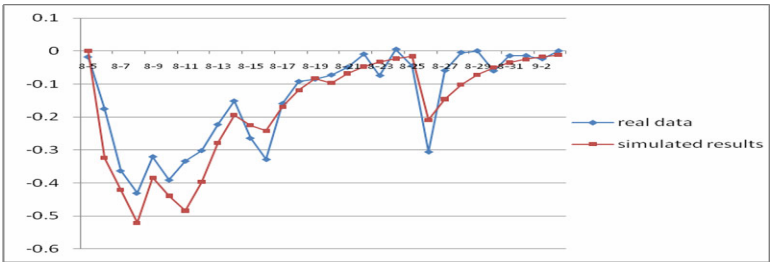
Synutra event is a food safety event happened in August 2010, in which Synutra baby milk powder was suspected to cause precocious puberty. Synutra event had drawn national attention and governmental investigation. At the beginning, three girls in Wuhan who had been fed with Synutra milk were suspected to be premature. Synutra denied the report later. Customers and large numbers of netizens do not believe the announcement given by Synutra and joined together to denounce Synutra. Finally Chinese Health Ministry was forced to initiate investigation. The investigation results showed that three girls were pseudo-precocious puberty. In this event, netizen group played an important role in the evolvement of the event.

We extract news reports and comments about Synutra event from three popular Chinese news sites, including QQ, 163, and Phoenix. Here we get 153, 104, and 580 reports, and 58307, 12146, and 13419 comments from QQ, 163, and Phoenix respectively.

In Synutra event, customers act as the main party, Synutra company is the opposite party and Health Ministry is the government. In the scenario we model the event from Aug 6 to Sep 3. Fig. 2 shows the comparison of simulation results and real data. In the figure x axis represents the date, and y axis denotes the positive/negative polarity of netizen group's opinion or the netizen group's volume of comments.

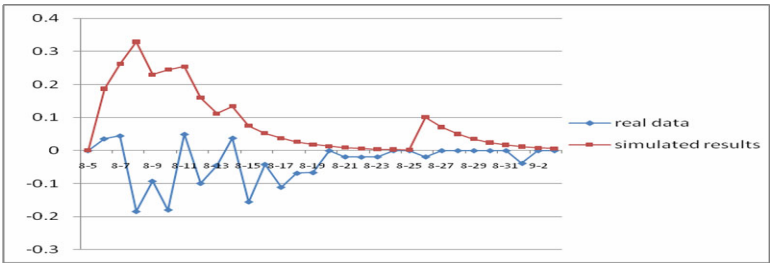


(a) Netizen group's concern

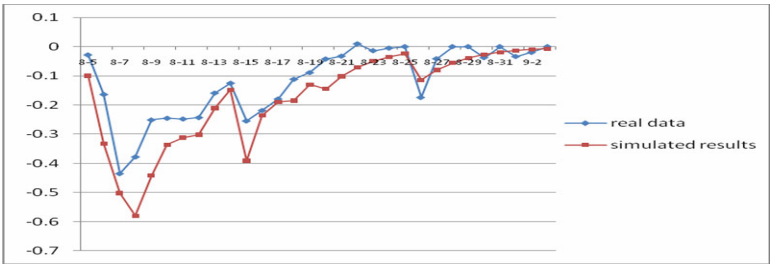


(b) Netizen group's opinion toward Synutra company

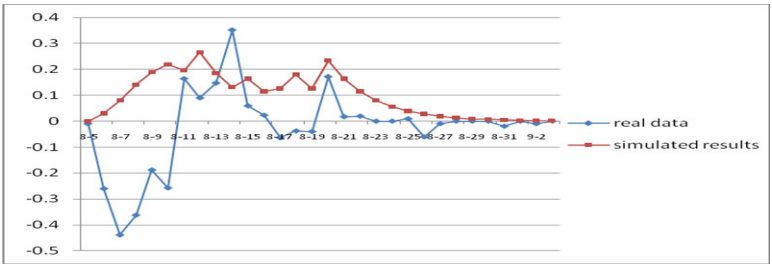
Fig. 2. Comparison between simulation results and real data



(c) Netizen group's opinion toward the customers



(d) Netizen group's opinion toward the government



(e) Netizen group's opinion toward media

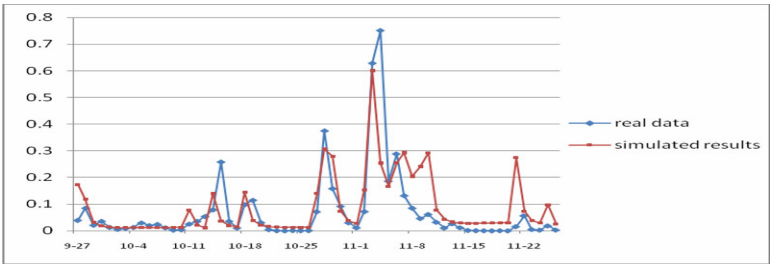
Fig. 2. (*continued*)

The curve of the simulated netizen group's concern matches well with the curve of the total amount of the comments, as shown in Fig. 2(a). From Fig. 2(b) we can see that the netizen group's negative opinion toward Synutra is rather strong on Aug 8 because a lot of victims of the Synutra milk powder are found. Fig. 2(d) shows the simulated opinion toward the government is similar to the real data. The opinion toward the government decreases at the beginning as the government refused to react to the event and thus raised the public rage. After the join of the Health Ministry, the opinion increases gradually. However, it drops suddenly on Aug 15 because the investigation results were not consistent with netizens' anticipation and they doubted the government's fairness. In general, the simulation results by our model fit the real social media data quite well. This basically verifies the effectiveness of the model.

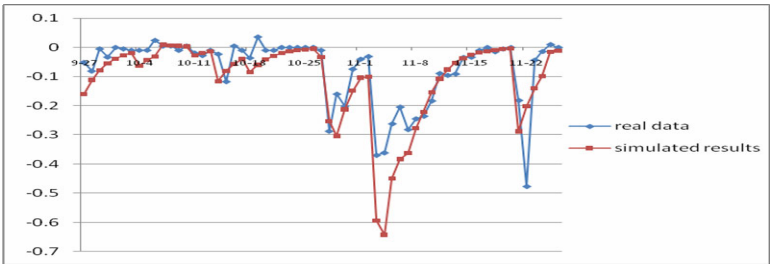
4.2 360 VS QQ

360 versus QQ is a business conflict between Qihoo and Tecent, two leading software companies in China. 360 published privacy protection software to monitor QQ software on September 27. Tecent announced QQ software would not compatible with 360 on personal computers. Netizens got rage soon and expressed their anger by posting threads online. The event ended after Industry and Information Technology Ministry and Public Security Ministry's engagement.

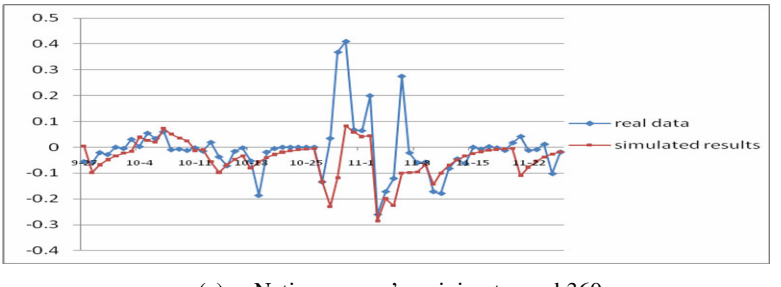
We downloaded 352 news reports and 169741 comments from 163. The y axis in Fig. 3(a) denotes the number of daily comments. The curves in the other figures show the netizens' opinions toward QQ, 360, the government and the media. We got these data according to the same procedure described in Synutra event case.



(a) Netizen group's concern

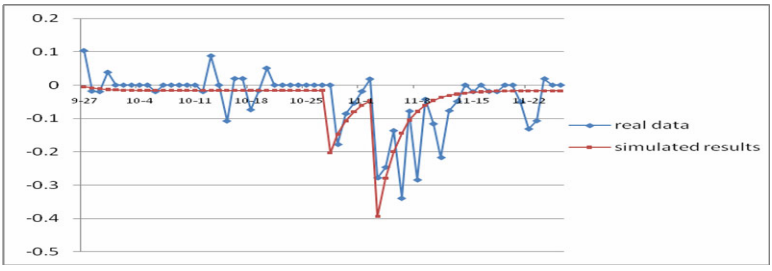


(b) Netizen group's opinion toward QQ

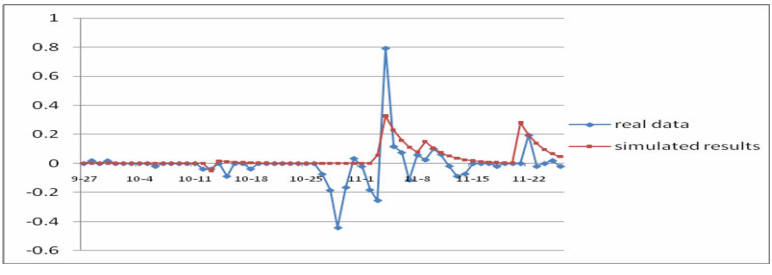


(c) Netizen group's opinion toward 360

Fig. 3. Comparison between simulation results and real data



(d) Netizen group's opinion toward the government



(e) Netizen group's opinion toward media

Fig. 3. (*continued*)

In this scenario, 360 is the main party, QQ is the opposite party, and Ministry of Industry and Information Technology and Ministry of Public Security are the government. Fig. 3 compares the simulated netizen group's concern and opinions with the real media data from Sep 27 to Nov 26. X axis and y axis denote the same meaning as those in Fig. 2.

Fig. 3(a) shows that the curve of the simulated concern matches well with the real data. In Fig. 3(b), it reaches its lowest point on Nov 4 because QQ exerted pressure on netizens and got criticism from media and netizens on that day. We can see from Fig. 3(c) that the simulated opinion toward 360 matches the real data well. In Fig. 3(d), the decreases on Oct 28 and Nov 3 are due to the direct conflicts between the two companies. In Fig. 3(e), opinion toward media in real data decreases on Oct 28 and Nov 3 for the same reason. The simulation results by our model generally fit the real data, which shows the model's effectiveness.

5 Conclusion

This paper proposes an agent-based model of netizen groups in Internet events. The proposed model considers all the main parties involved in an Internet event and captures how their interactions impact their opinions toward others. We focus on studying the evolution of netizen group's concern and opinion in the event and netizen group's interactions with the other parities. We select two typical Chinese Internet events as case studies and compare the experimental results with real social media data to verify the effectiveness of our model. This work contributes to the study of

social behavior in social computing [21]. Our future work shall explore the use of information extraction techniques to automatically extract media data and facilitate model construction.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grant Nos. 60875028, 60875049, 60921061 and 90924302.

References

1. Grimm, V., Revilla, E., Berger, U., et al.: Pattern-Oriented Modeling of Agent-based Complex Systems: Lessons from Ecology. Science 310(5750), 987–991 (2005)
2. Kaiser, J.: BIODEFENSE: Senate Bill Would Alter Biosafety, Select Agent Rules. Science 320(5883), 1573–1573 (2008)
3. Zacharias, G., MacMillan, J., Van Hemel, S.B.: Behavioral Modeling and Simulation: From Individuals to Societies. The National Academies Press, Washington, DC (2008)
4. Jiang, M.: Chinese Internet Events. In: The Internet in China: Online Business, Information, Distribution and Social Connectivity. Berkshire Publishing, New York (2010)
5. Synutra event (2010), http://www.bbc.co.uk/news/world-asia-pacific-10978702
6. 360 v. QQ (2010), http://en.wikipedia.org/wiki/360_v._Tencent
7. Hughes, A.L., Palen, L., Sutton, J., et al.: Site-seeing in Disaster: An Examination of Online Social Convergence. In: Proceedings of the 5th International ISCRAM Conference, Washington, USA (2008)
8. Sutton, J., Palen, L., Shklovski, I.: Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires. In: Proceedings of the 5th International ISCRAM Conference, Washington, USA, pp. 624–632 (2008)
9. Vieweg, S., Palen, L., Liu, S.B., et al.: Collective Intelligence in Disaster: Examination of the Phenomenon in the Aftermath of the 2007 Virginia Tech Shooting. In: Proceedings of the 5th International ISCRAM Conference, Washington, USA, pp. 44–54 (2008)
10. Blood, R.: How Blogging Software Reshapes the Online Community. Communication of the ACM 47(12), 53–55 (2004)
11. Zhang, H., Giles, C.L., Foley, H.C., et al.: Probabilistic Community Discovery using Hierarchical Latent Gaussian Mixture Model. In: Proceedings of the 22nd National Conference on Artificial Intelligence, Vancouver, British Columbia, Canada (2007)
12. Golbeck, J., Parsia, B., Hendler, J.: Trust Networks on the Semantic Web. In: Proceedings of Cooperative Intelligent Agents, Helsinki, Finland (2003)
13. Lim, E.P., Vuong, B.Q., Lauw, H.W., et al.: Measuring Qualities of Articles Contributed by Online Communities. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, pp. 81–87 (2006)
14. Wang, F.-Y., Zeng, D., Hendler, J.A., et al.: A Study of the Human Flesh Search Engine: Crowd-Powered Expansion of Online Knowledge. Computer 43(8), 45–53 (2010)
15. Nolker, R.D., Zhou, L.: Social Computing and Weighting to Identify Member Roles in Online Communities. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA (2005)

16. Niu, X.: Psychological Effects Analysis of the Dissemination of Internet Hot Events. Chinese Journal of Voice and Screen World 2009(1), 36–38 (2009)
17. Chen, Y., Yu, J.: Mechanisms of Mass Emergency's Occurrence, Response, and Prevention, and Public Opinion. Journal of Social Science Review 21(7), 135–137 (2006)
18. Plutchik, R.: The Nature of Emotions. American Scientist 89(4), 344–350 (2001)
19. Collier, N.: Repast: An Extensible Framework for Agent Simulation. The University of Chicago's Social Science Research (2003)
20. Zhang, C., Zeng, D., Li, J., et al.: Sentiment Analysis of Chinese Documents: From Sentence to Document Level. Journal of the American Society for Information Science and Technology (JASIST) 60(12), 2474–2487 (2009)
21. Wang, F.-Y., Carley, K.M., Zeng, D., et al.: Social computing: from social informatics to Social Intelligence. IEEE Intelligent Systems 2(22), 79–83 (2007)

Estimating Collective Belief in Fixed Odds Betting

Weiyun Chen[1], Xin Li[2], and Daniel Zeng[1]

[1] Institute of Automation, Chinese Academy of Sciences,
Beijing, China
{weiyun.chen,dajun.zeng}@ia.ac.cn
[2] Department of Information Systems
City University of Hong Kong, Hong Kong, China
xin.li@cityu.edu.hk

Abstract. Fixed odds betting is a popular mechanism in sports game betting. In this paper, we aim to decipher actual group belief on contingent future events from the dynamics of fixed odds betting. Different from previous studies, we adopt the prospect theory rather than the expected utility (EU) theory to model bettor behaviors. Thus, we do not need to make assumptions on how much each bettor stake on their preferred events. We develop a model that captures the heterogeneity of bettors with behavior parameters drawn from beta distributions. We evaluate our proposed model on a real-world dataset collected from online betting games for 2008 Olympic Game events. In the empirical study, our model significantly outperforms expert (bookmaker) predictions. Our study shows the possibility of developing a light-weight derivative prediction market upon fixed odds betting for collective information analysis and decision making.

Keywords: fixed odds betting; prediction markets; computational experiments.

1 Introduction

Fixed odds betting is a popular mechanism used in sports betting games. While providing gambling derivates, sports betting game is essentially a prediction market (PM) where participants' estimations on contingent future events can be combined.

A classic PM serves to trade virtual objects promising certain payoffs according to future events [1]. For the benefit of monetary awards, market participants would collect and analyze information to maximize their profit through trading. The pricing mechanism of prediction market can thus combine participants' belief and provide accurate prediction to the future events [2]. PM has been widely used in predicting US president election results, Iraq war consequences, movie box offices, etc.

Full-fledged PMs generally require significant mental efforts of participants, which may limit their applications. Fixed odds betting provides a simpler mechanism where participants can focus on judging future events along with odds set by bookmakers. Although its trading (i.e., betting) nature still supports aggregating group beliefs [3], the price of virtual object, posted as odds, is fixed at the beginning of the game and can no longer work as an indicator of participants' beliefs on future events as in classic prediction markets. The belief of bettors is biased by provided monetary incen-

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 54–63, 2011.
© Springer-Verlag Berlin Heidelberg 2011

tive odds and may not neutrally reflect future event probability either. Thus, to investigate group beliefs in fixed odds betting holds the potential to enable a lightweight PM mechanism for wider applications.

This paper aims to estimate market participants' group belief on future events in a fixed odds betting game. In light of the prospect theory, we model the impact of odds on individuals' betting choice to find out a more reliable measure on group belief. Our model can be extended to generic PMs. We evaluate the model upon a real-world dataset collected from online betting games for 2008 Olympic Game events.

2 Related Works

Previous literature on fixed odds sport betting mainly focus on two perspectives, to investigate market efficiency[4] and to examine how bettors act under uncertainty. Since the odds are set by bookmakers, most existing research on market efficiency takes a bookmaker perspective[5]. In general, it is argued that bookmakers need to make and have been making accurate assessments on future events in fixed-odds betting[6]. Nevertheless, some inefficiency such as the favorite-longshots bias is detected[7], which indicates that favorite events are undervalued and long shots, i.e. outcomes which are very unlikely, are overvalued. This promotes the second stream of study. Some researchers argue that this is due to the existence of risk loving bettors in such a game setting[8]. Alternatively, behavioral theories[9] suggest that cognitive errors and misperceptions of probability play a role in the market distortion, which is supported by recent empirical studies[7].

In terms of understanding group belief from the fixed odds betting, there have been limited studies previously. In classic PM literature, it is argued that the price of trading objects can work as an indicator of the group belief on future events. Wolfers and Justin's work[10] justified this argument through a model that considers betters as rational agents to maximize their profit. However, as we mentioned, irrational betters is a significant phenomenon in fixed odds betting. Thus, behavioral theories, such as Kahneman's prospect theory [11], may be more descriptive than rational agent theory to understand better behavior. For example, Bruno Jullien [12] investigates the bettors' attitude toward risk in British horse races, where the cumulative prospect theory shows higher explanatory power on bettors behavior under risk. However, such theories has not been utilized to estimate group belief in fixed odds betting.

3 Methodology

3.1 Game Setting

For the sake of simplicity, we consider binary game betting in this research, i.e., the game has only two possible results, either A or B happens. We assume the objective probability for events A and B are p_A and p_B, where $p_A+p_B=1$. In the beginning, the bookmaker sets the odds for A happen as o_A and for B happen as o_B, where $o_A, o_B \in (0,+\infty)$. It is necessary for $1/(1+o_A)+1/(1+o_B) = d > 1$, so that bettors can't gain through betting on both sides simultaneously. (In practice, d is often set to 1.11 [5].) In this research, we assume there large static population of bettors and each bettor

only bets once in each game. The shares of votes on A and B are denoted as b_A and b_B, respectively, with $b_A+b_B=1$.

As a naive measure, the odds represent bookmaker's estimation on the contingent event and shares of votes (b_A and b_B) represent bettor belief (under the influence of the odds). We denote $(o_B+1)/(o_A+o_B+2)$ as p_{odds} and b_A as $p_{bettors}$ to characterize the two parties' estimation on event A through our direct observation. If $o_A > o_B$, bookmaker prefers B than A. If $b_A > b_B$ participants prefer A than B under the influences of betting odds (i.e., A may bring them more profit).

3.2 Model Design

We assume each better make betting choice based on their perceived utility of the two events. Their decision model is:

$$\begin{cases} \text{if} \quad U(p_A, o_A) > U(p_B, o_B) & \text{then choose A} \\ \text{if} \quad U(p_A, o_A) < U(p_B, o_B) & \text{then choose B} \end{cases} \tag{1}$$

where $U(p_i, o_i)$ is the generalized expected utility for choosing event i, given odds o_i and objective probability $p_i i \in$ {A,B}.According to the prospect theory, human accesses the objective probability of events with a bias. Such bias can be represented by a probability weighting function in utility function. The generalized expected utility is:

$$U(p_i, o_i) = w^+(p_i)v(o_i) + w^-(1-p_i)v(-1) \tag{2}$$

where, $w^+()$ and $w^-()$ are probability weighting functions on objective probabilities for event i to happen (p_i) or not ($1-p_i$). $v()$ is the value function for certain gain or loss. In our case, the gain is o_i and the loss is -1 for unit betting.

In prospect theory, a commonly used probability weighting function is

$$w^+(p) = p^\gamma \Big/ \left[p^\gamma + (1-p)^\gamma \right]^{1/\gamma}, \quad w^-(p) = p^\tau \Big/ \left[p^\tau + (1-p)^\tau \right]^{1/\tau} \tag{3}$$

which models the fact that bettors often overestimate the small probability but underestimate the large probability. In practice, the sum of the two functions are close to 1, which is assumed so in this research. Since for each betting game, p_i is an constant number, we denote $w^+(p_i)$ as w_i and simplify above model to:

$$U(p_i, o_i) = w_i v(o_i) + (1-w_i)v(-1) \tag{4}$$

To model the heterogeneity in betters' bias, we assume the distribution of bettors on γ cause w_i follow a Beta distribution $f(w_i)$, with different shape parameters in different event predictions. Since any Un-regular distributions in [0,1] can be approximated as the linear combination of a set of Beta distributions, this assumption does not affect our model's generalizability. Thus, we have.

$$w_i \in \text{Beta}(\eta_i, \upsilon_i); \quad E(w_i) = \frac{\eta_i}{\eta_i + \upsilon_i}; \quad D(w_i) = \frac{\eta_i \upsilon_i}{(\eta_i + \upsilon_i)^2 (\eta_i + \upsilon_i + 1)} \tag{5}$$

The prospect theory argues that bettors often are risk averse in gain and risk love when against lost. Their value function can take a form:

$$v(x) = \left(x^{\alpha} \quad \text{when } x \geq 0; \quad -\lambda(-x)^{\beta} \quad \text{when } x < 0 \right) \tag{6}$$

Given this value function, the generalized expected utility is:

$$U(p_i, o_i) = w_i o_i^{\alpha} - (1 - w_i)\lambda \tag{7}$$

That leaves us two parameters λ and α to model the heterogeneity in betters' value functions. We assume that λ is fixed and α follows a Beta distribution h(α). For a given group of bettors, we assume the value of λ and the distribution of α is non-variant in different bets of sport games since they are features of human nature.

Given these simplifications, for a game with (unknown) p_i we can characterize the observed share of bettors on event i, i.e., $E(b_i)$, according to odds o_i, with the help of, 1) the distribution of w_i on bettor's belief bias; and 2) λ and the distribution of α on the congenital and heterogeneous decision nature of bettors under risk. More specifically, we expect to find out the relationship among these variables.

We tried to model this problem through 2-dimentional integration. However, there is no closed form solution on such a formulation. Thus, we turn to a computational experiment approach [13-14] for approximation.

3.3 Model Specification

In our model, the objective probability of event i (i.e., p_i) is coded in w_i. Thus, we use the expectation of weighted (biased) group belief $E(w_i)$ as an alternative and investigate how it moderate the relation between $E(b_i)$ and odds o_i. Since our game only has binary results, we can only investigate one event, say, A. In this a case $E(b_i) = p_{bettors}$. For the ease of analysis, we investigate the impact of p_{odds}, rather than o_i, on $E(b_i)$ in our computational experiments.

Table 1. Simulation process to find out the influences of model parameters on the experiment result

Given λ
For α~Beta(·) in a list of shape parameters
For w~Beta(·) in a list of shape parameters
For odds in a list of (o_A, o_B) satisfying $1/(1+o_A)+1/(1+o_B) = d > 1$, we get p_{odds}
Sample from the distribution and produce a large number of bettors
Each produced agent act according their decision parameters and make choice
Statistic the final bet share on each side, thus get $p_{bettors}$
Fit the curve between p_{odds} and $p_{bettors}$ using curve fitting
Evaluate how distribution of w influence the parameter of the formula between p_{odds} and $p_{bettors}$
Evaluate whether distribution of α influence the formula between p_{odds} and $p_{bettors}$

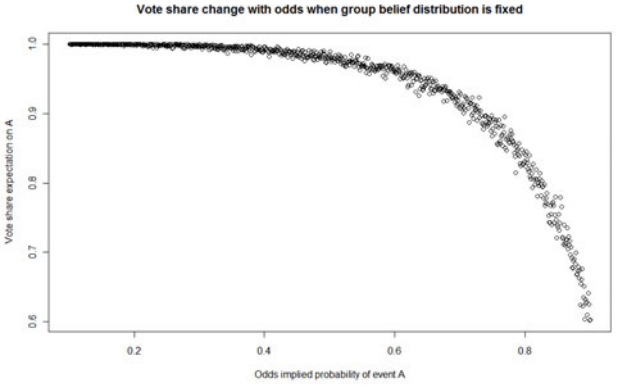


Fig. 1. $p_{bettors}$ change with p_{odds} given other parameters fixed

We conduct our computational experiments in the following steps. First, we generate a large amount of bettors by specifying their characteristics on λ, distribution of α, and distribution of w_i. We alternate parameters that characterize λ, α, w_i to have different types of bettors. Second we generate a group of games with various odds o_i. Thus we can generate these produced bettors' choices based on our discussion in section 3.2 and get the share of betters choosing each side. Based on these simulated data, we estimate the relationship between $p_{bettors}$ and p_{odds}, by considering the parameters specifying w_i and α. We include those parameters in a step by step manner. These steps are described in Table 1.

From the experiments, we notice that if we fix parameters of w_i and α to generate bettors, p_{odds} and $p_{bettors}$ show a polynomial relationship on games with different odds (Figure 1). We thus build the base function of relations between p_{odds} and $p_{bettors}$ as:

$$p_{bettors} = p_{odds}^{a} \tag{8}$$

Furthermore, we found that a change with p_{odds} exponentially in simulated data:

$$a = e^{b+c\cdot p_{odds}}, \text{ i.e., } p_{bettors} = p_{odds}^{e^{b+c\cdot p_{odds}}} \tag{9}$$

where b and c can be estimated from simulated data.

We then try to include bettors' belief bias in the model. We first adapt shape parameters of distribution of w, where the sum of two shape parameters, $(\eta_i + \nu_i)$is a constant. Based on simulated data we find that both b and c have a linear relationship with $E(w)/\sqrt{D(w)}$.

$$b = r + s\frac{E(w)}{\sqrt{D(w)}} \qquad c = l + m\frac{E(w)}{\sqrt{D(w)}} \tag{10}$$

Thus we have a formula on overall how bettors' belief biases affect the relationship between odds setting and final betting share:

$$p_{bettor} = p_{odds}^{\exp\left[\left(r+s\frac{E(w)}{\sqrt{D(w)}}\right)+\left(l+m\cdot\frac{E(w)}{\sqrt{D(w)}}\right)\cdot P_{odds}\right]}, \text{ i.e.,} \tag{11}$$

$$\log\left(\frac{\log p_{bettors}}{\log p_{odds}}\right) = \left(r + s\frac{E(w)}{\sqrt{D(w)}}\right) + \left(l + m\frac{E(w)}{\sqrt{D(w)}}\right) p_{odds} \tag{12}$$

We test above formula with a variety of shape parameters of w_i. It in general fits the data well.

Note that we have assumed the sum of shape parameters of w to be constant. So we further check whether formula (12) holds without this constraint. Specifically, we keep variance of w, $D(w)$, as fixed and change expectation of w, $E(w)$,. The results show that (12) also holds. Thus we can assert that this conclusion holds for arbitrary combination of distribution parameters of w since the degree of freedom for Beta distribution is 2. One can specify λ, distribution of α and $D(w)$ as known empirical value and utilize the simulation process in Table 1 to estimate the coefficients r, s, l and m. Then utilize these estimated coefficients and corresponding $D(w)$ in (12) to solve out $E(w)$.

Based on the formula of (12), we continue to change the distribution parameters of α and check its role in p_{odds} and $p_{bettors}$ relationship. In general, we do not observe any significant relations between the distribution parameters and coefficients r, s, l and m. For a given group of bettors, the coefficients r, s, l and m are mainly determined by their belief bias. In practice, we can estimate them from the entire population.

It should be noted that our ultimate goal was to estimate $E(w_i)$ as an indicator of p_i. If we have no sense of the empirical value of λ, distribution of α and $D(w)$. We can consider the characteristics of real application to further simplify this model. As we know, in different games, bettors' beliefs are likely from different distribution, hence different $D(w)$. However, we can assume $D(w)$ is constant across games for a same set of bettors and get:

$$\log\left(\frac{\log p_{bettors}}{\log p_{odds}}\right) = \left(r + sE(w)\right) + \left(l + mE(w)\right) p_{odds} \tag{13}$$

Where s and m actually absorb the effect of $1/\sqrt{D(w)}$ if we estimate them from data. Finally we can get:

$$E(w) = \frac{\log\left[\log P_{bettors} / \log P_{odds}\right] - r - l \cdot P_{odds}}{s + m \cdot P_{odds}} \sqrt{D(w)} \tag{14}$$

3.4 Model Parameters Estimation

Given a training data set, we can estimate the parameters in (14) and utilize them to forecast the perceived event probability. Assume there are N binary games with known odds and share of bets, we can estimate the parameters of (14) to minimize its difference with real event results, such as on Mean Square Error (MSE). In this research, we focus on the total prediction accuracy and adopt following objective:

$$\max_{r,s,l,m,\sqrt{D(w)}} \sum_{i=1}^{N} sign\left(E^i(w)-0.5\right)*R^i \qquad i=1,...,N \tag{15}$$

where $E^i(w)$ is the estimated group belief on game i, and R^i denotes the result of game i, i.e., 1 for A wins and -1 for B wins in a binary game. We can solve this maximization problem conveniently through numerical methods.

4 Empirical Study

4.1 Data

In August 2008, Sina.com hold a fixed odd betting game to the 2008 Olympic game on its web page. The game attracts 174,609 participants. We collect the votes on 482 betting games among which 167 are binary betting.

4.2 Benchmark Models

In order to verify the effectiveness of our Estimated Group Belief model $E(w)$, we compare its performance with two basic benchmark models:

First, Share of Betters Model: $P_{bettors}$. Final betting share on each side is an directly observable event probability from data, where majority choices indicates a winner. We expect our model overcome the biases introduced by odds incentives and achieve better performance than this one.

Second, Bookmaker Odds Model: P_{odds}. Odds setup by bookmakers indicates experts' judgment on event probability, where a smaller odds indicates a winner. Our model can partially address bettor's belief bias and opinions on risk and join the wisdom of crowds. We expect our model to be at least as effective as this one.

Although these two benchmark model seem naive, they are state-of-the-art methods on fixed odds betting. We compare our Estimated Group Belief model with them on predicting the probability for A to happen (A if larger than 0.5 and B less than 0.5). We compare their ratio of correct predicted games.

4.3 Results

First we estimate the parameter of (15) as explained in section 3.4. The estimated optimal r, s, l, m, $D(w)$ are shown in Table 2. Table 3 compares its performance with the two benchmark methods. In general, bookmaker's prediction accuracy is higher than directly observed group choice. Our estimated group belief model further outperforms bookmaker's model in this experiment.

We conduct further experiments to better understand the influence of user's action under risk on our model. By setting different $D(w)$ and bettors' decision parameters on α, we estimate parameters r, s, l and m from simulated data. We apply these parameters in estimating $E(w)$ for each game and get results in Fig. 2. In the figure, the prediction accuracies of two baseline model p_{bettor} and p_{odds} are shown as the gray dots. In general, we notice that the estimated group belief model always achieve better performance than the share of betters method. When $D(w)$ is selected appropriately,

the estimated group belief also have better performance than the Bookmaker Odds method. In many cases, the performances are not very sensitive to α, i.e., the performance may keep stage in a range of α.

Table 2. Parameters estimation using empirical data

Parameter	Estimated Value
r	26.207
s	-4.278
l	-18.245
m	3.735
D(*w*)	0.0053129

Table 3. Prediction accuracy

Methods	Correct Num.	Correct Ratio
Share of Betters	104	104/167=62.28%
Bookmaker Odds	107	107/167=64.07%
Estimated Group Belief	113	113/167=67.66%

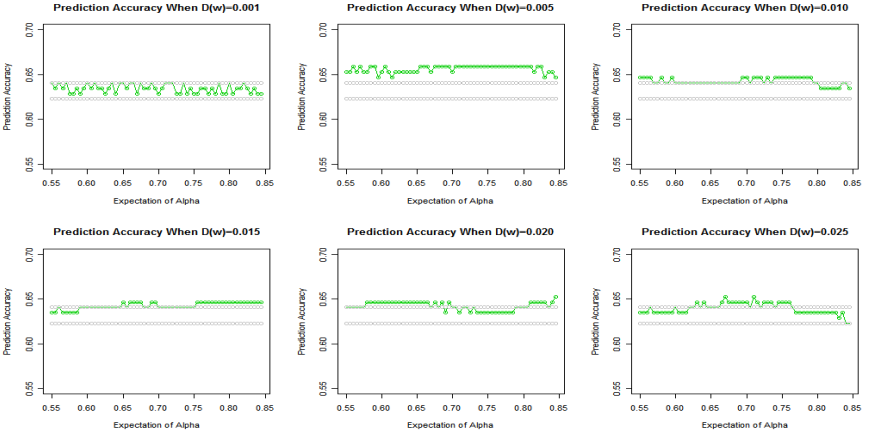


Fig. 2. Prediction results of the model given different parameters

The human nature which is reflected in the decision parameters, say expectation of α, seems play little importance in the ultimate system accuracy. This is opposite to our intuition because we often think that different bettor groups should give different prediction results. However, this phenomenon is consistent with findings in an online contest ProbabilitySports run by mathematician Brian Galebach[15]. In this experiment, most individual participants are poor predictors. However their aggregated (average) predictions are amazing good. Their result shows that the wisdom of crowds

can overcome some limitations in individual participants' decision functions, while our result has the same implications.

Meanwhile, in our settings we find D(w) play a relatively important role in affecting prediction performance. A explanation is that D(w) represents the divergence of bettors' belief to a certain set of sports games. This divergence reflects the group's opinion variance thus affect the effectiveness of the wisdom of crowds. The group's opinion variance is related to event itself. When the event is highly uncertain, the opinion variance of participants is larger, and vice versa. A more uncertain event is obviously more difficult to predict on a more certain one.

5 Conclusion

In this research, we investigated the group belief estimation problem in fixed odds betting. Based on the prospect theory, we include biased belief and people's different reaction to risk into our estimation model. Due to the nature of the problem, we introduce appropriate assumptions on the prospect theory to make is more applicable and simple to deal our problem. Taking a computational experiment approach, we establish a set formula to characterize the relationship between (perceived) group belief, user behavior under risk, odds set by bookmakers, and observed share of betters on each side. Our proposed model can work as an indicator to predict the probability of future events, just as price in dynamic price PMs. Our experimental study shows that by taking the heterogeneity of bettors in to consideration, our model can provide more effective predictions on future events than available methods. Our model hold the potential to enable setting up lightweight prediction markets based on the fixed odds betting mechanism and get the wisdom of crowds for prediction and decision making.

There are three main problems that need to be solved to further advance our model in future research: 1) to consider correlation between parameters in prospect theory for more precise model; 2) to consider the impact of amount of total bettors on the sensitivity of our model; and 3) to derive the analytical solution of the model for mathematical beauty.

Acknowledgments. This work was supported by the following grants: NNSF grants #60921061, #90924302, #71025001, #91024030; MNSTP grants #009ZX10004-315, #2008ZX10005-013; CAS grant #2F070C01; CityU SRG #7002625, #7002518. The authors thank our team member Mr. He Cheng for providing the Olympic Games data set.

References

1. Arrow, K.J., et al.: The Promise of Prediction Markets. Science 320(5878), 877–878 (2008)
2. Schrieber, J.M.: The Application of Prediction Markets to Business. Engineering Systems Division (2004)
3. Servan-Schreiber, E., et al.: Prediction Markets: Does Money Matter? Electronic Markets 14(3), 243–251 (2004)

4. Gray, P.K., Gray, S.F.: Testing Market Efficiency: Evidence from the NFL Sports Betting Market. Journal of Finance 52(4), 1725–1737 (1997)
5. Kuypers, T.: Information and Efficiency: An Empirical Study of a Fixed Odds Betting Market. Applied Economics 32(11), 1353–1363 (2000)
6. Steven, D.L.: Why are gambling markets organised so differently from financial markets? Economic Journal 114(495), 223–246 (2004)
7. Erik, S., Justin, W.: Explaining the Favorite-Long Shot Bias: Is it Risk-Love or Misperceptions? In: Chicago, IL, ETATS-UNIS, vol. 118, pp. 723–746. University of Chicago Press, Chicago (2010)
8. Quandt, R.E.: Betting and Equilibrium. The Quarterly Journal of Economics 101(1), 201–207 (1986)
9. Kahneman, D., Tversky, A.: Prospect Theory: An Analysis of Decision under Risk. Econometrica 47(2), 263–291 (1979)
10. Justin, W., Eric, Z.: Interpreting Prediction Market Prices as Probabilities. Institute for the Study of Labor, IZA (2006)
11. Tversky, A., Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5(4), 297–323 (1992)
12. Bruno, J., Bernard, S.: Estimating Preferences under Risk: The Case of Racetrack Bettors. Journal of Political Economy 108(3), 503–530 (2000)
13. Wang, F.-Y.: Toward a Paradigm Shift in Social Computing: The ACP Approach. IEEE Intelligent Systems 22(5), 65–67 (2007)
14. Wang, F.-Y., et al.: Social Computing: From Social Informatics to Social Intelligence. IEEE Intelligent Systems 22(2), 79–83 (2007)
15. Chen, Y., et al.: Information markets vs. opinion pools: an empirical comparison. In: 6th ACM Conference on Electronic Commerce. ACM Press, Vancsouver (2005)

Two Protocols for Member Revocation in Secret Sharing Schemes

Jia Yu[1], Fanyu Kong[2], Xiangguo Cheng[1], and Rong Hao[1]

[1] College of Information Engineering, Qingdao University,
266071 Qingdao, China
{yujia,hr,chengxg}@qdu.edu.cn
[2] Institute of Network Security, Shandong University,
250100 Jinan, China
fanyukong@sdu.edu.cn

Abstract. Secret sharing scheme plays a very important role in modern electronic applications. In actual circumstance, the members in secret sharing schemes may need to be changed. For example, some new members may join the system and some old members may leave the system. Therefore, how to construct the protocols satisfying these requirements is an important task. In this work, we discuss two protocols about revoking the old members in secret sharing schemes. In the first protocol, other members can make the share of one member leaving secret sharing scheme invalid. At the same time, the corresponding threshold value is unchanged. In the second protocol, publicly verifiable property is added to the first protocol. Thus the validity of the protocol can be verified by anyone besides the members executing the protocol. The both protocols are especially useful for the alterable circumstances such as ad hoc networks.

Keywords: threshold cryptography; secret sharing; publicly verifiable secret sharing; key revocation.

1 Introduction

The secret sharing scheme shares a secret among a group of members. Only qualified subsets of members are able to construct the secret, while unqualified subsets of members are not. For example, in a (t, n) secret sharing scheme [1,2], a secret is divided into n shares that are distributed into n shareholders, respectively. Any authorized subset with t honest shareholders can jointly reconstruct the secret. However, standard secret sharing scheme cannot identify which share is wrong during secret distribution and reconstruction. Verifiable secret sharing (VSS) schemes [3,4] are proposed to deal with this problem. In a verifiable secret sharing scheme, any incorrect share during distribution phase and reconstruction phase can be detected. Publicly verifiable secret sharing (PVSS) scheme [5-8] is one kind of special VSS scheme in which not only the shareholders but also anyone can verify the validity of shares that shareholders have received.

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 64–70, 2011.
© Springer-Verlag Berlin Heidelberg 2011

However, in a regular (publicly verifiable) secret sharing scheme, the shareholders group is always fixed. Some original shareholders sometimes, in fact, have to leave the system due to being corrupted or other reasons. How to dynamically revoke members from a secret sharing system is an important problem. Because the dealer is easy to become a weak point to be attacked, she will be offline after initiation. The motivation of this paper is to put forward protocols that revoke the members in a secret sharing scheme.

Previous works. Numerous dynamic secret sharing schemes have been proposed. Desmedt and Jajodia [9] proposed a protocol, in which the new members and the old members can be absolutely different and the threshold value can be changed. Wong *et al.* [10] gave an improved version to verify the validity of subshares and old shares, which is used for archival systems [11]. Some new protocols [12,13,14,15] have been proposed. All above protocols have not the property of public verifiability. Publicly verifiable member-join protocols were proposed in [16,17]. However, as far as we are concerned, there is little work about the protocols revoking the members in secret sharing schemes.

Our contribution. We discuss two simple protocols to revoke members in secret sharing schemes. In the first protocol, other members can make the share of one member leaving the threshold scheme invalid. At the same time, the corresponding secret and threshold value are unchanged. In the second protocol, publicly verifiable property is added to the protocol. Therefore the validity of the protocol can be verified by everyone. The original ideas of these protocols are taken from the shares renewal algorithm in proactive secret sharing scheme [18] and the verifiable encryption algorithm based on discrete logarithms in [6]. Our proposed protocols can supplementary the member-join protocols [14,15,16,17] that only allow members to join a secret sharing scheme. Thus our protocols can combine other member-join protocols to construct dynamic secret sharing schemes.

2 The Proposed Protocols

2.1 Notations and Assumptions

p and q are primes *s.t.* $q \mid p-1$. Let G denote a group with prime order p and g be a generator of group G. Let $h \in Z_p^*$ be an element of order q. The shareholders group P is composed by $P_1, P_2, ..., P_n$ and the secret s is shared by a (t,n) secret sharing scheme among them. Assume the revoked member is P_b. Initially set $B = \{b\}$.

For simplifying the description, we assume that there is a dealer during initiation. Furthermore, the system is synchronized, i.e., all members can send their messages simultaneously in the same round.

2.2 The First Protocol

The protocol is composed of two phases. The first phase is secret distribution phase. In this phase, a dealer distributes the shares of a secret into a group of shareholders.

This procedure is the same as Feldman's secret sharing scheme [3]. The second phase is member revocation phase. In this phase, each shareholder except the revoked one selects a random polynomial with the constant item equating to zero. Each random share computed by this polynomial is given to each shareholder except the revoked one. At last, each shareholder except the revoked one can update his share and the share of the revoked one is invalid. The both phases are described as follows:

① The Secret Distribution Phase

(1) The dealer randomly selects a polynomial $f(x) = s + \sum_{i=1}^{t-1} a_i x^i \in Z_p[x]$, and computes $s_i = f(i)$, $i = 1, 2, ..., n$. The dealer broadcasts g^s, $g^{a_i} (i = 1, 2, .., t-1)$.

(2) Each member $P_i (i = 1, 2, ..., n)$ verifies equation $g^{s_i} = g^s \prod_{j=1}^{t-1} (g^{a_j})^{i^j}$ holds or not. If it holds, P_i believes his share is correct and sets $E_i = g^{s_i}$. Otherwise, publishes s_i and broadcasts a complaint against the dealer.

② Member Revocation Protocol

When the system will revoke member P_b, other members except P_b execute the following protocol to disuse his share.

(1) Each member $P_j (j \neq b)$ selects a polynomial $f_j(x) = \sum_{l=1}^{t-1} a_{jl} x^l \in Z_p[x]$. He computes $s_{ji} = f_j(i)$, $i \in \{1, 2, ..., n\} / \{b\}$, sends s_{ji} to $P_i (i \neq j, i \neq b)$ secretly, and broadcasts $g^{a_{jl}} (l = 1, 2, .., t-1)$.

(2) Each member $P_i (i \neq j, i \neq b)$ verifies equation $g^{s_{ji}} = \prod_{l=1}^{t-1} (g^{a_{jl}})^{i^l}$ holds or not. If it holds, P_i believes the share s_{ji} he got is correct and sets $E_{ji} = g^{s_{ji}}$. Otherwise, he publishes s_{ji}, and broadcasts a complaint against P_j. If other members find P_j is dishonest, set $B = B \cup \{j\}$.

(3) Each member $P_i (i \neq b)$ updates his new share $s_i = s_i + \sum_{j \in \{1...n\} \setminus B} s_{ji} \pmod p$.

2.3 The Second Protocol

The second protocol uses verifiable encryption technique to add publicly verifiable property to this protocol. Each member $P_i (i = 1, 2, ..., n)$ randomly selects $x_i \in_R Z_p$, and then publishes $y_i = h^{x_i}$. Let $H : \{0,1\}^* \to \{0,1\}^l$ be a secure hash function.

① The Secret Distribution Phase

(1) The dealer randomly selects a polynomial $f(x) = s + \sum_{i=1}^{t-1} a_i x^i \in Z_p[x]$ and computes $s_i = f(i)$, $i = 1, 2, ..., n$. The dealer broadcasts g^s, $g^{a_i} (i = 1, 2, .., t-1)$.

(2) The dealer encrypts each s_i by a variation of ElGamal encryption algorithm:

She selects $l_i \in_R Z_q$, computes $\gamma_i = h^{l_i} (\text{mod } p)$ and $\delta_i = s_i^{-1} y_i^{l_i} (\text{mod } p)$, and publishes (γ_i, δ_i) as the ciphertext of the share s_i. And then selects $w_k \in Z_q$, $k = 1, 2, ..., l$, computes and broadcasts $T_{h,i,k} = h^{w_k} (\text{mod } p)$ and $T_{g,i,k} = g^{y_i^{w_k}}$, where $i = 1, 2, ..., n$.

She computes

$$c_i = H(g \| h \| \gamma_i \| \delta_i \| T_{h,i,1} \| T_{h,i,2} \| ... \| T_{h,i,l} \| T_{g,i,1} \| T_{g,i,2} \| ... \| T_{g,i,l}) \quad (1)$$

(3) Let $c_{i,k}$ denote the *k*-th bit of c_i. The dealer computes $r_{i,k} = w_k - c_{i,k} l_i$, where $k = 1, 2, ..., l$ and publishes $\text{Pr}oof_D = (c_i, r_{i,1}, ..., r_{i,l})$.

(4) Each member $P_i (i = 1, 2, ..., n)$ decrypts $s_i = \gamma_i^{x_i} \cdot \delta_i^{-1} (\text{mod } p)$ and verifies equation $g^{s_i} = g^s \prod_{j=1}^{t-1} (g^{a_j})^{i^j}$ holds or not. If it holds, P_i believes his share is correct and sets $E_i = g^{s_i}$. Otherwise, publishes s_i and broadcasts a complaint against the dealer.

(5) Each member $P_j (j = 1, 2, ..., n)$ checks the validity of share s_i $(i \neq j)$. She computes $E_i = g^s \prod_{j=1}^{t-1} (g^{a_j})^{i^j}$, $T_{h,i,k} = h^{r_{i,k}} \gamma_i^{c_{i,k}}$, and $T_{g,i,k} = (g^{1-c_{i,k}} E_i^{c_{i,k} \delta_i})^{y_i^{r_{i,k}}}$.

And then verifies whether equation (1) holds. If it holds, then believes s_i is correct. Otherwise, generates a complaint against the dealer.

② Member Revocation Protocol

When the system will revoke member P_b, other members execute the following protocol to disuse his share.

(1) Each member $P_j (j \neq b)$ selects a polynomial $f_j(x) = \sum_{l=1}^{t-1} a_{jl} x^l \in Z_p[x]$ and computes $s_{ji} = f_j(i)$, $i \in \{1, 2, ..., n\} / \{b\}$. $P_j (j \neq b)$ broadcasts $g^{a_{jl}} (l = 1, 2, .., t-1)$.

(2) Member $P_j (j \neq b)$ encrypts each s_{ji} by a variation of ElGamal encryption algorithm:

She selects $l_{ji} \in_R Z_q$, computes $\gamma_{ji} = h^{l_{ji}} (\text{mod } p)$ and $\delta_{ji} = s_{ji}^{-1} y_i^{l_{ji}} (\text{mod } p)$, and publishes $(\gamma_{ji}, \delta_{ji})$ as the ciphertext of the share s_{ji}. And then selects $w_k \in Z_q$, $k = 1, 2, ..., l$, computes and broadcasts $T_{h,j,i,k} = h^{w_k} (\text{mod } p)$ and $T_{g,j,i,k} = g^{y_j^{w_k}}$, where $j = 1, 2, ..., n$.

She computes

$$c_{ji} = H(g \| h \| \gamma_{ji} \| \delta_{ji} \| T_{h,j,i,1} \| T_{h,j,i,2} \| ... \| T_{h,j,i,l} \| T_{g,j,i,1} \| T_{g,j,i,2} \| ... \| T_{g,j,i,l}) \quad (2)$$

(3) Let $c_{j,i,k}$ denote the k-th bit of c_{ji}. The member computes $r_{j,i,k} = w_k - c_{j,i,k} l_{ji}$, where $k = 1, 2, ..., l$ and publishes $\Pr oof_D = (c_i, r_{j,i,1}, ..., r_{j,i,l})$.

(4) Each member $P_i (i = 1, 2, ..., n)$ decrypts $s_{ji} = \gamma_{ji}^{x_i} \cdot \delta_{ji}^{-1} (\mathrm{mod}\ p)$, and verifies the equation $g^{s_{ji}} = \prod_{l=1}^{t-1} (g^{a_{jl}})^{i^l}$ holds or not. If it holds, P_i believes his share is correct and sets $E_{ji} = g^{s_{ji}}$. Otherwise, publishes s_{ji}, broadcasts a complaint against P_j and makes the set of dishonest members $B = B \cup \{j\}$.

(5) Each member $P_i (i = 1, 2, ..., n)$ updates his new share $s_i = s_i + \sum_{j \in \{1...n\} \setminus B} s_{ji} (\mathrm{mod}\ p)$.

(6) Each member $P_m (m = 1, 2, ..., n)$ checks the validity of share s_{ji} $(i \neq j, i \neq b)$. She computes $E_j = \prod_{l=1}^{t-1} (g^{a_l})^{i^l}$, $T_{h,j,i,k} = h^{r_{j,i,k}} \gamma_{ji}^{c_{j,i,k}}$, and $T_{g,j,i,k} = (g^{1-c_{j,i,k}} E_j^{c_{j,i,k} \delta_j})^{y_j^{r_{j,i,k}}}$

And then verifies whether equation (2) holds. If it holds, then believes s_{ji} is correct.

3 Security Analysis

Theorem 1. After the member revocation phases in protocol 1 and protocol 2, any t honest shareholders can recover the real secret using their new shares.

Proof.
It is because:

$$s_i^{new} = s_i^{old} + \sum_{j \in \{1...n\} \setminus B} s_{ji} (\mathrm{mod}\ p)$$

We assume that the set of the t honest shareholders is $A(|A|=t)$

$$\begin{aligned} \sum_{i \in A} C_{Ai} s_i^{new} &= \sum_{i \in A} C_{Ai} (s_i^{old} + \sum_{j \in \{1...n\} \setminus B} s_{ji}) \\ &= \sum_{i \in A} C_{Ai} s_i^{old} + \sum_{j \in \{1...n\} \setminus B} \sum_{i \in A} C_{Ai} s_{ji} \\ &= s + \sum_{j \in \{1...n\} \setminus B} f_j(0) \\ &= s + \sum_{j \in \{1...n\} \setminus B} 0 \\ &= s \end{aligned}$$

Theorem 2. After the member revocation phases in protocol 1 and protocol 2, the share of the revoked shareholder cannot be valid with a non-negligible probability.

Theorem 3. In our proposed member revocation protocols, the adversary cannot get any useful message about the secret and the shares of the honest shareholders.

The formal proofs of theorem 2 and theorem 3 are omited in this version.

4 Conclusions

We discuss two protocols for revoking members in secret sharing schemes. These protocols can verifiably or publicly verifiably revoke members in secret sharing schemes. We make use of the shares renewal algorithm in proactive secret sharing scheme and the verifiable encryption algorithm based on discrete logarithms to reach this aim. These new protocols can combine other member-join protocols to construct dynamic secret sharing schemes.

Acknowledgments. This research is supported by National Natural Science Foundation of China (60703089) and the Shandong Province Natural Science Foundation of China (ZR2010FQ019, ZR2009GQ008, ZR2010FQ015).

References

1. Shamir, A.: How to Share a Secret. Communications of the ACM 22(11), 612–613 (1979)
2. Blakley, G.R.: Safeguarding cryptographic keys. In: Proc. AFIPS 1979 National Computer Conference, vol. 48, pp. 313–317. AFIPS Press, NJ (1979)
3. Feldman, P.: A Pratical Scheme for Non-Interactive Verifiable Secret Sharing. In: Proc. 28th Annual FOCS, pp. 427–437. IEEE Press, New York (1987)
4. Pedersen, T.P.: Non-Interactive and Information-Theoretic Secure Verifiable Secret Sharing. In: Feigenbaum, J. (ed.) CRYPTO 1991. LNCS, vol. 576, pp. 129–140. Springer, Heidelberg (1992)
5. Schoenmakers, B.: A simple Publicly Verifiable Secret Sharing Scheme and its Application to Electronic Voting. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 148–164. Springer, Heidelberg (1999)
6. Stadler, M.A.: Publicly verifiable secret sharing. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 190–199. Springer, Heidelberg (1996)
7. Fujisaki, E., Okamoto, T.: A practical and provably secure scheme for publicly verifiable secret sharing and its applications. In: Nyberg, K. (ed.) EUROCRYPT 1998. LNCS, vol. 1403, pp. 32–47. Springer, Heidelberg (1998)
8. Young, A., Yung, M.: A PVSS as Hard as Discrete Log and Shareholder Separability. In: Kim, K.-c. (ed.) PKC 2001. LNCS, vol. 1992, pp. 287–299. Springer, Heidelberg (2001)
9. Desmedt, Y., Jajodia, S.: Redistributing secret shares to new access structures and its application. Technical Report ISSE TR-97-01, George Mason University (1997)
10. Wong, T.M., Wang, C., Wing, J.M.: Verifiable secret redistribution for archive systems. In: Proceeding of the 1st International IEEE Security in Storage Workshop, pp. 94–106. IEEE Press, Los Alamitos (2002)

11. Wong, T.M., Wang, C.X., Wing, J.M.: Verifiable secret redistribution for archive systems. In: Proc. of the 1st International IEEE Security in Storage Workshop, pp. 94–106. IEEE Press, New York (2002)
12. Gupta, V., Gopinaph, K.: An Extended Verifiable Secret Redistribution Protocol for Archival Systems. In: The First International Conference on Availability, Reliability and Security 2006, pp. 8–15. IEEE Press, New York (2006)
13. Yu, J., Kong, F.Y., Li, D.X.: Verifiable Secret Redistribution for PPS Schemes. In: Proc. of the 2nd Information Security Practice and Experience Conference. Journal of Shanghai Jiaotong University (Science), vol. E-11(2), pp. 71–76 (2006)
14. Li, X., He, M.X.: A protocol of member-join in a secret sharing scheme. In: Chen, K., Deng, R., Lai, X., Zhou, J. (eds.) ISPEC 2006. LNCS, vol. 3903, pp. 134–141. Springer, Heidelberg (2006)
15. Yu, J., Kong, F.Y., Hao, R., Cheng, Z.: A Practical Member Enrollment Protocol for Threshold Schemes. Journal of Beijing University of Posts and Telecommunications 28(z.2), 1–3,8 (2006) (in Chinese)
16. Yu, J., Kong, F.Y., Hao, R.: Publicly Verifiable Secret Sharing with Enrollment Ability. In: The 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 194–199. IEEE Computer Society, New York (2007)
17. Yu, J., Kong, F.Y., Hao, R., Li, X.L.: How to Publicly Verifiably Expand a Member without Changing Old Shares in a Secret Sharing Scheme. In: Yang, C.C., Chen, H., Chau, M., Chang, K., Lang, S.-D., Chen, P.S., Hsieh, R., Zeng, D., Wang, F.-Y., Carley, K.M., Mao, W., Zhan, J. (eds.) ISI Workshops 2008. LNCS, vol. 5075, pp. 138–148. Springer, Heidelberg (2008)
18. Herzberg, A., Jarecki, S., Krawczyk, H., Yung, M.: Proactive Secret Sharing or: How to Cope with Perpetual Leakage. In: Coppersmith, D. (ed.) CRYPTO 1995. LNCS, vol. 963, pp. 339–352. Springer, Heidelberg (1995)

Dual-Verifiers DVS with Message Recovery for Tolerant Routing in Wireless Sensor Networks

Mingwu Zhang[1,2], Tsuyoshi Takagi[2], and Bo Yang[1]

[1] College of Informatics, South China Agricultural University, China
[2] Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan
{mwzhang,takagi}@imi.kyushu-u.ac.jp

Abstract. In wireless sensor networks, message authentication needs higher computation efficiency and lower communication cost since specific nodes are more vulnerable to various attacks due to the nature of wireless communication in public channel whereas sensor nodes are equipped with limited computing power, storage, and communication modules. Designated verifier signature scheme enables a signer node to convince the designated verifier that a signed message is authentic where neither the signer nor the verifier can convince any other party of authenticity of that message. In this paper, we explore a dual designated verifier signature scheme with message recovery for dual routing paths in a fault-tolerant manner. It supports the redundant routing path based message authentication under a dual route path to the goal base stations. The message is hidden in the signature that protects the privacy of both the sender's identity and message, thus only designated receivers can extract and verify the message. The proposed scheme can be applied to tiny and short signature requirements for lightweight message authentication in wireless sensor networks.

keywords: designated verifier signature, wireless sensor network, Diffie-Hellman, lightweight authentication.

1 Introduction

Wireless sensor networks (WSNs) are ad hoc mobile networks that include sensor nodes with limited computational and communication capabilities. WSNs have become an economically viable monitoring solution for a wide variety of applications. However, sensor nodes are equipped with limited computing power, storage, and communication modules, thus message authentication in such resource-constrained environment is a critical security concern. Providing authenticity and privacy in WSNs poses different challenges than traditional network computer. WSNs are more vulnerable to various attacks due to the nature of wireless communication[12], so authentication for sensed data with highly efficient and low cost communicating will have an especial space in this field. Lightweight digital signatures and key-exchange based on asymmetric algorithms should be very valuable manner in WSNs. There are several schemes in which RSA and

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 71–84, 2011.
© Springer-Verlag Berlin Heidelberg 2011

ECC are implemented for sensor nodes, though varius of innovative asymmetric algorithms exist[4]. Designated verifier signature schemes (DVS)[6,22,3] are constructed to address the privacy issues for signer and verifier by preventing the signature from being arbitrarily disseminated, which is suitable for WNSs since sensors usually have very constrained resources in computing, communication, storage, and battery power.

Security should be supported by fundamental operations in WSNs in order to provide a stable infrastructure that will be able to handle authentication effectively to prevent malicious nodes from compromising the WSNs. Traditionally, in a single path routing between source node and goal base station, a failed node may cause the path failure and date loss[26]. Thus effective authentication cannot be provided due to malicious activities or network problems. Storage within the network is therefore designed to support multiple paths and potentially long timescales, and not to require but to support end-to-end reliability[18].

Dual routing paths provide a chain fault-tolerance manner to transfer the sensor data[26,23]. Up to now, many of message authentication schemes provide single routing path authentication[5,24]. If a scheme fulfils the message authentication in dual paths, it needs to use broadcast[2], multi-verifier[14,3], or send the signature twice[7,17], which will consume extra communication resources.

1.1 Related Works

Eliana and Andreas[5] surveyed the current state-of-the-art of secure multipath routing protocols against failure of the single routing path in WSNs, classifid the protocols into categories in accordance with their security-related operational objectives, defined a new threat model in the routing procedure, and identified open research issues in the WSNs. Li et al.[13] reviewed two main categories of privacy-preserving techniques for protecting two types of private information: data-oriented privacy and context-oriented privacy. Some light-weight authentication models for wireless sensor networks have also been proposed[19,24].

Driessen et al.[4] investigated the efficiency and suitability of digital signature algorithms on the basis of innovative asymmetric primitives for WSNs under the single routing path. Liu et al.[17] presented an online/offline identity-based signature scheme for the wireless sensor network, which can be suitable with severely constrained resources due to significantly reduced in costs of computation and storage. It also provides multi-time usage of the offline storage.

DVS is a special type of digital signature. Formal definitions of security definitions for the DVS schemes are considered by Li et al.[15], including the notions of unforgeability, non-delegatability, and non-transferability. Unforgeability means that no third party(except the sender and designated receiver) can forge a valid signature with non-negligible probability, which is the indispensable property for a signature scheme. The notion of non-delegatability describes as a strong DVS scheme that neither A nor B can delegate his ability to generate signatures to the third party without giving the third party his secret key. Non-transferability means that given a message-signature pair that is accepted by a verifier, it is

infeasible for any probabilistic polynomial-time distinguisher to tell whether the message was signed by the signer or the designated verifier.

Saeednia et al.[22] formalized the notion of strong DVS (sDVS) and proposed an efficient scheme. Later, Lee and Chang [11] pointed out that Saeednia et al.'s scheme would reveal the identity of the signer if the secret key of this signer was compromised, where Saeednia et al.'s scheme prevents the third party from reading the message upon seeing the signature since the secret key of designated verifier is involved in the verification algorithm. Huang et al.[7] presented a tiny identity-based DVS scheme. The size of the signature of their scheme is short of all existing schemes that are better deploy in WSNs. However, in their scheme, the signature is short of randomicity for the same message and has the same signatures in every signature produced[21].

Message recovery need mean that the signature does not provide the message plaintext enclosed with the signature directly. So it may save on the communication cost, which has the advantage of being deployable in WSNs. Lee et al. proposed a sDVS scheme[10], that takes advantages of message recovery to hide messages. However regarding efficiency of Lee et al. protocol, expensive computation with a high volume of messages must be made, and thus efficiency could not achieved as expected. Yang et al.[25] proposed an efficient strong designated verifier signature scheme with message recovery, using key distribution mechanism where both sender and designated receiver share an encryption/decryption key. An obvious drawback of this scheme is that it enables messages only $|m| < log(p/q)$[21] in length to be signed.

1.2 Our Contribution

We provide a dual routing paths-based message authentication scheme with a failure-tolerant nature, which deploys a two-verifier DVS scheme. Under this circumstance, we propose a *dual* designated verifiers sDVS scheme (sDDVS) such that a sender A (sensor node) sends a message signature to two designated receiving base stations B and C. The sDDVS scheme is different from a designated multiple receivers signature scheme[14] lies in that sDDVS has a fixed signature size (with two verifiers) but the signature size is increasing with the receivers in multiple-receiver sDVS. Thus multiple-receivers sDVS is limited for large scale sensor nodes in WSNs.

The sDDVS scheme supports the validity verification of the message, and only the node A and receivers B and C can generate the indistinguishable signature in WSNs. In our scheme, the message is compacted in the signature and it can be recovered by a designated receiver . Our contribution is described as follows:

- Our proposed scheme supports a single sensor node A and two designated verifier base stations B and C in such a way that any base station can verify the validity of the signature signed by signer A, where no any other party (common sensor node or base station) can distinguish the signature validity by the sender or the receiver. It provides the privacy preservation between sending node and receiving base station in WSNs.

- The message can be recovered by VERI algorithm after the integrity and designated verifier properties are checked in the signature. Informally, our scheme needs not provide the message plaintext in the signature verification procedure so as to protect the confidentiality of the signed message and reduce the communication band.
- To achieve the higher efficiency and compact signature size, the tri-party including a signer A and two verifiers B and C will negotiate and decide a session key non-interactively by deploying a bilinear Deffie-Hellman session value.
- Compared with related schemes, our proposed scheme not only supports message recovery ability but also has higher computation efficiency and small signature size that can perfectly deployed for applications in WSNs.

The remainder of the paper is organized as follows. Section 2 gives some preliminaries including bilinear pairing and security assumptions. Section 3 describes the formal model and Section 4 provides the construction of sDDVS. Section 5 gives security analysis and Section 6 evaluates performance. Finally, Section 7 draws the conclusion.

2 Strong Dual-Designated Verifiers Signature in WSNs

In wireless sensor network, we consider a common sensor node A to send a message to two base stations B and C, and then base station B or C sends the message to control center S. In Fig.1, to avoid the single-path failure problem, the node A sends a message to the control center via two base stations B, C with different routing paths in a fault-tolerant manner. After receiving the signature from node A, either of station B or C can verify that the message come from source node A where other nodes can neither distinguish the message (hidden in the signature), nor know the sender and receiver of his re-transferred signature.

We present a generic definition of a strong dual designated verifiers signature (sDDVS) to implement message authentication in fault-tolerant routing paths. Let A be the signer (sender sensor node), and B, C be the designated verifiers (base stations). Let $\mathcal{M}$ be the message space. An sDDVS scheme is comprised of the following five algorithms:

- SETUP(k) is a probabilistic algorithm that outputs the public parameter pp, where k is the security parameter.
- KEYGEN(pp) is a probabilistic algorithm that takes the public pp as input and outputs secret key SK and public key pair PK.
  We can repeat call KEYGEN algorithm to produce key pairs all nodes and base stations. We denote by (PK_A, SK_A), (PK_B, SK_B) and (PK_C, SK_C) as the public and secret pairs of node A, and base stations B and C.
- SIGN(m, SK_A, PK_B, PK_C) takes as input signer's secret key, designated verifiers' public key, a message $m \in \mathcal{M}$, and outputs a signature σ, denoted as $\sigma \leftarrow$ SIGN(m, SK_A, PK_B, PK_C).

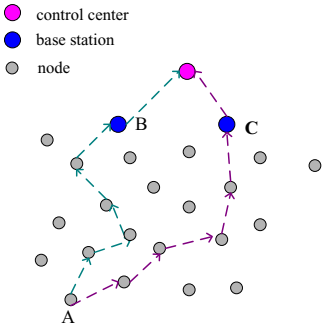


Fig. 1. Dual routing path for WSNs

- $\text{VERI}(PK_A, SK_B|SK_C, \sigma)$ is a deterministic algorithm that takes as input a signing public key PK_A, secret key of either designated verifier B or C, and a candidate signature σ, and returns 1 if signature is valid or $\perp$ otherwise.
- $\text{SIMU}(m, SK_B|SK_C, PK_A)$ is a probabilistic algorithm producing signatures indistinguishable from those produced by $\text{SIGN}(m, SK_A, PK_B, PK_C)$.

In particular, for every security parameter k, every tuple $(PK_A, SK_A, PK_B, SK_B, PK_C, SK_C)$ generated by KEYGEN, and every message $m \in \mathcal{M}$, the correctness should hold that

$$\begin{cases} pp \leftarrow \text{SETUP}(1^k) \\ (PK_A, SK_A) \leftarrow \text{KEYGEN}(pp) \\ (PK_B, SK_B) \leftarrow \text{KEYGEN}(pp) \\ (PK_C, SK_C) \leftarrow \text{KEYGEN}(pp) \\ 1 \leftarrow \text{VERI}(\text{SIGN}(m, SK_A, PK_B, PK_C), \\ \qquad SK_B|SK_C, PK_A, PK_C|PK_B) \\ 1 \leftarrow \text{VERI}(\text{SIMU}(m, PK_A, SK_B, PK_C), \\ \qquad SK_B|SK_C, PK_A, PK_C|PK_B) \end{cases}$$

Definition 1. Unforgeability: *Unforgeability requires that any third party other than signer A and designated verifiers B and C cannot forge a signature on behalf of A with non-negligible probability. Formally, unforgeability is defined by the following game played between a game challenger $\mathcal{C}$ and an adversary $\mathcal{F}$:*

1. Challenger $\mathcal{C}$ generates the key pairs of participants A, B, and C, i.e. key pairs as (PK_A, SK_A), (PK_B, SK_B) and (PK_C, SK_C) respectively, and gives (PK_A, PK_B, PK_C) to $\mathcal{F}$;
2. $\mathcal{F}$ performs the adaptively queries to the following oracles for polynomially bounded times:
   - Hash Oracles($\mathcal{O}_{Hash}$). Given a message m and an integer $r \in \mathbb{F}_q^*$, this oracle outputs a hash value of $\mathbb{F}_q^*$;
   - SIGN Oracles($\mathcal{O}_{Sign}$). Given a message m, this oracle returns a signature σ on m to $\mathcal{F}$, which is valid with respect to signer SK_A and designated verifiers PK_B, PK_C;

- SIMU Oracles($\mathcal{O}_{Simu}$). Given a message m, this oracle returns a signature σ on m to $\mathcal{F}$, which is valid with respect to PK_A, PK_B and SK_B;
- VERI Oracles($\mathcal{O}_{Veri}$). Given a query of the form σ, this oracle first recovers the message m and returns 1 if it is a valid signature on m with respect to PK_A and PK_B, PK_C, and $\perp$ otherwise.

3. Finally, $\mathcal{F}$ outputs its forgery σ^* on message m^* and wins the game if
   - $\mathcal{F}$ does not query σ^* and SIMU oracle on message m^* for designated sender and receivers, and
   - VERI$(PK_A, SK_B|SK_C, PK_C|PK_B, \sigma^*) = 1$.

To give the definition of Non-transferability. we first define the sDDVSgame for two designated verifiers' property as follows:

1. Run SETUP(k) to obtain system parameters;
2. Perform KEYGEN algorithm to generate sender A and receivers B and C's public/secret key pairs;
3. A random bit $b \leftarrow \{0, 1\}$ is chosen;
4. Adversary $\mathcal{F}$ is given SK_A, PK_B, PK_C and oracle access to either $\sigma_0 \leftarrow$ SIGN$(\cdot, SK_A, PK_B, PK_C)$ if b=0, or $\sigma_1 \leftarrow$ SIMU$(\cdot, SK_B, PK_A, PK_C)$ if b=1.
5. $\mathcal{F}$ outputs b' be the guess of experiment b is 1 if $b = b'$, and 0 otherwise.

Given a message-signature pair (m, σ) that is accepted by a designated verifier, and without access to the signer's secret key, it is computationally infeasible to determine whether the message was signed by the signer, or the signature was simulated by the designated verifier.

Definition 2. Non-transferability: *Let* Π = (SETUP, KEYGEN, SIGN ,VERI, SIMU) *be a sDDVS scheme for sender A and receivers B and C, we say that sDDVS has strong designated verifier property if for any probabilistic polynomial time adversary* $\mathcal{F}$ *there exists a negligible function* $\epsilon(k)$ *in sDDVSgame.*

$$Pr[sDDVSgame_{\mathcal{F},\Pi}(k)] = 1/2 + \epsilon(k)$$

Definition 3. Non-delegatability: *Let* $\kappa \in [0, 1]$ *be the knowledge error.* Π = (SETUP, KEYGEN, SIGN, VERI, SIMU) *is* (τ, κ)*-non-delegatable if there exists a black-box knowledge extractor* K *that, for every algorithm* $\mathcal{F}$ *and for every valid signature* σ*, satisfies the following condition: for every* $(SK_A, PK_A) \leftarrow$ KEYGEN*, a set of receivers* (SK_{B_i}, PK_{B_i}), $(SK_{C_i}, PK_{C_i}) \leftarrow$ KEYGEN*, for* $i \in [1, n]$*, and message* m*, if* $\mathcal{F}$ *produces a valid signature on* m *with probability* $\epsilon > \kappa$*, then on input* m *and on access to the oracle* $\mathcal{F}_m$*,* K *produces one of the secret keys* $(SK_A, \{SK_{B_1}, \ldots, SK_{B_n}\}|\ \{SK_{C_1}, \ldots, SK_{C_n}\})$ *in expected time* $\tau/(\epsilon - \kappa)$.

The formal definition for non-delegability was first proposed by Lipmaa, Wang and Bao in[16]. In sDDVS, non-delegatability requires that if one produces a valid signature on a message m, the secret key of the signer and the verifier must be known.

3 Construction

We propose an efficient strong designated verifier signature scheme with two verifiers(sDDVS) that supports tolerant routing path in WSNs. When the sender node wants to communicate with the designated base stations, the sender first generates a bilinear Diffie-Hellman key $S_a = \hat{e}(P_B, P_C)^{x_a}$ to encrypt the target message. Upon receiving the message, the designated base station B generates a decryption key $S_b = \hat{e}(P_A, P_C)^{x_B}$, then it recovers the target message and verifies its validity from the signature. It is easy to see that these third parties have the same negotiation session value that $S_A = S_B = S_C = \hat{e}(P, P)^{x_A x_B x_C}$.

Our proposed strong designated two verifiers signature scheme is comprised of the following algorithms:

- SETUP: PKG generates two groups $(\mathbb{G}_1, +)$ and $(\mathbb{G}_2, \cdot)$ of prime order $q(q \geq 2^k)$ where k is the system security parameter, and a bilinear pairing $\hat{e} : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$, together with an arbitrary generator $P \in \mathbb{G}_1$. It also chooses a cryptographically collision-resistant hash function $H : \{0,1\}^* \to \mathbb{F}_q^*$. Then the public parameters is $pp = (\hat{e}, \mathbb{G}_1, \mathbb{G}_2, q, P, H)$.
- KEYGEN: PKG picks random $x \leftarrow_R \mathbb{F}_q^*$, computes $PK = [x]P$. Its public and secret key pair is (PK, SK)=$([x]P, x)$.
  In particular, the sender A's key pair is $([x_A]P, x_A)$, and two receivers B and C's key pairs are $([x_B]P, x_B)$, $([x_C]P, x_C)$, respectively.
- SIGN: To sign a message $m \in \mathcal{M}$ to two designated receivers B and C, sender A performs as follows:
  1. Randomly picks $r \leftarrow \mathbb{F}_q^*$;
  2. Produces session value as $T = \hat{e}(PK_B, PK_C)^{x_A} = \hat{e}(P, P)^{x_A x_B x_C}$, where $PK_B = x_B P$ and $PK_C = x_C P$;
  3. Computes $y = H(m||r) \in \mathbb{F}_q^*$, and $W = (m||r) \oplus T^y$;
  4. Outputs the signature $\sigma = (y, W)$.
- VERI: Either of designated receiver B or C may recover and convince that the signature $\sigma = (y, W)$ comes from the sender A to be valid (we assume receiver B performs the recovery and verification procedure. Note that C can perform the similar procedure). Receiver B does:
  1. Uses the other two parties' public key PK_A , PK_C and B's private key to generate session value $\tilde{T} = \hat{e}(PK_A, PK_C)^{x_B y} = \hat{e}(P, P)^{x_A x_B x_C y}$;
  2. Recovers the message m with B's secret key x_B by computing $m||r = W \oplus \tilde{T}$;
  3. Verifies the validity of the message by the following equation

     $$H(m||r) = y$$

     If the above equation holds, B accepts the message m and convinced m is valid from signer A; otherwise outputs $\perp$ as failure.
- SIMU: The receiver B or C can simulate the designated verifier signature $\sigma = (y, W)$ of message m. He does (taking B as an example):

1. Picks a random value $r' \leftarrow \mathbb{F}_q^*$;
2. Computes $\tilde{y} = H(m||r')$;
3. Computes $\tilde{T} = \hat{e}(PK_A, PK_C)^{x_B}$;
4. Computes $\tilde{W} = (m||r') \oplus \tilde{T}^{\tilde{y}}$;
5. Generates the simulated signature $\sigma = (\tilde{y}, \tilde{W})$.

4 Proof of the Security

The sDDVS scheme supports the security properties such as unforgeability, non-transferability, strong designated verifier etc.

Theorem 1. Unforgeability. *Suppose H modeled as a random oracle, if an adversary $\mathcal{F}$ can forge a valid sDDVS signature $\sigma^* = (y^*, W^*)$ with a non-negligible advantage ϵ, then there exist an algorithm to solve bilinear Diffie-Hellman problem with non-negligible advantage $\epsilon' \geq \epsilon + 1/q_H$, where q_H is the number of hash queries.*

Proof. Suppose the sender A's public key be PK_A, the designated two verifiers B and C's public key be PK_B and PK_C, respectively. Publish public keys of A, B and C. We regard the hash function as the random oracle H.

Let $\mathcal{F}$ be a polynomial adversary attacking the proposed sDDVS scheme. We can construct an algorithm $\mathcal{B}$, which has oracle access ability and solves a bilinear Diffie-Hellman problem whenever $\mathcal{F}$ forges a valid signature. Informally, $\mathcal{B}$ receives quartuple $(P, [x]P, [y]P, [z]P)$, his goal is to compute $\hat{e}(P, P)^{xyz}$. The simulation is described as follows.

$\mathcal{B}$ simulates all the oracles in the proof to answer $\mathcal{F}$'s queries. In the simulation, $\mathcal{B}$ will maintain a list L (L-List) and records the H hash queries and the corresponding values. This L-List consists of the items (m, r, κ), where (m, r) is the input of the hash and κ is the output. L is initially empty.

Let q_H, q_S, q_V denote the number of queries to the H-hash, Sign oracle and Veri oracle. The requirement is that m^* must not be queried to the Sign oracle.

Thus Simu oracle is indistinguishable to Sign oracle (this is called strong designated verifier property and is analyzed in **Theorem 3**. We ignore the Simu query here).

1. $\mathcal{B}$ generates the parameters $pp = (\hat{e}, \mathbb{G}_1, \mathbb{G}_2, q, P, H)$ to $\mathcal{F}$. $\mathcal{B}$ also sets and sends the $PK_A = [x]P$, $PK_B = [y]P$, $PK_C = [z]P$ to $\mathcal{F}$;
2. $\mathcal{B}$ maintains a list L queried by $\mathcal{F}$ to H hash oracle. If $\mathcal{F}$ queries the H oracle $H(m, r)$, $\mathcal{B}$ answers as follows:
   - Searches entry (m, r, κ) in list L, answers with κ if such an entry exists in L;
   - Otherwise, random picks $\kappa \leftarrow_R \mathbb{F}_q^*$, answers with κ and store (m, r, κ) to L.
3. If $\mathcal{F}$ makes a query to its verification oracle Veri (y, W), $\mathcal{B}$ answers as follows:
   - Searches entry (m, r, κ) in list L, returns 0 if there is no such an entry exists;

- Computes $D = (W \oplus (m||r))^{1/\kappa} = \hat{e}(P,P)^{xyz}$ as the solution of the bilinear Diffie-Hellman problem and terminates the execution, otherwise returns 0 as an answer to $\mathcal{F}$'s verification oracle query.

4. At the end of $\mathcal{F}$'s execution, $\mathcal{F}$ outputs a pair (y^*, W^*). Check whether it is a valid signature similarly as in the previous step. If (y^*, W^*) is a valid signature, outputs C, otherwise outputs $\perp$.

We assume that each signature query SIGN must perform Hash query first. After bounded polynomial time queries, forger $\mathcal{F}$ forges a valid signature $\sigma^* = (y^*, W^*)$ which has not been queried by SIGN query. The advantage that $\mathcal{B}$ solves the bilinear Diffie-Hellman problem is $\epsilon' \geq \frac{1}{q_H} + \epsilon$.

Theorem 2. Non-transferability. *The sDDVS scheme is non-transferable.*

Proof. Suppose receiver B wants to prove the source of the message to a third party D (other than sender A and another receiver C). If B provides all secret messages to D, D can verify whether the signature provided by B is valid or not. However, D knows B can easily simulate the transcript of A's or C's signature. Therefore, no third party can distinguish whether the signer of this signature is produced from the sender or the designated verifier.

Theorem 3. *The sDDVS scheme is a strong designated verifier signature scheme. Formally, the algorithm* SIMU(m, PK_A, SK_B, PK_C) *produces identically distributed signatures as those produced by the algorithm* SIGN(m, SK_A, PK_B, PK_C).

Proof. We show that the transcript simulated by designated verifier B is indistinguishable from those that B receives from sender A, i.e., the following distributions are identical.

Signer A generates the signature $\sigma_A = (y, W)$ as follows:

$$\begin{cases} r \leftarrow_R \mathbb{F}_q^* \\ y = H(m||r) \\ W = (m||r) \oplus \hat{e}(PK_B, PK_C)^{x_A y} \end{cases}$$

Designated verifier B's simulated signature $\sigma_b = (\tilde{y}_b, \tilde{W}_b)$ is as follows:

$$\begin{cases} r' \leftarrow_R \mathbb{F}_q^* \\ \tilde{y}_b = H(m||r') \\ \tilde{W}_b = (m||r') \oplus \hat{e}(PK_A, PK_C)^{x_B y_b} \end{cases}$$

Also, designated verifier C's simulated signature $\sigma_c = (\tilde{y}_c, \tilde{W}_c)$ is as follows:

$$\begin{cases} r'' \leftarrow_R \mathbb{F}_q^* \\ \tilde{y}_c = H(m||r'') \\ \tilde{W}_c = (m||r'') \oplus \hat{e}(PK_A, PK_B)^{x_C y_c} \end{cases}$$

Let $\widehat{\sigma} = (\widehat{y}, \widehat{W})$ be a signature that is randomly chosen in the set of all valid A's signatures intended to B and C, and then we have the following distributions of probabilities:

$$
\begin{aligned}
&Pr[\widehat{\sigma} = \sigma_a] \\
&\quad = Pr\begin{bmatrix} y = H(m||r) = \widehat{y}, \\ W = (m||r) \oplus \hat{e}(PK_A, PK_B)^{x_C y} \\ = (m||r) \oplus \hat{e}(P,P)^{x_A x_B x_C y} = \widehat{W} \end{bmatrix} \\
&\quad = \tfrac{1}{q-1} \\
&\text{and} \\
&Pr[\widehat{\sigma} = \sigma_b] \\
&\quad = Pr\begin{bmatrix} y_b = H(m||r') = \widehat{y}, \\ W = (m||r') \oplus \hat{e}(PK_A, PK_C)^{x_B y_b} \\ = (m||r') \oplus \hat{e}(P,P)^{x_A x_B x_C y_b} = \widehat{W} \end{bmatrix} \\
&\quad = \tfrac{1}{q-1} \\
&Pr[\widehat{\sigma} = \sigma_c] \\
&\quad = Pr\begin{bmatrix} y_c = H(m||r'') = \widehat{y}, \\ W = (m||r'') \oplus \hat{e}(PK_A, PK_B)^{x_C y_c} \\ = (m||r'') \oplus \hat{e}(P,P)^{x_A x_B x_C y_c} = \widehat{W} \end{bmatrix} \\
&\quad = \tfrac{1}{q-1}
\end{aligned}
$$

It is obviously to see that two designated verifiers B and C can simulate transcriptions of the messages and signatures by sender A with the identical distributions of probabilities where no other third party can verify the signature. Hence, the sDDVS scheme has the strong designated verifier property.

5 Deployment and Performance Evaluation

5.1 Deployment

We consider arduino WSNs [20] as the deployment for our scheme. In arduino system, it uses the 64-bit address in data-link layer, and 15-byte will be transmitted for each outgoing packet excluding the data field (In Table 1).

Table 1. Arduino WSNs frame

segment	field	size (Byte)
1	Delimiter	1
2-3	Length field	2
4	Api-ID	1
5	Frame-ID	1
6-13	Destination Address	8
14	Options	1
15	Checksum	1

Each data transmission costs additional energy. Therefore it should be tried to keep the number of necessary data transmissions as low as possible. Normally the data field is about 16-byte to 100-byte in WNSs[1,20]. Obviously, this data field contains data cleartext (if the scheme cannot provide message recovery) and signature when deploys the DVS scheme.

Table 2. Security comparison with related schemes

scheme	security (y: Yes, n: No)				
	unforgeability	non-transferability	message confidentiality	message recovery	support tolerant routing path
[3]	y	y	n	n	y
[7]*	y	y	n	n	n
[9]-(4.7)	y	y	n	n	y
[9]-(4.8)	y	y	n	n	y
[10]	y	y	y	y	n
our scheme	y	y	y	y	y

*[7] cannot support signature randomicity.

5.2 Performance Comparison

In this section, we compare the security, computing efficiency and and communication cost for related schemes. Table 2 compares the security properties of our sDDVS with the previously known sDVS[3,7,9,10] that have small signatures deployed in WSNs. Lal and Verma[9] constructed several two-verifier sDVS schemes derived from general sDVS schemes. We choose two optimized schemes (4.7) and (4.8) in [9] to make a comparison. Chow[3], Lal and Verma[9] and our scheme provide the support for tolerant routing path for WSNs, but [3] and [9] cannot provide message recovery. Lee and Chang[10] supports the message recovery that has the similar security property but supports non-tolerant routing path. Thus, double computation time are needed to implement dual routing paths message verification if we deploy the Lee and Chang scheme in WSNs.

Table 3 concretely compares the computing complex between the previous schemes and ours. We assume that the pairing $\hat{e} : \mathbb{G}_1^2 \rightarrow \mathbb{G}_2$ is implemented using elliptic curves AES 80-bit security. In particular, q is 160 bits, $\mathbb{G}_1$ is 160 bits and $\mathbb{G}_2$ is 1024 bits, respectively. As shown in Table 3, our scheme has a comparable advantage in computation efficiency over those of Chow[3], and Lal and Verma[9](4.7).

In Lee and Chang[10] and our scheme, they need not provide the data cleartext because of the message recovery ability from the signature. But Chow[3], Huang[7], and Lal[9] schemes have to attach the cleartext. So the sizes of signatures plus the cleartexts are $960 + |m|$, $600 + |m|$, and $480 + |m|$ in [3], [7] and [9], respectively, where 960, 600 and 480 are signature size. As shown in[1,20], the data size is approximately 16-byte to 100-byte in WSNs. Fig.2 indicates the

Table 3. Computation performance comparison

scheme	Signature generation				Verification			
	H	A	E	P	H	A	E	P
[3]	2	7	0	1	2	2	2	4
[9]-(4.7)	1	6	2	1	0	0	2	2
[9]-(4.8)	1	5	0	0	1	2	1	2
[10]*	2	6	4	0	1	0	3	1
our scheme	1	0	1	1	0	0	1	1

H: Hash; A: Addition of $\mathbb{G}_1$; E: Exponent of $\mathbb{G}_2$; P: Pairing.

signature data packet size corresponding to the data length increasing from 16 bytes to 100 bytes. Lee and Chang's scheme[10] only supports a single routing path with 640-bit signature size and the cleartext size is constrainted to 160-bit (20-byte) since the message space is defined in an elliptic curve group $\mathbb{G}_1$. In the other schemes[10,9,3], the signature data packets size will increase with the message length increasing. Our scheme has the fix size of signature and the message recovery ability such that protects the confidentiality of message cleartext, the privacy of the sender and the designated recipient, as well as the privacy of cleartext length.

By the comparison with related schemes, our proposed scheme supports strong dual designated verifiers and message recovery as well as cleartext privacy protection, and also has a comparably advantage in computation cost and signature packet size. Our scheme can also protect the signed cleartext privacy. Although our scheme supports at most 864-bit (108-byte) cleartext size, this also has the advantage for deployment in the security requirement in WSNs since the data length is about 16-byte to 100-byte.

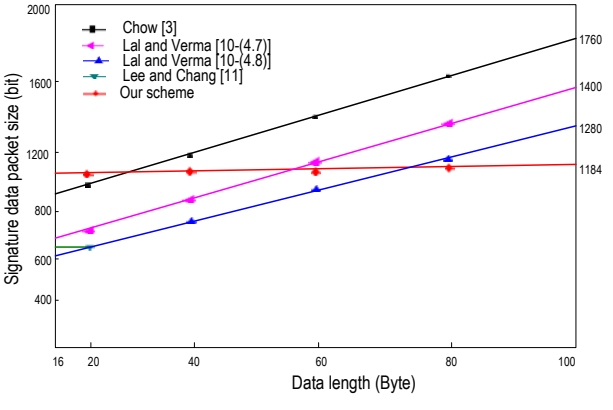


Fig. 2. Signature data packet

6 Conclusion

In this paper, we proposed an efficient strong designated verifier signature scheme with dual verifiers as well as the message recovery. The proposed scheme supports the message authentication in tolerant routing path deployed in wireless sensor networks. The proposed scheme has a comparably higher efficiency and lower communication cost, and thus can be used in resource constrainted environments.

Acknowledgment

The authors grateful thank the anonymous reviewers for their helpful comments and suggestion. This work is supported by the National Natural Science Foundation of China under Grants 60973134, the Natural Science Foundation of Guangdong under Grants 10351806001000000 and 10151064201000028 , the Foundation for Distinguished Young Talents in Higher Education of Guangdong under grants wym09066 , and supported by Grant-in-Aid for JSPS Fellows under grant 22·P10045.

References

1. Aslam, N., Robertson, W., Phillips, W.: Performance analysis of WSN clustering algorithms using discrete power control. IPSI Transactions on Internet Research 5(1), 10–15 (2009)
2. Cao, X., Kou, W., Dang, L., Zhao, B.: IMBAS: Identity-based multi-user broadcast authentication in wireless sensor networks. Computer Communications 31(4), 659–667 (2008)
3. Chow, S.S.M.: Identity-based strong multi-designated verifiers signatures. In: Atzeni, A.S., Lioy, A. (eds.) EuroPKI 2006. LNCS, vol. 4043, pp. 257–259. Springer, Heidelberg (2006)
4. Driessen, B., Poschmann, A., Paar, C.: Comparison of innovative signature algorithms for WSNs. In: The First ACM Conference on Wireless Network Security, pp. 30–35. ACM, New York (2008)
5. Eliana, S., Andreas, P.: A survey on secure multipath routing protocols in WSNs. Computer Networks 54(13), 2215–2238 (2010)
6. Huang, Q., Yang, G., Wong, D.S., Susilo, W.: Identity-based Strong Designated Verifier Signature Revisited. The Journal of Systems and Software 84(1), 120–129 (2011)
7. Huang, X., Susilo, W., Mu, Y., Zhang, F.: Short designated verifier signature scheme and its identity-based variant. International Journal of Network Security 6(1), 82–93 (2008)
8. Kang, B., Boyd, C., Dawson, E.: Identity-based strong designated verifier signature schemes: attacks and new construction. Computer Electrical Engineering 35(5), 49–53 (2009)
9. Lal, S., Verma, V.: Some identity-based strong bi-designated verifiers signature schemes, ePrint/2007/193.pdf (2007)
10. Lee, J.S., Chang, J.H.: Strong designated verifier signature scheme with message recovery. Advanced Communication Technology 1, 801–803 (2007)

11. Lee, J.S., Chang, J.H.: Comment on Saeedniia et al.'s strong designated verifier signature scheme. Computer Standards & Interaces 31(1), 258–260 (2009)
12. Li, B., Batten, M., Doss, R.: Lightweight authentication for recovery in wireless sensor networks. In: 5th Mobile Ad-hoc and Sensor Networks, pp. 465–471 (2009)
13. Li, N., Zhang, N., Das, S.K., Thuraisingham, B.: Privacy preservation in wireless sensor networks: a state-of-the-art survey. Ad Hoc Networks 7(8), 1501–1514 (2009)
14. Laguillaumie, F., Vergnaud, D.: Multi-designated Verifiers Signatures. In: López, J., Qing, S., Okamoto, E. (eds.) ICICS 2004. LNCS, vol. 3269, pp. 495–507. Springer, Heidelberg (2004)
15. Li, Y., Susilo, W., Mu, Y., Pei, D.: Designated Verifier Signature: Definition, Framework and New Constructions. In: Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (eds.) UIC 2007. LNCS, vol. 4611, pp. 1191–1200. Springer, Heidelberg (2007)
16. Lipmaa, H., Wang, G., Bao, F.: Designated Verifier Signature Schemes: Attacks, New Security Notions and a New Construction. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) ICALP 2005. LNCS, vol. 3580, pp. 459–471. Springer, Heidelberg (2005)
17. Liu, J.K., Baek, J., Zhou, J., Yang, Y., Wong, J.W.: Efficient online/offline edentity-based signature for wireless sensor network. International Journal of Information Security 9(4), 287–296 (2010)
18. Lu, R., Lin, X., Zhu, H., Ho, P.-H., Shen, X.: ECPP: effi-cient conditional privacy preservation protocol for secure vehicular communications. In: Infocom 2008, pp. 15–17 (2008)
19. Oscar, D.M., Amparo, F.S., Sierra, J.M.: A light-weight authentication scheme for wireless sensor networks. Ad Hoc Networks (2010), doi:10.1016/j.adhoc.2010.08.020
20. OPEN WSN- An open source multihop sensor network based on arduino (2010), http://www.openwsn.net/index.php
21. Rjaško, M., Stanek, M.: On designated verifier signature schemes. Cryptology ePrint Archive: Report 2010/191 (2010)
22. Saeednia, S., Kramer, S., Markovitch, O.: An efficient strong designated verifier signature scheme. In: ICISC 2003. LNCS, vol. 5984, pp. 40–54. Springer, Heidelberg (2003)
23. Sarma, N., Nandi, S.: A multipath QoS routing with route stability for mobile Ad hoc networks. IETE Technical Review (27), 380–397 (2010)
24. Sun, B., Li, C.C., Wu, K., Xiao, Y.: A lightweight secure protocol for wireless sensor networks. Computer Communications 29(13-14), 2556–2568 (2006)
25. Yang, F.Y., Liao, C.M.: A provably secure and efficient strong designated verifier signature scheme. International Journal of Network Security 10(3), 220–224 (2010)
26. Zubair, Z.A., Baig, Khan, A.I.: A fault-tolerant scheme for detection of DDOS attack pattern in cluster-based wireless sensor networks. In: Sensor and Ad-hoc Network. LNEE, vol. (7), pp. 1–20 (2008)

A Geospatial Analysis on the Potential Value of News Comments in Infectious Disease Surveillance

Kainan Cui[1,2], Zhidong Cao[2], Xiaolong Zheng[2], Daniel Zeng[2], Ke Zeng[1,2], and Min Zheng[1,3]

[1] The School of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China
[3] Department of Communication, Engineering College of Armed Police Force, China
kainan.cui@live.cn

Abstract. With the development of Internet, widely kind of web data have been applied in influenza surveillance and epidemic early warning. However there were less works focusing on the estimation of geospatial distribution of influenza. In order to evaluate the potential power of news comments for geospatial distribution estimation, we choose the H1N1 pandemic in the mainland of China in 2009 as case. After collecting 75878 comments of H1N1 related news from www.sina.com(a famous news site in the mainland of China), we compared the geospatial distribution of comments against surveillance data. The result shows that the comments data share a similar geospatial distribution with the epidemic data(a correlation of 0.848 $p<0.01$), especially with a larger data volume(a correlation of 0.902 $p<0.01$). It suggests that extracting geospatial distribution from comments data for estimation could be an important supplementary method when the surveillance data are incomplete and unreliable.

Keywords: H1N1, infectious diseases, surveillance, geospatial analysis, open source information.

1 Introduction

The prevalence of internet and the threat of emerging infectious disease have driven growing interest in web-based public health surveillance[1, 2]. There are already lots of attempts and applications using different web data sources such as search engine logs[3, 4], news[5], blogs[6, 7], micro-blog[8, 9], wiki[10] and so on. Unfortunately considerable part of the existing works ignored the geographic information contained in the web data and focused on the analysis the temporal dynamic of influenza[11].

The geospatial distribution of infectious diseases is critically important for disease monitoring and control. Although geographic information system (GIS) has already been widely used on visualization and analysis for both epidemic data and web data [12-17], there were little work about evaluating the correlation of web data against

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 85–93, 2011.
© Springer-Verlag Berlin Heidelberg 2011

epidemic data. The uncertain of the correlation weaken the significance of GIS based on web data. In another word, quantitative the correlation of geospatial distribution of web data against epidemic data is the first step for geospatial distribution estimation, and is crucial for the GIS of public health surveillance based on web data.

Fortunately, the development of open source information provides us increasing amount of data with geographic information, which enable us analyze those data from a spatial perspective. News comment is an example of web data with geographic information. The available of geographical information enable us study news comments from a spatial perspective. One of the advantages of news comments is that news comments are straightforward to retrieve. Compared to other web data such as blog and query logs，the news comments are more relevant to certain topic, which means we do not have to apply nature language processing technology for data filtering. When big event happens, the news site will offer a special report containing all related news, which will serve as comments aggregator. In another word, the web sites have done the classification and clustering of comments before we retrieve them. In this paper, we collected comments from www.sina.com and compared the geospatial distribution of news comments against the epidemic data of H1N1 outbreaks in the mainland of China in 2009. The result shows a high positive correlation, which revealed the potential power of news comments for geospatial distribution estimation.

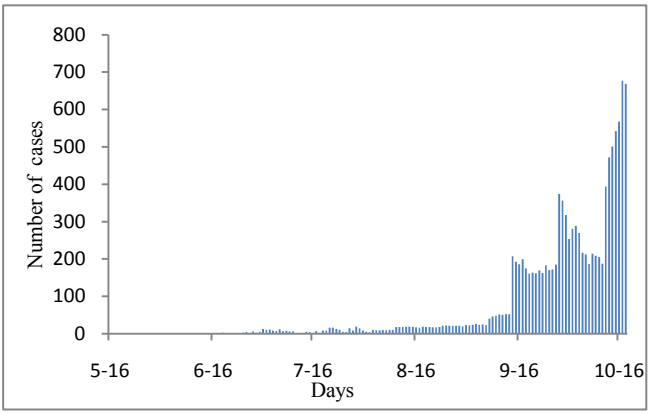
The remainder of this paper is structured as follows. In Section 2, we introduce the epidemiological data, comments data and the analysis method used in our study respectively. In Section 3 we show the result of our geospatial correlation analysis. Based on the result we discuss the possible usages and further work in Section 4. At last we present a conclusion of this paper in Section 5.

2 Data and Methods

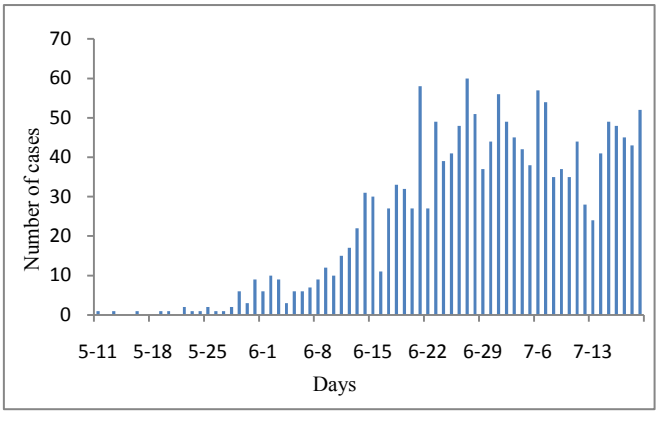
2.1 Epidemic Data

We chose the H1N1 influenza outbreak in the mainland of China in 2009 as case to evaluate the correlation of the geospatial distribution of comments data against epidemic data. The epidemic data used in our study is constituted by two parts. One dataset is obtained through an authority website, which serves 31 province-level regions in the mainland of China (we refer that dataset as CH-ED). The other dataset came from the Beijing Center for Disease Control and Prevention (CDC), which contains the number of reported H1N1 cases in Beijing over 15 geographic divisions (referred as BJ-ED). There are 14 districts and 2 countries in Beijing. For the consideration of reducing the space size difference among regions, we merged Dongcheng District and Xicheng District as city center region in this study.

The CH-ED covers the period from May 11, 2009 to July 9, 2009, and the BJ-ED covers the period from May 16, 2009 to October 18, 2009. Figure 1 show the series plots of the two datasets. Figure 1 (a) is for CH-ED and Figure 1 (b) is for BJ-ED. There is another part of data after October 18, 2009 from CDC, which was estimated based on sample survey. We do not use that part of data with the consideration for accuracy.



(a)



(b)

Fig. 1. Time series plots of number of H1N1 cases, (a) is the reported case number in the mainland of China covering the period from May 11, 2009 to July 9, 2009; (b) is the reported case number in Beijing from May 16, 2009 to October 18, 2009

2.2 Comments Data

We obtained comments dataset from www.sina.com (referred as SINA-CD). www.sina.com is a famous news service offering a full array of Chinese-language news. A special report is a universal entrance for all news related to certain topic in a site, which means we can retrieve all comments of one topic by traversal. The special report for H1N1 is available at http://news.sina.com.cn/z/zhuliugan/index.shtml. We developed a customized crawler for collecting data. After removing duplicated data, there are 75878 comments covering the period from May 27, 2009 to August 11, 2010. Figure 2 shows the time series plot of SINA-CD. Each record in comment dataset contains the following elements: Comment ID, Post time, News title, News URL, Content and Location. Prior to analyze the data, we normalized the comments for duplication. The

comments ID were checked and those records whose id has already appeared were removed. The post time was used for time series plot. After a simple analysis of the comments number at different time interval on a day, we found that users like to replay news in morning. The maximum value appears in the period between 9 am and 10 am and the minimum value appears in 3 am and 4 am. In this study we mainly explore the location information. Other elements such as the content can be used for further study.

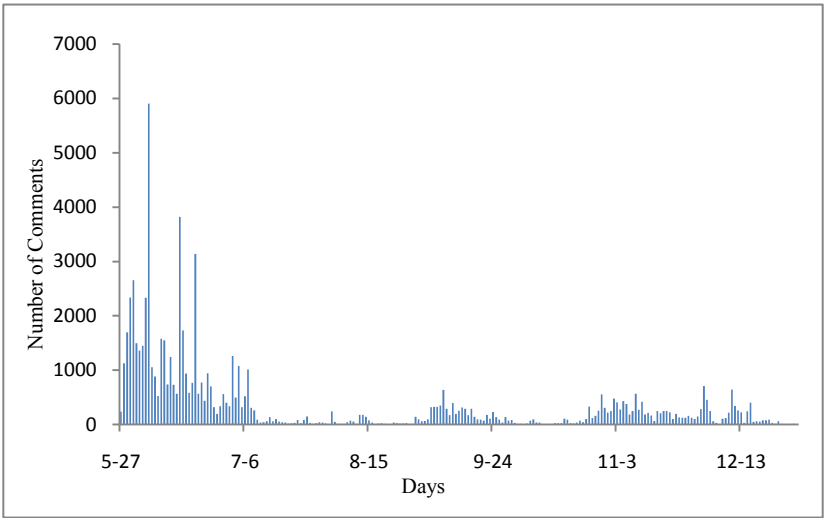


Fig. 2. Time series plots of number of H1N1 related news comments from www.sina.com covering period is from May 27, 2009 to August 11, 2010

The form of location information in comments dataset is string. For most comments the geographical accuracy is in prefecture-level, however for some comments from Beijing, the geographical accuracy is in country-level. We count the comments number in different geographical level by string matching. In particular, we get the comments number for 31 province-level regions in the mainland of China and the comments number for 14 districts and 2 countries in Beijing. After adding the comments number in Dongcheng District and Xicheng District together as comments number in city center region, we get comments number for 15 geographic divisions in Beijing.

2.3 Correlation Analysis

Two geospatial correlation analysis were performed in different geographical levels. We chose the Pearson's Correlation Coefficient to measure the difference between comments and epidemic data in both analyses. In particular, we computed the correlation value of daily counts of comments in SINA-CD against influenza surveillance data in CH-ED and BJ-ED respectively. We defined $P < 0.01$ as statistically significant. After the correlation analyses, several geospatial distribution plots were generated for visual comparison on geospatial patterns of comment data and epidemic data.

The comparison between CH-ED and SINA-CD were referred as SP-C1. This analysis is performed in province-level for the reason that the CH-ED is in

province-level. Although the SINA-CD can reach to prefecture-level, we cannot perform further analyses unless the epidemic data with higher accuracy is available.

The comparison between BJ-ED and SINA-CD were referred as SP-C2. This analysis is performed in prefecture-level. In this case the geographical accuracy of BJ-ED is higher than SINA-CD, which means if we could get comments data in township-level, a more detailed comparison is available. For the inconsistent of data period, we align both start date and end date to the earlier one respectively, which means for SP-C1 the data period is form May 11, 2009 to July 9, 2009, for SP-C2 the data period is form May 16, 2009 to October 18, 2009.

3 Result

3.1 SP-C1: Spatial Correlation of CH-ED against SINA-CD

Figure 3 depicts the geospatial distribution of the number of reported H1N1 case before July 19, 2009 and the number of comments of H1N1 news from

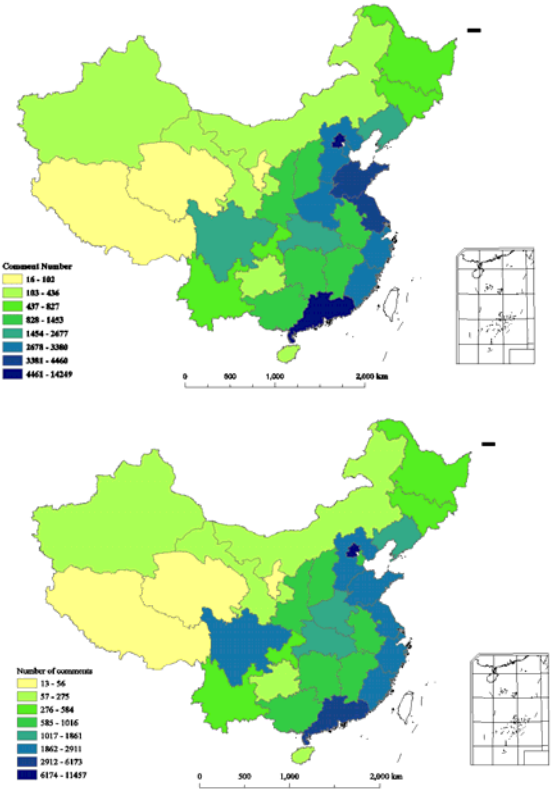


Fig. 3. A comparison of the number of reported H1N1 cases before July 19, 2009 against the number of comments that reply to H1N1 news from www.sina.com. A correlation of 0.848 ($p<0.01$) was obtained over 31 province-level regions.

www.sina.com, which shows that the two datasets have a similar geospatial distribution. From a qualitative perspective, the two datasets have a high correlation of 0.848 ($p<0.01$) which was obtained by correlation analysis over 31 province-level regions.

3.2 SP-C2: Spatial Correlation of BJ-ED against SINA-CD

Figure 4 depicts the geospatial distribution of the number of reported H1N1 cases before October 18, 2009 and the number of comments of H1N1 news from www.sina.com in Beijing, which shows that the two datasets have a similar geospatial distribution too. From a qualitative perspective, the two datasets have a higher correlation of 0.902 ($p<0.01$).

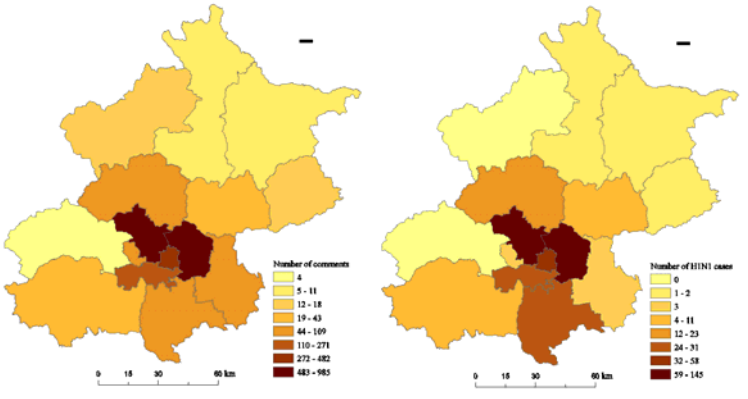


Fig. 4. A comparison of the number of reported H1N1 cases before October 18, 2009 against the number of comments that reply to H1N1 news from www.sina.com. A correlation of 0.902 (*p<0.01)* was obtained over 15 geographic divisions.

Table 1. A comparison for the two spatial correlation analyses

ID	Epidemic data	Number of days covered by epidemic data	Spatial correlation
SP-C1	CH-ED	59	0.848
SP-C2	BJ-ED	155	0.902

3.3 A Comparison of SP-C1 against SP-C2

Table 1 shows the comparison for SP-C1 against SP-C2, the BJ-ED contains data of 155 days has a correlation of 0.902 ($p<0.01$) with SINA-CD over 15 districts, and the CH-ED contains data of 59 days has a correlation of 0.848 ($p<0.01$) with SINA-CD over 31 provinces.

4 Discussion

In this paper we attempt to show the correlation of comments of public health related news with epidemic data in a spatial perspective, we used 2009 H1N1 outbreak in the

mainland of China as case and perform a correlation analysis to qualitative the relationship. In this section we will first discuss the features and possible usage of comments, and then we will discuss the limitation and future work.

4.1 Features and Possible Usage of News Comments

As a kind of web data, news comments have features affect the way we use it for public health surveillance. Take comments from H1N1 news in www.sina.com as example, there are two main features

- Topic related
- Anonymity

Contrary to other web data like blogs, advertisement click records and so on, the most notable feature of news comments is the topic related. Firstly the special reports were arranged by editors. Those editors will ensure all news in a special report was surrounded to certain topic. Secondly the administrator will review the comments and delete those unrelated posts. This feature means we do not need a keyword matching for retrieving data. General speaking, keyword matching result may includes unrelated noisy record that will lead error in subsequent analysis. In order to attract more comments, the www.sina.com allows user post comment without registration, which means we cannot analyze the relationship among users by applying complex network theory.

The correlation analysis result shows that the comments data have a similar geospatial distribution with epidemic data. This is the main finding in this paper. It is worth to note and that the correlation value is higher when we applied the correlation analysis to the epidemic dataset with larger data volume. In particular, the BJ-ED contains data of 155 days has a correlation of 0.902 ($p<0.01$) with SINA-CD over 15 districts, and the CH-ED contains data of 59 days has a correlation of 0.848 ($p<0.01$) with SINA-CD over 31 provinces. Because of the high geospatial correlation, we may use the geospatial patterns extracted from comments data to estimate the epidemic situation in certain area with survey data in other regions, which could reduce the cost for epidemiological investigations, especially when the surveillance data are incomplete and unreliable for target area.

4.2 Limitations and Future Work

Contrary to other web data the main limitation of news comments for public health surveillance is the timeliness. In H1N1 outbreak in the mainland of China, the first comment appears several days after the report case. Although the development of smart phones and cloud computing will allow user post comment more easily and increase the response speed indirectly, the comments are still disadvantage as data source for early warning systems of public health.

Another limitation is the media interest and public interest for a certain infectious disease. H1N1 is a brand-new disease affecting the globe which leads news site's special report and thousands of comments by web users. However for other diseases the web users concerned not very much, there may be not enough comments in the news site for research.

There are two main directions for the future work. The first one is to confirm the result that news comments have a high geospatial correlation against epidemic data with more data, we plan to collect more comments data and build geospatial model for prediction. Through comparing the output of model with epidemic data, we will quantitative the prediction power of news comments. Another direction is to mine more information in the comments, such as applying natural language processing technology to analyze the content of comments. The semantic information contained in comments such as the emotion or opinion about certain disease control measure may help us archiving a more comprehensive understand of influenza outbreak.

5 Conclusion

The high spatial correlation against epidemic data shows the news comments data is a potential data source for influenza surveillance. Besides the containing of geographic information, another important feature of comments is topic related, which decrease the difficulty for data retrieve. When epidemiological investigation is too costly or unable to perform in certain area, we may collect comments data and then perform an estimation based on the spatial patterns extracted from the comments as an alternative.

Acknowledgments. This work was supported by Major National Science and Technology Project (Grant No. 2009ZX10004-315 and 2008ZX10005-013), Chinese Academy of Sciences (Grant No. 2F070C01 and 2F08N03), National Natural Science Foundation of China (Grant No. 60921061, 71025001, 91024030, 90924302, 60621001 and 40901219), and U.S. DHS through Grant 2008-ST-061-BS0002.

References

1. Chen, H., Zeng, D.: AI for Global Disease Surveillance. IEEE Intelligent Systems 24, 66–82 (2009)
2. Brownstein, J.S., Freifeld, C.C., Madoff, L.C.: Digital disease detection–harnessing the Web for public health surveillance. New England Journal of Medicine (2009)
3. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2008)
4. Wilson, K., Brownstein, J.S.: Early detection of disease outbreaks using the Internet. Canadian Medical Association Journal 180, 829 (2009)
5. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K.: BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 24, 2940 (2008)
6. Corley, C.D., Cook, D.J., Mikler, A.R., Singh, K.P.: Text and structural data mining of influenza mentions in web and social media. International Journal of Environmental Research and Public Health 7, 596 (2010)
7. Corley, C.D., Mikler, A.R., Singh, K.P., Cook, D.J.: Monitoring influenza trends through mining social media. In: Conference Monitoring Influenza Trends Through Mining Social Media (Year)

8. Culotta, A.: Detecting influenza outbreaks by analyzing Twitter messages. Arxiv preprint arXiv:1007.4748 (2010)
9. Signorini, A.: Social Web Information Monitoring for Health (2009)
10. Laurent, M.R., Vickers, T.J.: Seeking health information online: does Wikipedia matter? Journal of the American Medical Informatics Association 16, 471–479 (2009)
11. Dailey, L., Watkins, R.E., Plant, A.J.: Timeliness of data sources used for influenza surveillance. Journal of the American Medical Informatics Association 14, 626 (2007)
12. Tatem, A.J., Campiz, N., Gething, P.W., Snow, R.W., Linard, C.: The effects of spatial population dataset choice on estimates of population at risk of disease. Popul. Health Metr. 9, 4 (2011)
13. Zhang, Z., Chen, D., Chen, Y., Liu, W., Wang, L., Zhao, F., Yao, B.: Spatio-temporal data comparisons for global highly pathogenic avian influenza (HPAI) H5N1 outbreaks. Plos One 5, e15314 (2010)
14. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences 106, 21484 (2009)
15. Freifeld, C.C., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. Journal of the American Medical Informatics Association 15, 150 (2008)
16. Cao, Z.D., Zeng, D.J., Wang, Q.Y., Zheng, X.L., Wang, F.Y.: An epidemiological analysis of the Beijing 2008 Hand-Foot-Mouth epidemic. Chinese Science Bulletin 55, 1142–1149 (2010)
17. Cao, Z.D., Zeng, D.J., Zheng, X.L., Wang, Q.Y., Wang, F.Y., Wang, J.F., Wang, X.L.: Spatio-temporal evolution of Beijing 2003 SARS epidemic. Science China Earth Sciences, 1–12 (2010)

Using Spatial Prediction Model to Analyze Driving Forces of the Beijing 2008 HFMD Epidemic

JiaoJiao Wang[1,*], ZhiDong Cao[2], QuanYi Wang[3], XiaoLi Wang[3], and HongBin Song[4]

[1] College of Geoscience and Surveying Engineering, China University of Mining & Technology (Beijing), No. 11, XueYuan Road, HaiDian District
100083 Beijing, China
[2] Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, No.95, ZhongGuanCun East Road, HaiDian District
100190 Beijing, China
[3] Beijing Center for Disease Control and Prevention,
100013 Beijing, China
[4] PLA Institute of Disease Control and Prevention,
100071 Beijing, China
{JiaoJiao Wang,ZhiDong Cao,QuanYi Wang,XiaoLi Wang,
HongBin Song,LNCS}@Springer.com

Abstract. Based on the spatial units of community, village and town in Beijing, the relationship betweent HFMD morbidity and the potential risk factors has been examined. According to the 6 selected risk factors (namely population density, disposable income of urban residents, the number of medical and health institutions, the number of hospital beds, average annual temperature and average annual relative humidity) significantly related to HFMD morbidity, the prediction performance of Classical Linear Regression Model(CLRM) and Spatial Lag Model(SLM) has been compared. The results showed that SLM achieved better effect and R square reached 0.82. It was showed that spatial effect played the crucial role in the HFMD morbidity prediction and its contribution attained 88%. However, CLRM showed low prediction accuracy and bias estimation. It was demonstrated that including spatial effect item into CLRM could greatly improve the performance of HFMD morbidity prediciton model.

Keywords: Hand-foot-mouth disease(HFMD) morbidity, risk factor, Beijing, Classical Linear Regression Model, Spatial Lag Model

1 Introduction

Hand-foot-mouth disease (HFMD) prevalence has been reported in a majority of the countries and regions all over the world. HFMD is caused by the intestinal virus and

* This work was supported by National Natural Science Fundation of China (Grant No. 91024030, 90924302, 40901219 and 71050001), Major National Science and Technology Project (Grant No. 2009ZX10004-315 and 2008ZX10005-013) and Chinese Academy of Sciences (Grant No. 2F070C01 and 2F08N03).

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 94–100, 2011.
© Springer-Verlag Berlin Heidelberg 2011

easily infectious among the crowds. HFMD cases are mostly made of the infants and the children. Adults HFMD cases unusually appeared, but they could take and transmit the virus. HFMD occurred in China in 1981 and became prevalence in most of the provinces all over the country. Recently, the scale of HFMD outbreak has gradually increased and greatly threatened the health security of the nation. In 2008, the administration of HFMD was strengthened further[1].

So far, most research on HFMD has been focused on the fields such as molecular biology[2], clinical medicine[3], pathogeny[4] and epidemiology[5]. Although some scholars investigated the factors that could cause HFMD prevalence from the view of physical, social, economic and humanity, the epidemiology analysis was carried out mostly based on classical statistics. At the end of 1960s, spatial correlation was firstly used in the research on pathogeny analysis[6]. In 1973, Cliff and Ord[7] put forward the concept and framework of spatial autocorrelation. Since then, a series of exploration and study on the theory and methodology about spatial data analysis has come forth[8-10]. Over the recent years, the technology of Geographical Information System(GIS) in which spatial analysis is the key point has developed very rapidly and been increasingly used in the field of public health[11-14].

2 Data

2.1 HFMD Morbidity

HFMD prevalence can be measured by HFMD morbidity which means the ratio of the number of HFMD cases in a year to the number of total population in the same region.

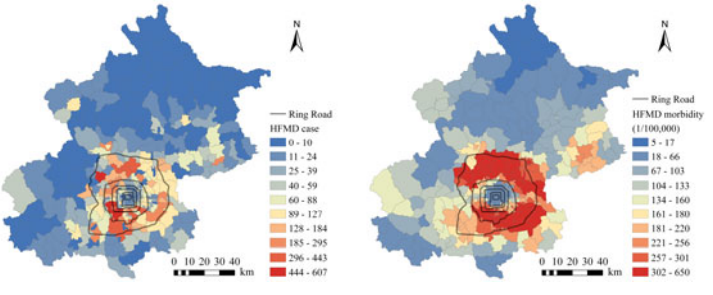


Fig. 1. Spatial distribution of 2008 Beijing HFMD case (left) and HFMD morbidity (right)

Beijing Center for Disease Control and Prevention (CDC) received 18,445 observed HFMD cases covering 309 towns and villages all over the 18 administrative counties in Beijing City, with the period from October 12, 2007 to October 31, 2008[15].

2.2 Risk Factors

There are many potential risk factors to influence HFMD epidemic, such as environmental, demographic, socioeconomic and human factors. The preceding study

has indicated that health service could play an important role in HFMD prevalence[15]. According to the data obtained, we divided the 16 potential factors into the 6 labels as Table 1 shows.

Table 1. Potential risk factors of 2008 Beijing HFMD prevalence. The calculation of population density was detailed in literature [15]. The climate materials were from meteorological observation stations under China Meteorological Administration including 137 stations located in Beijing City and the surrounding areas 250 km away from Beijing (detailed in literature [16]).The data source of the other factors was from Beijing Regional Statistical Yearbook 2009.

Lable	Factor	Note
Urbanization	agricultural land cultured land constructive land unused land	the ratio of the area of every kind of land to the total area of the land (%)
Socioeconomic	GDP per capita	the ratio of total regional GDP to resident population at the end of a year (yuan)
	unit GDP energy consumption	the ratio of total GDP to comprehensive energy consumption (tce/million)
	disposable income of urban residents	the ratio of total income of the sampled residents to the number of those residents (yuan/per capita)
Health service	healthcare orgnizations	the ratio of the number of healthcare organizations to that of the residents at the end of a year (1/10,000)
	medical practitioners	the ratio of the number of medical practitioners to that of the residents at the end of a year (1/1,000)
	registered nurses	the ratio of the number of registered nurses to that of the residents at the end of a year (1/1,000)
	beds in hospital	the ratio of the number of beds in hospital to that of the residents at the end of a year (1/1,000)
Environmental	green coverage	the ratio of the area of forest land, shrubs and all around tree to the total area of the land (%)
Demographic	kindergarten	the ratio of kindergartens to the number of the children enrolled in kindergartens (1/1,000)
	population density	the ratio of the residents at the end of a year to the total area of the land (persons/square km)
Climate	annual average temperature	the ratio of the sum of daily temperature in a year to the number of days in a year (℃)
	annual average relative humidity	the ratio of the sum of daily relative humidity in a year to the number of days in a year (%)

In order to select the factors which would be relatively significantly associated with HFMD morbidity, Pearson correlation analysis and Stepwise Regression method was adopted to remove the redundant factors. Finally we got the 6 factors as follows, disposable income of urban residents (Dis Inc), healthcare orgnizations (Hea Org), beds in hospital (Hos Bed), population density (Pop Den), annual average temperature (Temp) and annual average relative humidity (Rel Hum). In order to preserve the same spatial scale of the different variables, Kriging interpolation method [16,17] was adopted to realize spatial data scale conversion.

3 Methodology

Classical Linear Regression Model (CLRM)[17] assumes explanatory variable and dependent variable is independent from each other. CLRM is given by,

$$Y=a+bX+e . \quad (1)$$

Where a and b are parameters, e is a vector of i.i.d. error terms. Y is a vetor of observations on the dependent variable. X is a matrix of observations on the explanatory variable.

Unlike CLRM approach which considers OLS (Ordinary Least Squares) Estimation, Spatial regression model usually adopts Maximum Likelihood Estimation. Spatial Error Model (SEM) includes a spatial autoregressive error term and Spatial Lag Model (SLM) includes a spatially lagged dependent variable[18]. SEM is given by,

$$Y= cX+e . \quad (2)$$

$$e=dWe+u . \quad (3)$$

Where W is the spatial weights matrix, e is a vector of spatially autocorrelated error terms, u is a vector of i.i.d. errors, c and d are parameters.

SLM uses eigenvalues of the weights matrix and is well suited to the estimation in situations with very large data sets. Formally, SLM is given by,

$$Y=fWY+gX+e . \quad (4)$$

Where WY is a spatially lagged dependent variable for weights matrix W, f and g are parameters.

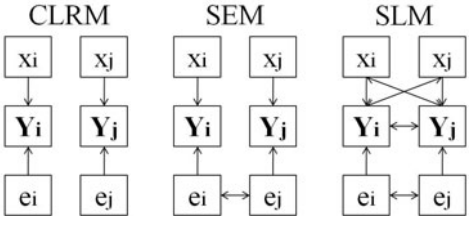


Fig. 2. Relationships among variables in CLRM, SEM and SLM

Lagrange Multiplier test statistics are used to suggest the alternative specification, where LM-Lag and Robust LM-Lag pertain to the spatial lag model as the alternative and LM-Error and Robust LM-Error refer to the spatial error model as the alternative. The important issue is to only consider the Robust versions of LM statistics when the standard versions (LM-Lag or LM-Error) are significant. The spatial regression specification is summarized as Fig.3. Moran's I statistic is used to detect spatial autocorrelation and misspecifications in the model.

There are three classic specification tests[18-20] on the spatial autoregressive coefficient, where LM-Lag test is based on OLS residuals, the Likelihood Ratio (LR)

test is one of the three comparing the null model (the classic regression specification) to the alternative SLM, and the Wald test is the square of the asymptotic t-value (or, z-value).

Log-Likelihood, AIC (Akaike information criterion) and SC (Schwarz criterion) are the proper measures of fit, which are based on an assumption of multivariate normality and the corresponding likelihood function for the standard regression model. The higher the log-likelihood, the better the fit. For the information criteria, the direction is opposite, and the lower the measure, the better the fit.

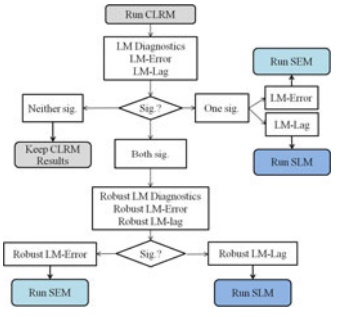


Fig. 3. Spatial regression decision procedure[18], where sig. means significant

4 Results

Even though LM-Lag (z=219.3059, p=0) and LM-Error (z=183.4574, p=0) statistics were both highly significant, Robust LM-Lag statistic (z=36.1088, p=0) was highly significant while Robust LM-Error statistic (z=0.2603, p=0.6099) was clearly not, SLM was finally chosen to predict HFMD morbidity according to spatial regression decision process and achieved high R square (0.82). For CLRM, Moran's I statistic for residual spatial autocorrelation was positive and highly significant (0.4787, $p<0.01$), while for SLM the statistic was not significant (0.0297, $p>0.05$).There was still multicollinearity, abnormality of error distribution, heteroscedasticity and low R square (0.50) caused by CLRM.

Log-Likelihood (-1696.71) obtained by SLM was higher than that (-1820.62) by CLRM, AIC (3409.42) and SC (3439.29) obtained by SLM was lower than that (AIC=3655.24, SC=3681.37) by CLRM, which proved the better fitness of SLM.

Besides, the relationship among the three classic tests was as below, Wald test (1027.1350) > LR (247.8147) > LM-Lag (219.3059), indicating that LM Estimation was significant and the prediction got by SLM was reasonably effective. In SLM equation obtained,

$$Y=0.879WY+175.619-3.675X_1-0.001X_2+0.004X_3-6.516X_4-3.665X_5-0.374X_6$$

Where Y was HFMD morbidity, $X_{1\text{-}6}$ were the selected 6 risk factors, the parameter associated with the spatial lag reached 0.879 which indicated that communities would be expected to have higher HFMD morbidity if on average, their neighbors had high HFMD morbidity. The observation and prediction map of HFMD morbidity showed as Fig. 4.

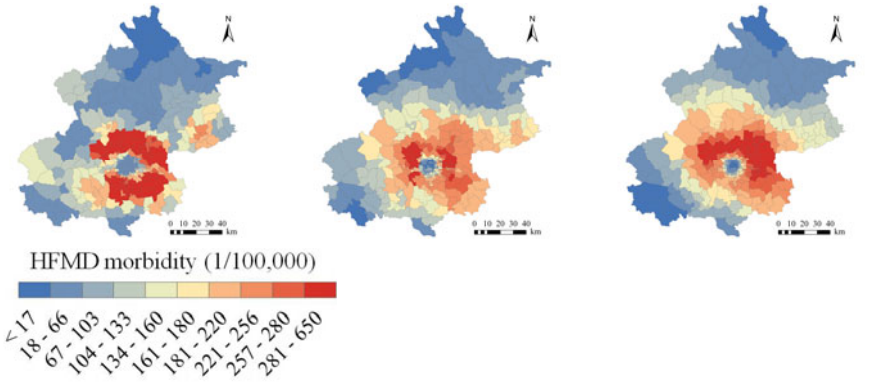


Fig. 4. Comparison of spatial distribution of the Beijing 2008 HFMD morbidity. The left plot is observed value, the middle plot is predicted value by CLRM and the right plot by SLM.

5 Discussion

The results above suggested that spatial effect included in SLM could greatly improve the performance of prediction model. Spatial correlation should be especially considered in epidemiology study. There are more and more applications indicating the prevalence of many infectious diseases such as SARS, H1N1, Bird Flu, HFMD and AIDS presented spatial correlation. Ignoring spatial effect would result in bad prediction precision and bias estimation which would both mislead the control and prevention of epidemic.

To some extent, even though Kriging interpolation method adopted in data preprocessing caused error and uncertainty in the experiment, the reliability of the results would not be pulled down. Generally speaking, the limitation will not weaken the precious value of the study, which well contributed to the application of spatial prediction model in the field of public health and infectious disease informatics.

References

1. He, Y.: Hand-foot-month disease (HFMD) prevention and control manual (2008 edition). J. China Medical News 23, 20–22 (2008)
2. Chen, S.C., Chang, H.L., Yan, T.R., et al.: An eight-year study of epidemiologic features of enterovirus 71 infection in Taiwan. J. The American Society of Tropical Medicine and Hygiene 77, 188–191 (2007)
3. Tseng, F.C., Huang, H.C., Chi, C.Y., et al.: Epidemiological survey of enterovirus infections occurring in Taiwan between 2000 and 2005: analysis of sentinel physician surveillance data. Journal of Medical Virology 79, 1850–1860 (2007)
4. Chang, G.-h., Lin, L., Luo, Y.-j., et al.: Sequence analysis of six enterovirus 71 strains with different virulences in humans. J. Virus Research 151, 66–73 (2010)
5. Ali, M., Emch, M., Yunus, M., et al.: Modeling spatial heterogeneity of disease risk and evaluation of the impact of vaccination. J. Vaccine 27, 3724–3729 (2009)

6. Mantel, N.: The detection of disease clustering and a generalized regression approach. J. Cancer Research 27, 209–220 (1967)
7. Cliff, A.D., Ord, J.K.: Spatial autocorrelation. Pioneer, London (1973)
8. Anselin, L.: Spatial econometrics: methods and models. Kluwer Academic Publishers, Dordrecht (1988)
9. Griffith, D.A.: Advanced spatial statistics. Kluwer Academic Publishers, Dordrecht (1988)
10. Haining, R.P.: Spatial data analysis in the social and environmental sciences. Cambridge University Press, London (1993)
11. Crighton, E.J., Elliott, S.J., Moineddin, R., et al.: A spatial analysis of the determinants of pneumonia and influenza hospitalizations in Ontario (1992-2001). J. Social Science & Medicine 64, 1636–1650 (2007)
12. Zeng, D., Yan, P., Li, S.: Spatial regression-based environmental analysis in infectious disease informatics. In: Zeng, D., Chen, H., Rolka, H., Lober, B. (eds.) BioSecure 2008. LNCS (LNBI), vol. 5354, pp. 175–181. Springer, Heidelberg (2008)
13. Wu, P.-C., Lay, J.-G., Guo, H.-R., et al.: Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical Taiwan. J. Science of The Total Environment 407, 2224–2233 (2009)
14. Cao, Z., Zeng, D., Zheng, X., et al.: Spatio-Temporal Evolution of Beijing 2003 SARS Epidemic. J. Science in China Series D: Earth Sciences 53, 1017–1028 (2010)
15. Cao, Z., Zeng, D., Wang, Q.: An epidemiological analysis of the Beijing 2008 Hand-Foot-Mouth epidemic. J. Chinese Sci. Bull. 55, 1142–1149 (2010)
16. Cao, Z., Zeng, D., Wang, F.: Weather Conditions and Spatio-Temporal Spreading Risk of the Beijing 2009 Influenza A (H1N1) Epidemic. J. Science & Technology 28, 26–32 (2010)
17. Gujarati Damodar, N.: Basic Econometrics. McGraw-Hill, New York (1995)
18. Anselin, L.: Exploring Spatial Data with Geoda: A Workbook (2005)
19. Anselin, L., Florax, R. (eds.): Small sample properties of tests for spatial dependence in regression models: Some further results. Springer, Berlin (1995)
20. Anselin, L., Bera, A.K., Florax, R., et al.: Simple diagnostic tests for spatial dependence. J. Regional Science and Urban Economics 26, 77–104 (1996)

An Online Real-Time System to Detect Risk for Infectious Diseases and Provide Early Alert*

Liang Fang[1,2] and ZhiDong Cao[1]

[1] Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, No.95, ZhongGuanCun East Road, HaiDian District 100190 Beijing, China
[2] University of Science and Technolgoy of China, No.96, Jinzhai Road, Hefei, Anhui Province 230026, China

Abstract. The purpose of this research was to design and develop an online real-time system to detect risk for infectious diseases and provide an early alert to improve the ability to deal with epidemics. The system is composed of report submission module for collecting data through web form, a report reception module for delivering real-time epidemic intelligence on emerging infectious diseases for a diverse audience, and an epidemic early alert module suggests an approach for detecting an epidemic outbreak at an early stage through time and spatial analysis. Advanced data analysis on the data may detect predominant numbers of incidences, indicating a possible outbreak. This gives the health authorities the possibilities to take actions to limit the outbreak and its consequences for all the inhabitants in an affected area. In field experiments, the system has been proven to be of substantial value in visualizing the epidemic data and perceiving the infectious diseases out-break.

Keywords: ArcGIS, real-time, early alert, infectious diseases, Mashup, geographical information system.

1 Introduction

The epidemics are a major public health concern, such as seasonal influenza epidemics caused tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year [1]. So rapid identifying an infectious disease outbreak is critical, both for effective initiation of public health intervention measures and timely alerting of government agencies and general public.

The Internet, however, is revolutionizing how epidemic intelligence is gathered, it offers fast way to detect the epidemics. A great many of real-time information about infectious disease outbreak is found in various forms of Web-based data stream [2]. These range from official public health reporting to informal news coverage to blogs

* This work was supported by Ministry of Health of the People's Republic of China(Grant No. 2009ZX10004-315 and 2008ZX10005-013), Chinese Academy of Sciences(Grant No. 2F070C01 and 2F08N03), and National Natural Science Foundation of China(Grant No. 71050001, 90924302, 60621001, 40901219 and 91024030).

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 101–107, 2011.
© Springer-Verlag Berlin Heidelberg 2011

and chat rooms [3-5]. During the global spread of pandemic influenza A(H1N1) in the 2009, researchers found that, on average from country to country, there was a 12-day lag period [6]. So building an online, web-based, real-time surveillance system is significant.

There are many Web-based surveillance system worldwide. One is GPHIN(Global Public Health Intelligence Network), although automation is a key component, but still need expert people to analyze [7]. Another is ProMEM-mail, which was founded in 1994 and has grown into a large, publicly available reporting system, with more than 45000 subscribers in 188 countries [8]. ProMEM uses the Internet to disseminate information on outbreaks by e-mail.

More recently, there are some website combining the ArcGIS and epidemic disease surveillance. One is HealthMap, which is an openly available public health intelligence system that uses data from disparate sources to produce a global view of ongoing infectious disease threats [9-10], which was mostly used in crisis emergency.

From what has been mentioned above, we know that some systems are not totally automated systems, and some can only be used in crisis emergency, and some are not map interactive system. More importantly, they don't suitable for the requirement for monitoring the epidemics breakout in China, because their data all came from aboard, seldom came from China. So we need our own epidemic monitor system to do early alert in China.

To solve the issues mentioned above, we established an online real-time system to detect risk for infectious diseases and provide early alert. The system utilized ArcGIS and Mashup technology, and the statistic method to do the spatial-temporal analysis, which is a freely accessible, online real-time system that monitors, organizes, integrates, visualizes and disseminates online information about emerging disease. so the system can be greatly used in online perceiving and early alert, and help us to take effective and corrective measures when outbreaks of infectious disease.

2 Material and Method

2.1 Multiple Data Sources

In order to satisfy the system demand for accuracy and real time, the data collection is essential. The Internet, however, is revolutionizing how epidemic intelligence is gathered, it offers fast way to detect the epidemics.

The system integrates outbreak data from multiple electronic sources, including official public health reporting(e.g., hospital, school, public health agency.), web form data(users submit report through web form), email and mobile messages, and online web-based data(e.g., news from Baidu, Google, and some large portal.) which is fetched using a web crawler. The free available data sources are came from two ways, one is from the web submission, the other is from Internet, both are non-structured data, we should extract useful data, then store these data in the database after format. Fig. 1 shows the structured data in database. In this system, we mainly fetch data from Sina of China.

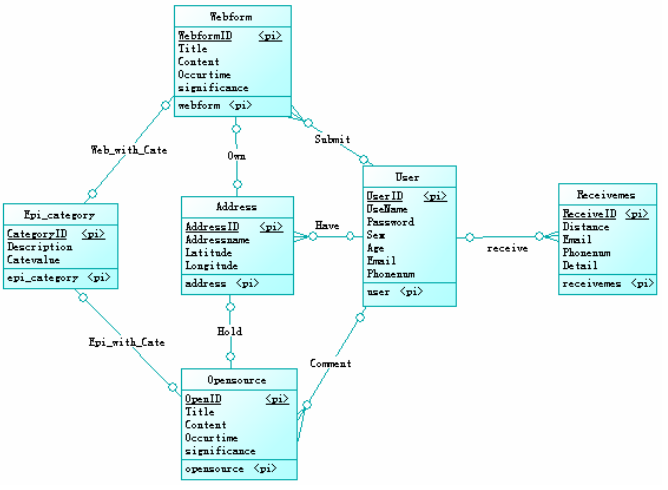


Fig. 1. The conceptual model sketch map of structured multiple data sources in database

2.2 Spatial-temporal Analysis

Time analysis mainly utilizes the method of Exploratory Data Analysis which is an approach using visualization technique to observe the information in each side [11]. The system adopts visualization technique of the histogram and timing diagram. As time analysis we use the statistical analysis methods to get the temporal distribution of epidemic diseases. We use Java to implement the simple statistic algorithm, and get the histogram and timing diagram.

In spatial analysis, we use the Spatial Description Statistics and Spatial Cluster Analysis [12] and the ArcGIS technology which is good at presenting the distribution of the disease outbreaks. The ArcGIS technology is implemented through Google map [13].

2.3 Perceiving the Risk of Epidemic

Perceiving the risk of epidemic is constructed for evaluating the contemporary epidemics situation and predicting the epidemic diseases. The data used in epidemic early alert module were took from an epidemiological survey of 18,445 HFMD-infected persons in Beijing in 2008, conducted by the Beijing Centers for Diseases Control and Prevention (Beijing CDC). The survey instruments covered information such as patients' sex, age, home address, onset date, and so on. The onset dates range from December 24, 2007 to December 31, 2008. Fig.2 shows the interface of the epidemic early alert interface of the system.

The right of the picture presents the histogram that shows the dates range from January 1, 2008, to December 31, 2008. We can figure that the April is the most severe months for Hand, foot and mouth disease out-break.

The left of the picture was demonstrated the 18448 epidemic record tagged on the map. From the map, we can clearly know that the Caoyang, Fengtai, and Haidian district were the most susceptible to infectious disease.

The epidemic early alert module provide an early warning for an epidemic. From research, China is doing epidemic alert through the connection of hospital online, then the Ministry of Health will alert when the people for some disease is reach to a threshold. However this method is time delaying. Our system gets real-time information from Internet, it is able to combine data from all sources and put it into one place, thus the system can provide a earlier alert than traditional epidemic alert. The system will remind people or government when the number of some infectious disease have increased at a fast speed. So the system can do early alert base on the real-time online information and it will be more faster than traditional methods [14].

Fig. 2. The epidemic early alert module interface of the system

2.4 Mashup Technique

In Web development, a Mashup [15] is a Web page or application that uses and combines data, presentation or functionality from two or more sources to create new services. The term implies easy, fast integration, frequently using open APIs and data sources to produce enriched results that were not necessarily the original reason for producing the raw data.

The main characteristics of the Mashup are combination, visualization, and aggregation. It is important to make existing data more useful, moreover for personal and professional use.

We have implemented the system using the map Mashup technology, it is a way to combine the local data and the Google map APIs and gives a way to make various of epidemic influenza data visualized.

2.5 System Implementation Technology

An online real-time system to detect risk for infectious diseases and provide early alert was developed base on My Eclipse8.0, and Mysql was used to construct the database, and Tomcat was used to web service. The system used the J2EE application

model and MVC architecture technology. The system was installed and tested on both windows and Linux operating system. The system is a very user-friendly interface, allowing users to utilize the system through the interface without having to read the help documentation. Fig.3 is the system architecture diagram. The free available data sources are came from two ways, one is from the web submission, the other is from Internet, both are non-structured data, we should extract useful data, then store these data in the database after format. The collected data is currently being stored in a relational database implemented with MySQL Community Server 5.5. The data contain author, source, address, score information. The frontend consists of a web interface that will provide a dynamic graphic environment for the user to explore the data.

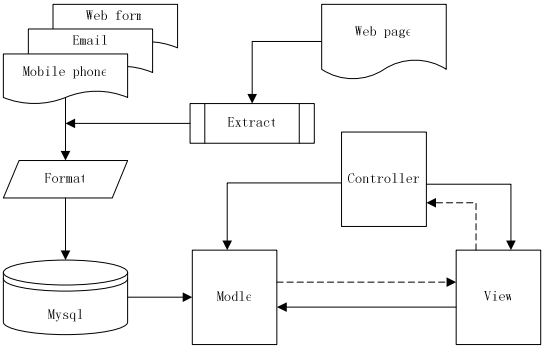


Fig. 3. System architecture diagram

3 Result

The system provides the function we need. Fig.4 is the system interface. The initial interface of the system presents the H1N1 data we get from Sina range from January1, 2011 to March 3, 2011. Firstly we parse the address information, and then tag the disease information on the map according to address. From the map, we can clearly see the H1N1 spatial distribution on the map. The map below shows the detail information about the epidemic diseases, which contain the time, address, and title information. We can track the source of each information. The presented electronic map is Google map which is help to visualize the epidemic data collected from various sources. the map shows the H1N1 outbreak in China, the blue makers on the map are detected of H1N1 diseases, and the number in the markers presents the number of the disease outbreaks, then we can figure out the most serious areas suffered from the epidemic diseases. The right panes shows the various of infectious diseases can be detected by the system, such as flu influenza, H1N1, HFMD disease and so on. The system can visualize the selected disease data after chose a type of disease. We can trace the message from the detail information about infectious disease which contain title, location and time information. The message from map correspond to the detail data showed below of the system, users can identify the map data from the report list, and report list data from the map. The data can sort from time or location, then we can

figure out what time or which place is the most serious place where suffered from infectious diseases.

The system consists of three functional modules, which are report submission module, report reception module and epidemic early alert module. The report submission module is provided user with submission reports of infectious diseases, through people share their message, the others could know the trends of diseases, about area what where the diseases occur and time when the disease detected. The reception report module facilitate the ordinary people and office agency to know the real time information about epidemic diseases. User only provide the email address, telephone, and the location they care, will receive the real-time messages. The epidemic alert module is for surveillance of the epidemic diseases on real time.

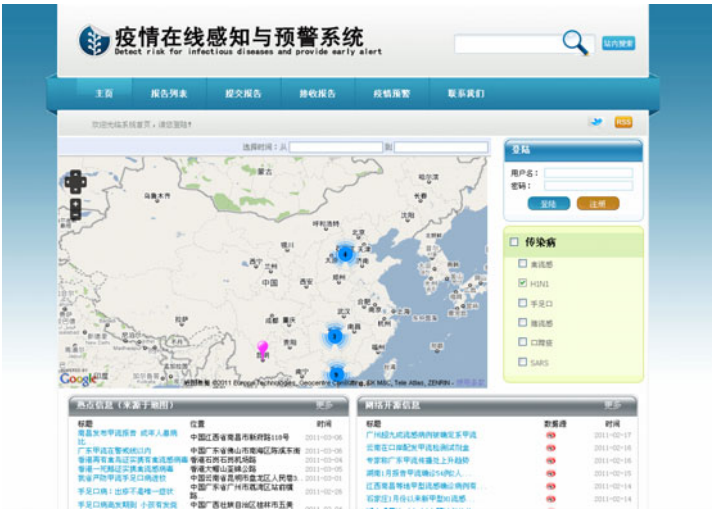


Fig. 4. The interface of the system

4 Discussion

An online real-time system to detect risk for infectious diseases and provide early alert is Web-base, real-time system using the technology of ArcGIS and Mashup designed to collect and display information about new outbreaks according to geographic location, time, and infectious agent. The system has multiple data source, including web extracted data and user submission data, the data are aggregated by source, disease, and geographic location and then overlaid on an interactive map for user-friendly access to the original reports. Through spatial-temporal analysis of the online data, the system can do early alert. The system delivers real-time epidemic intelligence on emerging infectious diseases to diverse audience, from public health officials to inter-national travelers. Most importantly, these technologies may provide important benefits to outbreak control at local, national level, ultimately reducing the healthy consequences of these outbreaks.

As the advent of web2.0, Microblogging(e.g., Twitter) and social network(e.g., Facebook) have involved in our life. As a next step, we will detect infectious diseases from these data which may present early evidence of an infectious disease. However the data in website increasing at tremendous speed, so cloud computing becomes more and more important. In future we will implement the system in a new way, which combine the technology of the cloud computing and the artificial intelligence that can be used in process the intelligence of epidemic diseases [16]. Thus our system will become more faster and smarter.

References

1. World Health Organization. Influenza fact sheet (2003), http://www.who.int/mediacentre/factsheets/2003/fs211/en/
2. Brownstein, J.S., Freifeld, C.C., Reis, B.Y., Mandl, K.D.: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med. 5(7), e151 (2008)
3. Grein, T.W., Kamara, K.B., Rodier, G., et al.: Rumors of disease in the global village: outbreak verification. Emerg. Infect. Dis. 6, 97–102 (2000)
4. Heymann, D.L., Rodier, G.R.: Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect. Dis. 1, 345–353 (2001)
5. M'Ikanatha, N.M., Rohn, D.D., Robertson, C., et al.: Use of the Internet to enhance infectious disease surveillance and outbreak investigation. Biosecur Bioterror 4, 293–300 (2006)
6. Brownstein, J.S., Freifeld, C.C., Chan, E.H., et al.: Information Technology and Global Surveillance of Cases of 2009 H1N1 Influenza. New England Journal of Medicine 362, 1731–1735 (2010)
7. Mykhalovskiy, E., Weir, L.: The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health. Can J. Public Health 97, 42–44 (2006)
8. Brownstein, J.S., Freifeld, C.C., Lawrence, C.: Madoff. Digital Disease Detection—Harnessing the Web for Public Health Surveillance. New England Journal of Medicine 360, 2153–2157 (2009)
9. HealthMap, http://healthmap.org
10. Ushahidi, http://www.ushahidi.com
11. Tukey, J.W.: Exploratory Data Analysis. Addision-Wesley Publishing Company, London (1997)
12. Kaufan, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York (1990)
13. Google map API, http://code.google.com/intl/zh-N/apis/maps/documentation/javascript/v2
14. Zhidong, C., Zeng, D.J., Wang, Q.Y., Zheng, X.L., Wang, F.W.: An Epidemiological Analysis of the Beijing 2008 Hand-Foot-Mouth Epidemic. Chinese Science Bulletin 55(12), 1142–1149 (2010)
15. Enterprise Mashups: The New Face of Your SOA, http://soa.sys-con.com
16. Zheng, X., Zeng, D., Li, H., Wang, F.: Analyzing Open-source Software Systems as Complex Networks. Physica A: Statistical Mechanics and its Applications 387(24), 6190–6200 (2008)

The Impact of Community Structure of Social Contact Network on Epidemic Outbreak and Effectiveness of Non-pharmaceutical Interventions

Youzhong Wang[1], Daniel Zeng[1,2], Zhidong Cao[1], Yong Wang[3], Hongbin Song[3], and Xiaolong Zheng[1]

[1] The Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] MIS Department, The University of Arizona, Tucson, Arizona, U.S.A.
[3] Institute of Disease Control and Prevention, Academy of Military Medical Sciences, Beijing, China
{youzhong.wang,dajun.zeng,zhidong.cao}@ia.ac.cn, ywang7508@gmail.com, hongbinsong@263.com, xiaolong.zheng@ia.ac.cn

Abstract. The topology structure of social contacts network has a big impact on dynamic patterns of epidemic spreading and effectiveness of non-pharmaceutical interventions. Corresponding to individuals' behavioral or functional units, people are commonly organized in small communities, meaning that most of social contacts networks tend to display community structure property. Through empirical investigation and Monte-Carlo simulation on a big H1N1 outbreak in a Chinese university campus, this paper explores the impact of community structure property of social contacts network on epidemic spreading and effectiveness of interventions. A stochastic model based on social contacts networks among students is constructed to simulate this outbreak, revealing that epidemic outbreaks commonly occur in local community. Moreover, effectiveness of three quarantine-based interventions is quantitatively studied by our proposed model, finding that community structure of social networks determines the effects these measures.

Keywords: community structure, social contact network, epidemic outbreak, non-pharmaceutical Interventions, H1N1.

1 Introduction

For many kinds of communicable diseases, contagion is spread among the population mainly through close contact such as talking and physical touch. As intensity and frequency of close contact are determined primarily by social relationships between individuals, studying patient social networks has obvious advantage to understand which reasons determined the epidemic outbreak and trend.

Over the last few years, there have been an increasing number of efforts to build variety of theoretical models that couple the classical compartmental models such as

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 108–120, 2011.
© Springer-Verlag Berlin Heidelberg 2011

susceptible-infectious-recovery (SIR) model with complex social network methods. These combined models borrow the conception from compartmental models that individuals are divided into different compartments according to their states and their states are transformed with the time [1]. On the other hand, these methods overcome the limitation that assumes populations are homogeneously mixed in standard compartmental models. Recent advances in social network study provide useful tools to simulate and analyze the epidemic dynamics among the real population in which contacts between individuals are not simply selected randomly but possessing many of complex properties [2-4].

One of important properties that possessed by many real social contacts network is community structure. Community structure, also known as hierarchical organization or modularity, refers to the fact that a network can be divided into sub-groups. Nodes within these sub-groups are densely connected while the connection between sub-groups is much sparser [5]. In real world, people are commonly organized in many of small communities corresponding to their behavioral or functional units such as residence district, work place, and university campus. It is found that the close and frequent contacts among population in these social groups make these local areas be highest-risk environments for outbreaks of infectious diseases, especially for respiratory infectious diseases such as H1N1 virus [6, 7]. How to protect people from being infected in such communities has been a public health priority to many countries.

On the other hand, the community structure of social contact networks determines the effectiveness of non-pharmaceutical intervention methods to slow down infection spreading in local communities. Non-pharmaceutical interventions have the added benefit of lessening the worry that pharmaceutical interventions may induce viral drug resistance. As such, from the very beginning of the 2009 H1N1 pandemic, many countries, including China, had adopted control strategies such as quarantining travelers from foreign countries, closing the schools, among others. These control strategies gain great benefits to cut down many of social contacts among population in communities and thus prevent the epidemic outbreak. In effect, this strategy, i.e., quarantining close contacts of the infected persons, was proved to be effective against SARS infections in 2003 [8, 9]. Obviously, it would cost more when we take the strategies earlier or more stringent. A quantitative evaluation of the effectiveness of different control measures is essential for us to choose when and how to conduct intervention methods. As most of contacts are within local social groups, it is essential to study the impact of community structure on the effectiveness of control strategies.

In this paper, we conduct a case study of a large 2009 pandemic influenza A (H1N1) outbreak to characterize transmission patterns of H1N1 virus in a school campus that possesses the property of community structure. According to the social contact information from a detailed epidemiologic investigation of the outbreak, we construct a hierarchical social network among students in one apartment building where a severe H1N1 outbreak occurred. A compartmental based stochastic model is proposed to simulate the spreading process of the epidemic in the network, finding community outbreaks within small social groups. Moreover, we compare the effect of three quarantine-based non-pharmaceutical interventions (including dormitory building quarantine only, plus between-room visits prohibition, and plus relocation of all people at risk to a treatment center) on transmission, aiming to find out the most efficient control measures with acceptable cost. In practical, it needs to respond to the

infectious outbreak timely, we compare the strategies to advise how to control the spreading in an emergency.

2 Related Work

The past several years have seen dramatic advances in exploring the transmission pattern of infectious disease and evaluating public health response and control strategies [10]. From a methodological standpoint, these studies typically employ either case-study or simulation methods. We argue that these two types of methods can complement each other very well. Studying real-world outbreaks through epidemiological and case studies can reveal critical insights to help understand the dynamics of disease transmission and identify risk factors [11-14]. These insights can also serve as the base for the follow-up analytical modeling and computational simulation work and facilitate an understanding of the impact of various public health responses [15, 16]. However, using case studies and real-world observations alone cannot guide the selection of the optimal control or response strategies because of lack of support for comparing various alternative measures. Simulations, on the other hand, offer a complementary framework enabling us to quantitative study the impact of various strategies [17]. In particular, detailed cost and benefit analyses are possible through simulation. [18, 19].

A commonly used framework for simulation of epidemic spreading is social network. After a great development in last two decades, various kinds of complex properties are found in social network structure [20]. These system properties, which emerge from individuals' behaviors and interactions, have a big impact of epidemic transmission among persons in the network. For example, recent study [21] shows that infection would tend to spread over and stay at a steady state in a heterogeneous networks (i.e., a network in which a small fraction of individuals possess most of contacts and many of other individuals only have a few contacts, this property is present in many real networks such as [22, 23]).

It has been shown that many real social networks are hierarchical organized by linking many of small communities [5, 24]. This network property is commonly called as community structure, which greatly influences the dynamic effects on real networks [25, 26]. For infectious disease transmission, the community structure property of social contacts network would arouse community outbreaks within small social groups, which have attracted public attention for intervention methods design in order to control epidemic spreading [6, 7].

3 Data and Methodology

3.1 Data

From August 28, through September 17, 2009, a big H1N1 outbreak occurred on a campus of a Chinese university near Beijing. 206 students are infected during this outbreak in total, and about a half (105) of them are living in a six-storied apartment building. In this paper, we are mainly focus on H1N1 virus spreading in this one

building, as it was the main infection hotspot and a center of student interactions during the outbreak.

Students living in this building belonged to the same academic department, and were divided into 14 classes. The students of the same class were assigned to adjacent rooms and each room typically housed six students. Therefore, students in this dormitory are organized hierarchically from a single dormitory room to a class to the whole building. To further understand the social contacts between students, we conduct a preliminary epidemiological investigation of these confirmed cases to collect their close contacts before onset of the illness. In this investigation, 30 respondents recalled their close contact history, involving in total 45 virus carriers. We list the relationship between these respondents and their infected contacts in Table 1. Before the building was isolated, 9 respondents were mainly infected by outsiders. After the building was isolated, 21 respondents were mainly infected by roommates and in a lesser degree by classmates in other rooms. Moreover, we investigate all infected cases about their time delay from infected to be hospitalized, finding that it took about 1.5 days on average.

In order to control this outbreak, the administration of the school took several emerging interventions, including 1) isolating the apartment building to stop the inter-building transmissions on August 31; 2) prohibiting students in the apartment building to contact other students living in different dormitory rooms on September 3; 3) quarantining all probable close contacts when a infected student is sent to the hospital in the entire outbreak.

Table 1. The relationship between the respondents and their contacts to H1N1 virus carriers

	Respondents	contacts	Roommates	Classmates	Different classes	Different buildings
Whole period*	30	45	16	11	2	16
Before isolate building	9	13	0	0	0	13
After isolate building	21	32	16	11	2	3

3.2 Method

As shown in Table 1, social contacts among the students in the apartment building show a community structure property, as a student has a biggest probability to be infected by his (her) roommates and has a moderate probability to be infected by his (her) classmates in different rooms and has a smallest probability to be infected by students in different classes. In order to model social contacts among students in the building, we construct a complex network by representing students as nodes and social relationships among them as links. Due to the community structure property of social contacts among students in the apartment, we allocate to each student different probabilities that contacting to his roommate, neighbors in the same classes, and others in the same building, respectively. The probabilities are preset based on the epidemiologic investigation: we assume that the contact histories of the 30 respondents in Table 1 can reflect the behaviors of other students. As the persons living in

the same room should contact with each other determinately in each day, we obtain the average number of one's contactors who are roommates, classmates but living in other rooms, others living in the same building is $p_1 = 5$, $p_2 = 3.5$ and $p_3 = 0.5$, respectively. The network model is shown as Fig. 1.

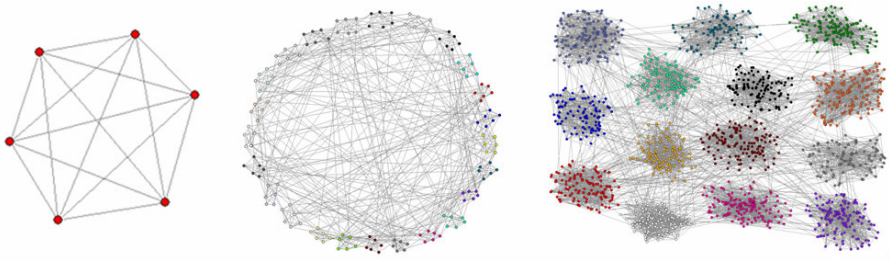


Fig. 1. (a) The social contacts network in one room. All students living in one room are fully connected. (b) The social contacts network in one class. Many edges exist in each room, and fewer edges between rooms. (c) The social contacts network in the building. The students are divided into different classes; many edges exist in each class, and fewer edges between classes.

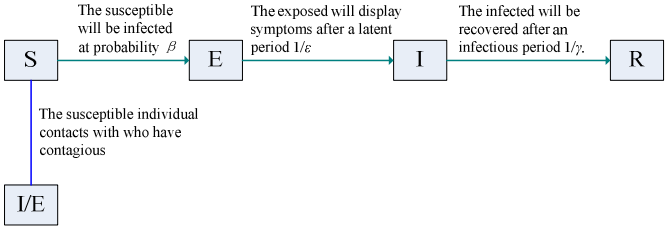


Fig. 2. The state change process of epidemic spreading in SEIR model

A stochastic SEIR model was applied to simulate the disease spreading process on the network. In the SEIR model, people are classified into four classes: Susceptible, Exposed (Incubation), Infected, and Recovered [27]. The state change process of epidemic spreading in SEIR model is shown in Fig. 2. We set the latent period $1/\varepsilon = 2$ days according to some empirical studies of H1N1 transmission [28, 29] and theinfectious period as $1/\gamma = 1.5$ days based on our epidemiologic investigation. The spreading rate β was set to 0.18, which leads that the coefficient of determination, a measurement of how well the real data can be explained by the model, between the daily counts of new infections in the real outbreak and simulation results is maximum.

Following the control strategies the school administrator taken during the outbreak, we divide the outbreak into three periods: 1) From August 28 to August 31, the apartment building was not isolated, and students in the building may be infected by infectors outside the building with a probability $\lambda = 0.004$; 2) From September 1 to September 3, infection can only be transmitted within the apartment building; 3) After September 3, the contacts between two students in different rooms are cut off with probability $\eta = 1$, meaning that infection would not be spread inter-rooms.

During the outbreak, the university administration investigated the infected to identify their close contacts for quarantining. In practice, however, identify all contacts successfully is difficult as it is very time-consuming to collect contacts histories of infected students. As a proxy, the roommates of a patient should be viewed as automatic close contacts and considered as candidates to be quarantined. In our simulation, we assume that all roommates of an infected patient would be quarantined when the infected patient is hospitalized, and preset the probability of identifying and quarantining a neighbor exposed to an infected is θ =0.8 according to the investigation taken by the administration. Related parameters are listed in Table 2.

Using our network-based model, we have studied the impact of three different control measures on H1N1 spread. We adjust the date to implement these measures in the simulation on one hand and the parameters λ, η, θ on the other hand, in order to quantitatively examine the impact of timeliness and seriousness of control measures on control effects. The outcomes of control measures are measured as the number of infected and isolated students reduced by the strategies.

Table 2. The parameters used in the social contacts network and SEIR model

Parameter	value	Interpretation
p_1	5	Average number of one's contactors who are roommates
p_2	3.5	Average number of one's contactors who are classmates but are living in other rooms
p_3	0.5	Average number of one's contactors who are students belong to other classes
β	0.18	Spreading rate
$1/\gamma$	1.5	Average infectious period
$1/\varepsilon$	2	Average latent period
λ	0.004	The probability that a student is infected by outsiders in the first period
η	1	The probability that edges between rooms are removed in the third period
θ	0.8	The probability that close contactors living in other rooms can be found during the outbreak

4 Result

4.1 Social Contact Network and Stochastic Simulation

The number of the newly infected in the building each day is shown in Fig. 3. In the first period, the infection was mainly introduced into the building from the outside and spread over more and more students. The transmission peaked in the second period and then reduced rapidly. If we take into account the 2-day latent period of the H1N1 virus, we conclude that the number of students exposed to infection decreased quickly after the administration isolated the building. Furthermore, the infection rate maintained at an extremely low level after the contacts among rooms were prohibited.

We simulated the H1N1 virus spreading process using the social contact network with a community structure. The related parameters are listed in Table 2. As shown in Fig. 1, the students first form small groups through roommate relationships, then form larger clusters through connecting rooms belonging to the same classes, and at last form the entire contact network for the building. The simulation results, averaged over 100 experimental runs, turn out to fit the overall outbreak pattern very well, as shown in Fig. 3. The coefficient of determination between the number of infected students on each day in the real outbreak and simulation results is 0.79. Fig. 4 illustrates the relationship between infection transmission and social networks based on a

typical single simulation run. It shows that inter-class transmission rarely exists. Within each class, frequent internal transmissions in rooms exist and the inter-room transmissions lead to infection spreading over the class.

We further observe that the transmission mode of the H1N1 outbreak is aligned with the community structure of the social contact network. Infection spread quickly among densely connected groups of people. Fig. 5(a) shows the distribution of infected number in dormitory rooms. There were a total of 68 dormitory rooms in which students have been got infected. Among these rooms, 44(64.7%) room with only one infected case, the other 24(35.3%) rooms possessed more than two infected cases despite the fact that the roommates of the infected students were quarantined promptly. The results indicate that rooms are relatively independent units where the infection is rapidly spreading.

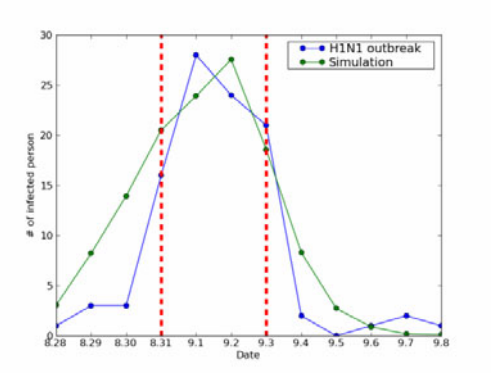


Fig. 3. The spreading process of the infection. Each point on the blue and green lines represents the number of the newly infected in the building each day of the situation of real outbreak and simulation, respectively. The outbreak is divided into three periods by the red dashed line.

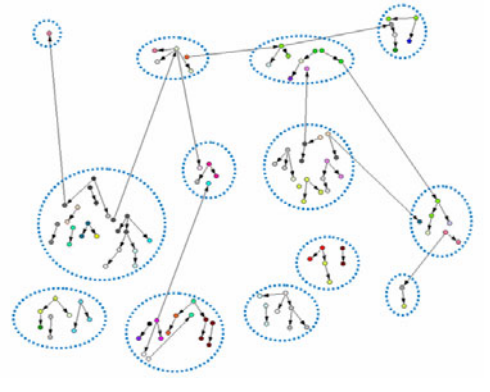


Fig. 4. The spreading process on the social contact network based on a typical single simulation run. Each point represents a student, and each edge represents a spreading path of the infection. The students are divided into various classes, denoted as circles. The students living in one room are marked with the same color.

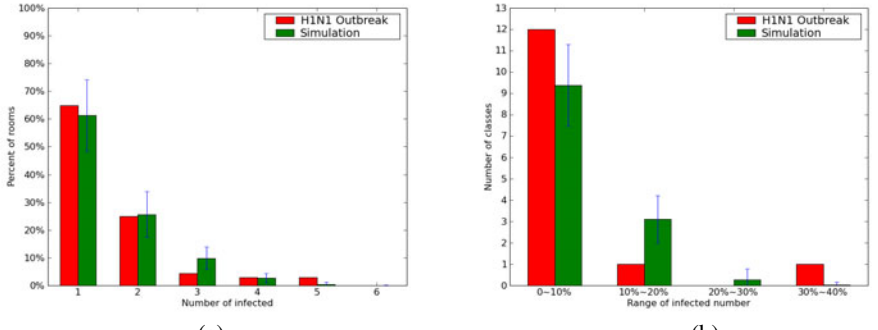
During the outbreak, the infection situation differed greatly across classes: the average infected number in each class is 7.5 whereas the standard deviation is 9.0. Note that the number of students across classes is quite uniform with the average being 136.7 and the standard deviation 18.2. As shown in Fig. 5(b), 12 classes housed less than 10% of infected students and two classes experienced larger outbreaks with 15(10% to 20%) cases and 32 cases (30%-40%). This imbalance implies that an outbreak could be occurring in individual classes with limited inter-class interference.

In general, the simulation produced H1N1 transmission in small communities such as rooms and classes very similar to what was observed in the real outbreak. We measured the Pearson's correlation coefficient between the number of dormitory rooms (classes) which possessed different infection number in real outbreak and simulation results. We obtained a value of 0.99 (0.95 for the class, both with $p<0.005$).

4.2 Evaluation of Control Measures

Isolating the whole building. When an outbreak occurred on a campus, isolation of buildings could help prevent epidemic spreading among buildings. As shown in Fig. 6 (a-b), comparing with isolating the whole building on the seventh day, taking the strategy on the first day can reduce the number of infected and isolated students in the entire outbreak by 85%~92%. The finding indicates that it is essential to strictly control visitations and identify the infected patients as soon as possible.

In practical terms, due to the high social and economic costs associated with building isolation, one needs to make careful decisions as to the timing of the isolation after H1N1 infection is reported in a school. In this model, the number of infected isolated students in the entire outbreak is 67 (300) fewer if the whole building is isolated on the first day than on the seventh day when $\lambda = 0.001$. While, when $\lambda = 0.007$, the building isolation taken on the first day would prevent 234 (988) students to get infected (isolated) compared to taking the strategy on the seventh day. In fact, we isolated the building on first day, only 26 infected and isolated 110 close contacts. The simulation results show that the more probable a student gets infected by an outsider, the more beneficial an earlier isolation.



(a) (b)

Fig. 5. (a) The percent of rooms versus the number of infection cases in one room. (b) The number of classes versus the ratio of infected number in one class to the sum infected in the building.

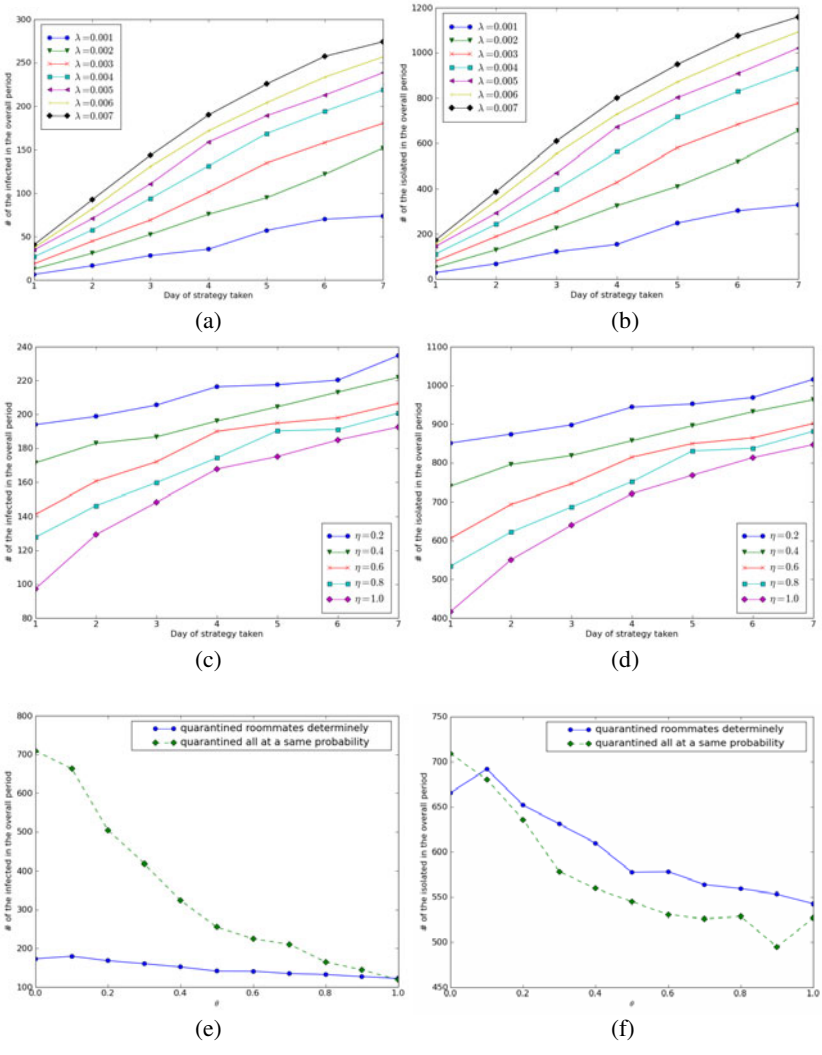


Fig. 6. The comparison of effectiveness of different control strategies under different setting. (a-b): The sum of (a) infected and (b) isolated versus the day that building is isolated, with λ = 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007 from bottom to top, respectively. (c-d) The sum of (c) infected and (d) isolated versus the day that contact among rooms is cut after isolation of building, with η = 0.2, 0.4, 0.6, 0.8, 1.0 from bottom to top, respectively. (e-f): The sum of (e) infected and (f) isolated versus the probabilities that close contacts are found. The solid line represents the roommates are determinedly quarantined and others are quarantined with specific probability denoted as θ; the dash line represents a comparative method that all contacts are quarantined with the same probability denoted as θ.

Cutting off the contacts across rooms. We have found that frequent visitations between dormitory rooms can trigger clustered outbreaks within classes. Fig. 6 (c-d) shows the effect of reducing inter-room contacts on different dates after building isolation became effective on August 31, 2009. If we take the strategy on the first day, cutting off all contacts across rooms would prevent 97 (435) (η=1.0 vs η=0.2) students to get infected (isolated) compared to the situation that 20%(η=0.2) contacts are cut off. While, the decreased ratio of infected/isolated students was 49.4%/50%(42/169) (η=1.0 vs η=0.2) when the strategy taken on the seventh day. The simulation results reveal that the impact of contact reduction measures on disease spread is minimum if these measures are taken after the outbreak is already progressing fast. From the policy perspective, the timing of the implementation needs to be carefully considered.

Quarantining the close contacts. Due to the existence of the latent period of the H1N1 virus, the close contacts of an infection case may be infected and contagious but without symptoms temporarily. As such, one possible outbreak containment strategy would be to identify and quarantine the close contacts exposed to the infection before symptoms show. We have found that the roommates have frequent contact with each other and need to be quarantined immediately when an infection case is found in a room. This observation leads us to experiment the control measure that isolate the infected person and his or her roommates during the simulation. Fig. 6 (e-f) shows the effectiveness of this measure. Without the previous knowledge of quarantining roommates of infected students, the incremental number of infected students is 536 when quarantining none of possible contacts; this drop to 114 when half of possible contacts are quarantined, regardless of the difficulties in accurately identifying possible contacts in time. Almost close contacts were found and quarantined, so the infected case (105) in this outbreak less than the simulated number(122) when θ=1.0(Fig. 6 (e)).Although the cost of quarantining all the roommates is high, the total number of students being isolated using this measure is almost the same as that of the alternative measure during the entire course of the outbreak(Fig.5 (f)) There were no difference(P>0.05) on number of isolated whether quarantined the roommates and all close contacts or only probable contacts(542vs526) when θ=1.0(Fig. 6 (f)). As a result, we note that if the public health officials can quarantine the roommates of infected persons in a very short period of time, it is likely that the total number of infections can be significantly reduced.

4 Discussion

The dynamic of epidemic outbreak influenced by many factors including host immunity, virus virulence, human behavior, environmental change, social and economical situation, etc [10]. These factors influenced the virus transmission from human society, school to household, but social relationship as a special factor maybe influenced the transmission dynamic of epidemic outbreak in high density population [2-4, 6, 7]. In our research, we have investigated a real-world H1N1 outbreak occurred on a Chinese university campus. We have identified the community structure in students' social contact network and concluded that the community structure strongly affects the H1N1 spread and triggers clustered outbreaks. Based on our empirical findings, we have constructed a hierarchical social contact network model. Our computational

experiments indicate that our model fits the real-world outbreak pattern very well and displays the clustering property of spreading through Monte-Carlo methods.

In this study, we investigated and identified a more comprehensive set of social factors influencing disease transmission of 2009 influenza A(H1N1) outbreak in density population special in school. The epidemic spread among susceptible people influenced by many social factors and identified by social network. For instance, the social contacts among students may be strongly influenced by their perception of the outbreak during the outbreak. Another significant result of the control measures used in this outbreak according the social relationship determined by models was effective and coordinated with models.

Using this social contact network model organized as a hierarchical network, we have conducted simulation studies to evaluate three public health response measures relevant to localized outbreaks: isolating the entire building, cutting off the contacts among rooms, and quarantining the close contacts of the infected. We have found that cutting off contacts among buildings and rooms can help contain the epidemic spread. As the H1N1 infection typically grows exponentially at first and tends to stabilize later [2], implementing the control measures earlier can be more beneficial.

The community structure of social contacts on campus leads to clustered outbreaks. On practical terms, we recommend that the roommates of the infected students be isolated as early as possible. The quickly infection spreading in the school requires emergency interventions, and the public measures implemented by the school administrator and our simulation, according to the community structure of social contacts, performs efficiently to respond to H1N1 outbreak in a localized area in time. On the contrary, prevaccination would be not practicable because vaccine cannot take effect instantly.

A limitation of our study is the lack of a quantitative analysis of cost-effectiveness of various measures. Real costs of a response measure are difficult to estimate, which include both social costs and economic costs caused by reducing contacts, providing isolation spaces, investigating contacts of the infected, handing out mask, taking protective measures and so on. The economic and social benefit of containing an outbreak is also hard to quantify, as it will be dependent on the virulence and clinical severity of illness, and the productivity and economic loss due to closed schools and so on. Nonetheless, we believe that our study provides a starting point to assist public health response decision-making during epidemic outbreaks in localized areas.

Acknowledgments. This work was supported in part by the Major-projects of Science and Technology Research (Grant No. 2009ZX10004-315, 2008ZX10004-008, 2008ZX10005-013), the Chinese Academy of Sciences (Grants No. 2F07C01, 2F08N03), National Natural Science Foundation of China (Grant No. 90924302, 60621001, 60921061, 40901219, 71050001, 91024030), and the US National Science Foundation (Grant No. IIS-0839990).

References

1. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. Proceedings of the Royal Society of London 115, 700–721 (1927)
2. Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A.: Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of Theoretical Biology 235, 275–288 (2005)

3. Kiss, I.Z., Green, D.M., Kao, R.R.: The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. Journal of The Royal Society Interface 5, 791–799 (2008)
4. Moore, C., Newman, M.E.J.: Epidemics and percolation in small-world networks. Physical Review E 61, 5678 (2000)
5. Newman, M.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103, 8577 (2006)
6. Yang, Y., Sugimoto, J., Halloran, M., Basta, N., Chao, D., Matrajt, L., Potter, G., Kenah, E., Longini Jr., I.: The transmissibility and control of pandemic influenza A (H1N1) virus. Science 1177373v1177371 (2009)
7. Cauchemez, S., Bhattarai, A., Marchbanks, T.L., Fagan, R.P., Ostroff, S., Ferguson, N.M., Swerdlow, D.: Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proceedings of the National Academy of Sciences (2009)
8. Lipsitch, M., Cohen, T., Cooper, B., Robins, J.M., Ma, S., James, L., Gopalakrishna, G., Chew, S.K., Tan, C.C., Samore, M.H., Fisman, D., Murray, M.: Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. Science 300, 1966–1970 (2003)
9. Ou, J., Dun, Z., Li, Q., Qin, A., Zeng, G.: Efficiency of the quarantine system during the epidemic of severe acute respiratory syndrome in Beijing. Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi 24, 1093 (2003)
10. Grassly, N.C., Fraser, C.: Mathematical models of infectious disease transmission. Nat. Rev. Micro. 6, 477–487 (2008)
11. Han, K., Zhu, X., He, F., Liu, L., Zhang, L., Ma, H., Tang, X., Huang, T., Zeng, G., Zhu, B.: Lack of airborne transmission during outbreak of pandemic (H1N1) 2009 among Tour Group Members. Emerging Infectious Disease 10, 1578–1581 (2009)
12. Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D., Zhu, Z., Schuchat, A.: Superspreading SARS events, Beijing. Emerging Infectious Diseases 10, 256–260 (2004)
13. ZhiDong, C., DaJun, Z., QuanYi, W., XiaoLong, Z., FeiYue, W.: An epidemiological analysis of the Beijing 2008 Hand-Foot-Mouth epidemic. Chinese Science Bulletin 55, 1142–1149 (2010)
14. Cao, Z.D., Zeng, D.J., Zheng, X.L., Wang, Q.Y., Wang, F.Y., Wang, J.F., Wang, X.L.: Spatio-temporal evolution of Beijing 2003 SARS epidemic. Science China Earth Sciences, 1–12 (2010)
15. Bondy, S., Russell, M., Lafleche, J., Rea, E.: Quantifying the impact of community quarantine on SARS transmission in Ontario: estimation of secondary case count difference and number needed to quarantine. BMC Public Health 9, 488 (2009)
16. Yasuda, H., Suzuki, K.: Measures against transmission of pandemic H1N1 influenza in Japan in 2009: simulation model. European Communicable Disease Bulletin 14 (2009)
17. Hsieh, Y., King, C., Chen, C., Ho, M., Hsu, S., Wu, Y.: Impact of quarantine on the 2003 SARS outbreak: A retrospective modeling study. Journal of Theoretical Biology 244, 729–736 (2007)
18. Dan, Y., Tambyah, P., Sim, J., Lim, J., Hsu, L., Chow, W., Fisher, D., Wong, Y., Ho, K.: Cost-effectiveness Analysis of Hospital Infection Control Response to an Epidemic Respiratory Virus Threat. Emerging Infectious Disease 15, 1909–1916 (2009)
19. Dasgupta, K., Menzies, D.: Cost-effectiveness of tuberculosis control strategies among immigrants and refugees. European Respiratory Journal 25, 1107 (2005)
20. Newman, M., Barabasi, A., Watts, D.: The structure and dynamics of networks. Princeton Univ. Pr., Princeton (2006)

21. Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86, 3200 (2001)
22. Zheng, X.L., Zeng, D., Li, H.Q., Wang, F.Y.: Analyzing open-source software systems as complex networks. Physica A 387, 6190–6200 (2008)
23. Yin, H., Rong, Z., Yan, G.: Development of friendship network among young scientists in an international Summer School. Physica A: Statistical Mechanics and its Applications 388, 3636–3642 (2009)
24. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99, 7821–7826 (2002)
25. Danon, L., Arenas, A., Díaz-Guilera, A.: Impact of community structure on information transfer. Physical Review E 77, 036103 (2008)
26. Wu, J.-j., Gao, Z.-y., Sun, H.-j.: Cascade and breakdown in scale-free networks with community structure. Physical Review E 74, 066111 (2006)
27. Hethcote, H.: The mathematics of infectious diseases. SIAM Review 42, 599–653 (2000)
28. Tuite, A.R., Greer, A.L., Whelan, M., Winter, A.L., Lee, B., Yan, P., Wu, J., Moghadas, S., Buckeridge, D., Pourbohloul, B.: Estimated epidemiologic parameters and morbidity associated with pandemic H1N1 influenza. Canadian Medical Association Journal 182, 131 (2010)
29. Canini, L., Carrat, F.: Population Modeling of Influenza A/H1N1 Virus Kinetics and Symptom Dynamics. J. Virol. 85, 2764–2770 (2011)

Modeling and Simulation for the Spread of H1N1 Influenza in School Using Artificial Societies

Wei Duan[1,2], Zhidong Cao[2], Yuanzheng Ge[1], and Xiaogang Qiu[1]

[1] College of Mechatronics Engineering and Automation,
National University of Defense Technology, 410073 Changsha, China
[2] Key Laboratory of Complex Systems and Intelligence Science, Institute of Automation,
Chinese Academy of Sciences, 100190 Beijing, China
duanwei@nudt.edu.cn

Abstract. According to the outbreak of H1N1 influenza on campus in Langfang city, Hebei province, at north of China in 2009, this paper constructed an artificial society model of the school, and simulated the spread of H1N1 influenza at the fifth floor in dormitory building. Firstly, it built the geographic environment model in accordance with the real dormitory building and a social relationship network model, including classmates, roommates and playmates. Secondly, it designed the behaviors and activities of students during a day, and built role based agent models of student. Each agent student had three roles, which were susceptible, infectious and recovered student. Finally, it conducted simulation experiments to compare the emergency measures of segregating non-classmates and segregating non-roommates.

Keywords: Public Health Security; Modeling and Simulation; H1N1 Influenza; Emergency Management.

1 Introduction

Public health security is closely related to national economy and social stability. So the sudden outbreak of epidemic will cause serious harm to the health of people, and damage national economic development. For example, the outbreak of SARS [16] in 2003 and H1N1 influenza [1] in 2009 has seriously affected the lives of people and economic development, and even led to the death of some patients. Therefore, it is an urgent issue that how to manage the sudden outbreak of epidemic effectively and fleetly. Owing to the fact that the outbreak of epidemic is always sudden and unpredictable, it is required to make reasonable and effective emergency management strategies beforehand. So the traditional research mode of emergency management, "forecast-response", cannot meet the demand. However the research mode based on modeling and simulation [2], [3], "scenario-response [4]", has become an important way to resolve sudden outbreak of epidemic.

The spread of epidemic is complex and variable that has many related causes [5]. So the traditional simulation aiming at approximating the single reality cannot simulate the complex and variable scene of epidemic emergencies. Whereas artificial

M. Chau et al. (Eds.): PAISI 2011, LNCS 6749, pp. 121–129, 2011.
© Springer-Verlag Berlin Heidelberg 2011

societies [6], [7] based modeling and simulation, as a bottom-up approach, could come forth complex social phenomena through evolution, interaction and cooperation between multi-agents. So it is an effectual resolution for emergency drills, emergency mental training and assessment of emergency strategies to research epidemic emergencies management.

In the spread of epidemic, the school is very special high-risk places. Because in school, students are high concentrated, and keep high consistency in behaviors and frequently close contact, such as living, learning, eating, playing and so on, it results in that the school is vulnerable to the outbreak of epidemic [8], [9], [10]. For instance, in the spread of H1N1 pandemic influenza in Beijing in 2009, more than 70% of outbreaks took place in schools. So protecting students from epidemic is an important task in public health emergency management.

This paper constructed an artificial society to simulate the spread of H1N1 influenza at the fifth floor in dormitory building according to the outbreak of H1N1 influenza on campus [11] in langfang city, Hebei province, at north of China in 2009. It made the foundation for simulating the spread of H1N1 influenza in any school and decision making for epidemic emergencies.

The remainder of this paper is organized as follows. Section 2 built the geographic environment model of the fifth floor, social relationship network model among students, behaviors and activities model of students during a day, and role models of students. Section 3 conducted simulation experiments for different measures of segregation, and compared these results of experiments.

2 Modeling with Artificial Societies

2.1 Geographic Environment Model

According to the real structure of the fifth floor and the population distribution of students in dormitory building, geographic environment model is built as figure 1. There are three classes in the fifth floor, including Class 4, Class 7 and Class 13, and 64 dorms, 3 WC, 3 washrooms, 3 storages, 3 Classrooms, 1 balcony. There are 8 dorms belong to Class 4, whose names are the odd number from 501 to 505 and the even number from 502 to 510, and 25 dorms belong to Class 7, whose names are the odd number from 507 to 525 and even number from 512 to 540, and 21 dorms belong to Class 13, whose names are the odd number from 527 to 543 and the even number from 542 to 564. And there are 6 students live in each dorm.

2.2 Social Relationship Network Model

The spread of H1N1 influenza is propelled by close contacts among students. And the probability and intensity of close contacts among students are depended on the mutual social relations [12], [13]. For instance, the probability and intensity of close contacts among roommates is higher than the one among non-roommates, and the same as classmates and non-classmates. In this paper it researched three relations between students to build social relationship network model. These relations are classmates,

roommates and playmates. Classmates could be roommates and non-roommates, but non-classmates must be non-roommates. And playmates could be classmates and non-classmates, as well as be roommates and non-roommates. The closeness of relations among students is different and their proportions are shown in table 1. In simulation experiment, these proportions determine the probability distribution of contact objects.

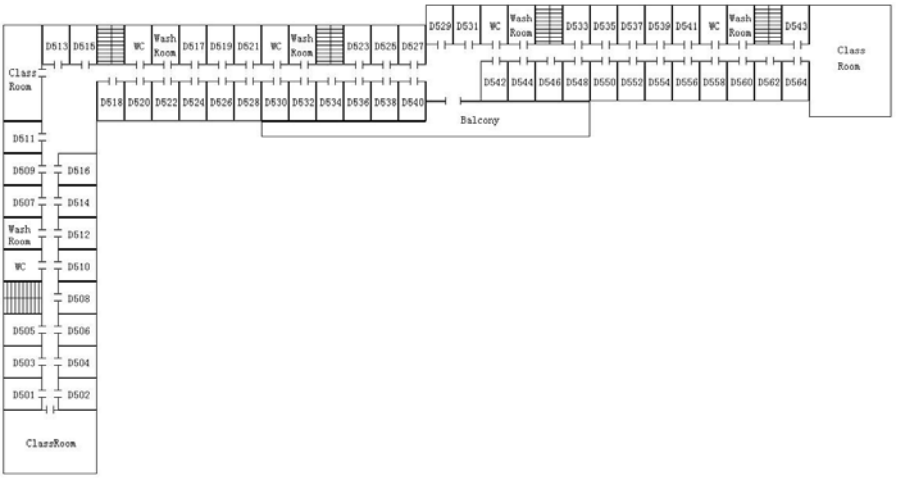


Fig. 1. The fifth floor model

Table 1. Closeness of relations between students

	Classmates		Non-classmates
	Roommates	Non-roommates	
Playmates	40%	20%	10%
Non-playmates	20%	10%	0%

The relations of classmates and roommates are confirmed by dorms which students live in. If students live in the same one dorm, they are roommates and classmates. If students live in different dorms, they are non-roommates, and they are classmates when these dorms belong to the same one class, or they are non-classmates. The relation of playmates is not confined to living place. In simulation experiments, it assumed that each student has 0 to 5 playmates, which are randomly generated at the beginning of experiment. Figure 2 shows random relationship among playmates.

In the evolution of artificial societies, relationships among agents may change continually [14]. It makes the configuration of artificial societies variable so as to come forth complex and diverse social phenomena. In this social relationship network model, the relations of classmates and roommates are unchangeable. However playmates may change with the evolution of artificial societies. In simulation experiment students who contact continually in certain time may become playmates. Contrarily,

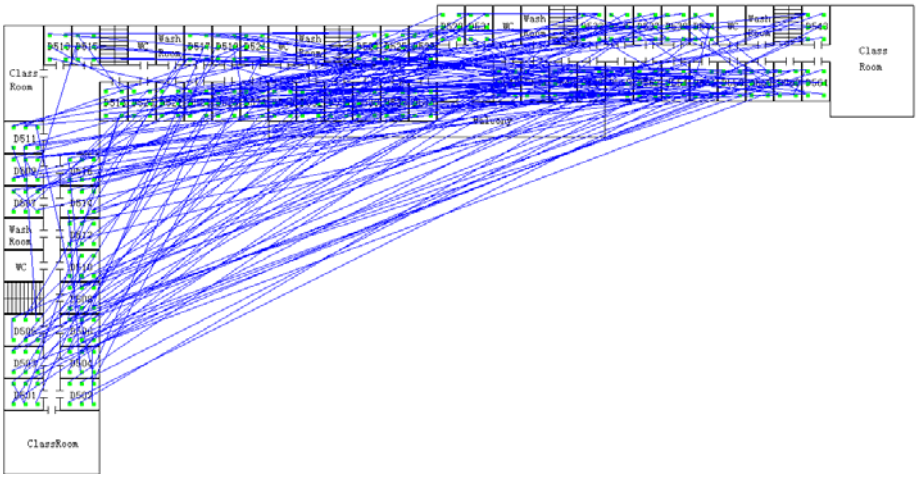


Fig. 2. Contact relationships among playmates

Table 2. Behaviours and activities model

Time	Location	Contact probability	Time	Location	Contact probability
7:00-8:00	dorm, washroom, WC	60%	12:00-13:30	dorm, balcony, WC	75%
8:00-8:50	classroom	10%	13:30-14:30	dorm, WC	6%
8:50-9:00	classroom, dorm, balcony, WC	80%	14:30-15:20	classroom	10%
9:00-9:50	classroom	10%	15:20-15:30	classroom, dorm, balcony, WC	80%
9:50-10:10	classroom, dorm, balcony, WC	80%	15:30-16:20	classroom	10%
10:10-11:00	classroom	10%	16:20-22:00	classroom, dorm, balcony, WC	75%
11:00-11:10	classroom, dorm, balcony, WC	80%	22:00-23:00	dorm, washroom, WC	60%
11:10-12:00	classroom	10%	23:00-7:00	dorm, WC	2%

students may become non-playmates when they have not contacted with each other for a long time.

2.3 Behaviors and Activities Model

The Spread of H1N1 influenza is closely related to the daily behaviors and activities of students. The main behavior which directly causes the spread of H1N1 influenza is close contacts among students. In daily life of students, the spread occurs with diverse probabilities of different behaviors and activities, and students have various behaviors

and activities during a day. In this paper, it designed the following behaviors and activities, including getting up, washing, going to toilet, going to classroom, taking classes, playing in dorms, playing in balcony, going to bed, talking, and learning in classroom. It defined behaviors and activities model during a day as table 2. As shown in table 2, it determines different action location in each time phase during a day, and contact probability derived from empirical data. In simulation experiment, according to the contact probability, agent decides whether have close contact with other agents in each simulation step, such as playing, talking, etc.

2.4 Role Models

In order to achieve that attributes and behaviors in agent model could dynamically change, we builds various role models for the same one agent [15]. During the evolution process of artificial societies, when the states of agent satisfy role changing conditions, the value of attributes and parameters of behaviors transfer from one role model to another role model. So it realizes role evolution of agents. In this paper, we build three roles, such as susceptible, infectious, and recovered student, which are shown in figure 3. As seen from figure 3, the green square represents susceptible student who has never been infected by H1N1; the red circle represents student infected by H1N1 influenza; the blue hollow box represents recovered student who has ever been infected by H1N1 and have antibody to H1N1 influenza.

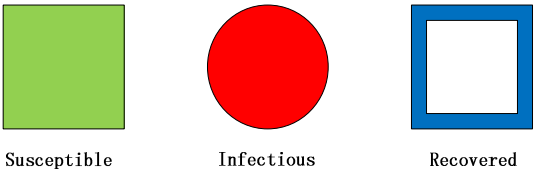


Fig. 3. Role models

Three role models have different attributes and behaviors. The infectious role model has special attributes of epidemic type and infected time, and the behavior of infecting other students. However, the recovered role model has special attribute of antibody type and the behavior of preventing from H1N1 influenza.

3 Simulation Experiments

With these models above, we could design and conduct simulation experiments. In simulation experiments, we set the walking speed of each agent to be one meters per second, and set simulation step to be one second. And the beginning time is at 7 am when students are getting up, and the whole simulation time is 30 days. Moreover, we assume that there is one infectious student in class 4 at the beginning of experiments, and it takes emergency measures for the outbreak of H1N1 influenza when the school

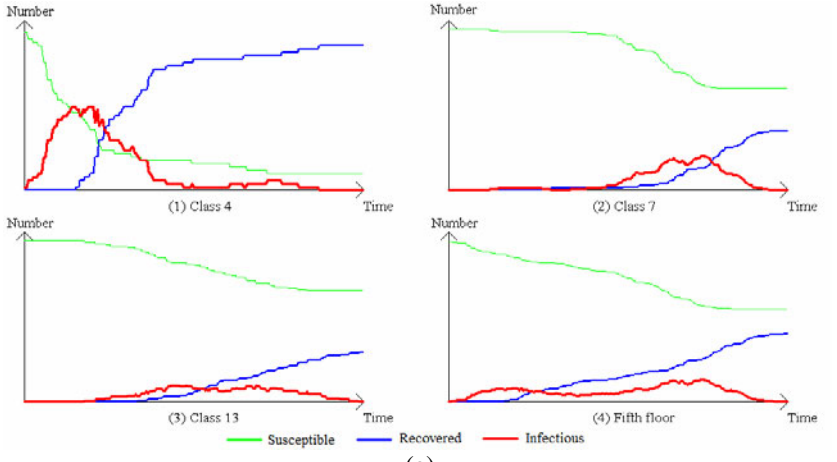
detects more than 10 students infected by H1N1 influenza. The emergency measures include segregating non-classmates and segregating non-roommates.

We conduct three simulation experiments, experiment (a), experiment (b) and experiment (c), to compare different measures for preventing student from H1N1 influenza. In experiment (a), no measures are taken. In experiment (b), students are segregated from non-classmates when the school detects more than ten students infectious by H1N1 already. In experiment (c), students are segregated from non-roommates. They have been confided to their dorms, and not allowed to go to classroom, balcony and other dorms. Figure 4 is a scene of running simulation experiment (a) at time from 16:20 to 22:00.

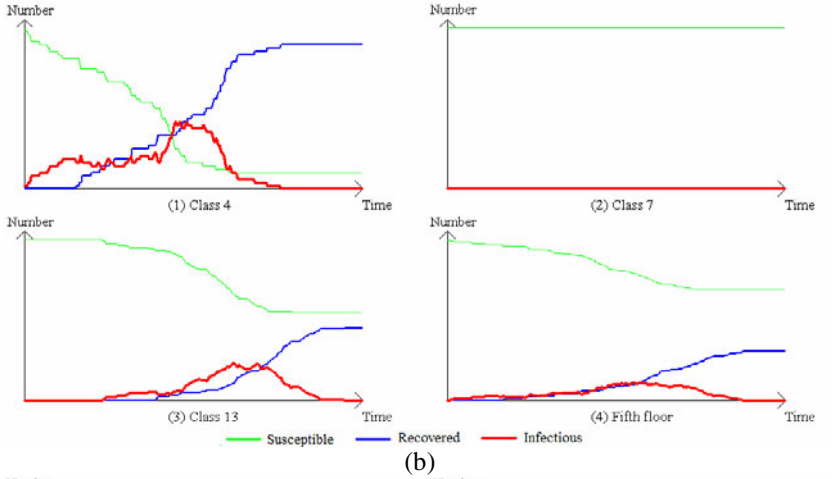
Fig. 4. Simulation experiment scene

The results of these experiments are respectively shown in figure 5 (a), (b) and (c). Each experiment has 4 graphs for class 4, class 7, class 13 and the fifth floor, to depict the changes of the number of susceptible, infectious and recovered students with three different kinds of curves.

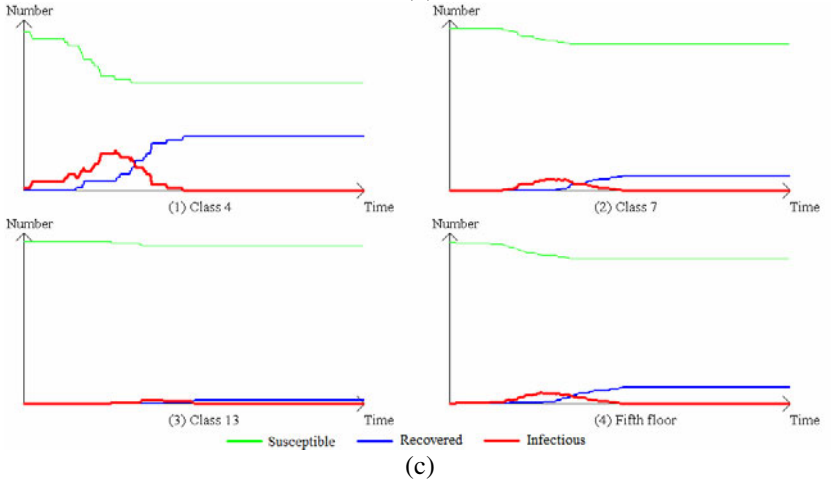
As figure 5 (a) shown, the only one infectious student in class 4 at the beginning of experiment, has spread H1N1 virus abroad to class 7 and class 13. And we can see from figure 5 (b) (2) that no students are infected by H1N1 influenza in class 7. It proves that emergency measure of segregating non-classmates has prevented class 7 from infectious students. From figure 5 (a) (4), (b) (4), and (c) (4), we know that the number of infectious students in the fifth floor decline, and become zero more and more quickly. It could conclude that the emergency measures of segregating non-classmates and non-roommates could effectively protect students from the spread of H1N1 influenza, and that segregating non-roommates is a much better measure than segregating non-classmates.



(a)



(b)



(c)

Fig. 5. Student number curves in three roles

4 Summary and Outlook

According the case of the outbreak of H1N1 influenza on campus in langfang city, this paper built models of the fifth floor and behaviors of students in the school. It simulated the spread of H1N1 influenza and reproduced the scene of the outbreak of H1N1 influenza. Besides, it designed a variety of random parameters in simulation experiments, and two kinds of segregation strategies, and analyzed the results of experiments with different segregation strategies. In the following works, we will build a whole school model using artificial societies on the basis of what we get from this paper to simulate the outbreak and spread of epidemic in school, and to support decision making in emergency management.

Acknowledgments

This work was supported by National Natural Science Foundation (No. 91024030, 90924302, 40901219, 71050001), and Key Projects of National Science and Technology (No. 2009ZX10004-315, 2008ZX1005-013), and Projects of Chinese Academy of Science (No. 2F07C01, 2F08N03).

References

1. Cai, C., Zhong, N.S.: Study of 2009 influenza A (H1N1) virus epidemic. Chinese Journal Critical Care Medicine 29, 553–555 (2009) (in Chinese)
2. Ozgur, M.A., John, W.F., Tim, W.L., Megan, J.: A Pandemic Influenza Simulation Model for Preparedness Planning. In: Proceedings of the 2009 Winter Simulation Conference, pp. 1986–1994. IEEE Press, Los Alamitos (2009)
3. Liu, T., Li, X., Liu, M.W.: Multi-Agent System Simulation of Epidemic Spatio-temporal Transmission. Journal of System Simulation 21, 5874–5877 (2009) (in Chinese)
4. Li, S.M., Liu, J.J., Wang, B., Xiao, L.: Unconventional Incident Management Research Based on Scenarios. Journal of University of Electronic Science and Technology of China 11, 1–3 (2010) (in Chinese)
5. Lioyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Gets, W.M.: Superspreading and the effect of individual variation on disease emergence. Nature 438, 335–359 (2005)
6. Joshua, J.E., Robert, A.: Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, Cambridge (1996)
7. Wang, F.Y.: A computational framework for decision analysis and support in ISI: Artificial societies, computational experiments, and parallel systems. In: Proceedings of Intelligence and Security Informatics, pp. 183–184. IEEE Press, Los Alamitos (2006)
8. Simon, C., Neil, M.F., Claude, W., Anders, T., Guillaume, S., Ben, D., Angus, N.: Closure of schools during an influenza pandemic. The Lancet Infectious Diseases 9, 473–481 (2009)
9. Zhao, C.Z., Wu, Y.: Etiology Detection and Investigation of A (H1N1) influenza outbreak in a middle school. International Journal Epidemiological Infect disease 37, 37–38 (2010) (in Chinese)
10. Justin, L., Nicholas, G.R., Derek, A.T.: Outbreak of 2009 Pandemic Influenza A (H1N1) at a New York City School. The New England Journal of Medicine 361, 2628–2636 (2009)

11. Cao, Z.D., Zeng, D.J., Song, H.B.: Scientific Problem of Emergency Management for the Influenza A (H1N1) Cluster Outbreak Incidents. In: International Symposium on Emergency Management, Beijing, pp. 187–283 (2009)
12. Stephen, E., Hasan, G., Anil Kumar, V.S.: Modeling disease outbreaks in realistic urban social networks. Nature 429, 180–183 (2004)
13. Simon, C., Achuyt, B., Tiffany, L.M.: Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proceeding of the National Academy of Sciences of the United States of America, 2825–2830 (2011)
14. Zu, Z.H., Zhang, T., Xu, Q.: Countermeasures simulation research for pandemic influenza A (H1N1) based on dynamic social contact network. Application Research of Computers 27, 3334–3337 (2010) (in Chinese)
15. Wang, X.K., Li, Z.X., Zhong, H.M.: Mechanism of Agent Organization Structure Evolution Based on Role. Computer System & Applications 19, 57–61 (2010) (in Chinese)
16. Cao, Z.D., Zeng, D.J., Zheng, X.L.: Analyzing spatio-temporal evolution of Beijing 2003 SARS Epidemic. Science in China Series D-earth Sciences 53, 1017–1028 (2010)

Author Index