Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network

Preetish Ranjan
Indian Institute of Information Technology
Allahabad, India
rs105@iiita.ac.in

Abhishek Vaish
Indian Institute of Information Technology
Allahabad, India
abhishek@iiita.ac.in

Abstract—**Social network analysis is a basic mechanism to observe the behavior of a community in society. In the huge and complex social network formed using cyberspace or telecommunication technology, the identification or prediction of any kind of socio-technical attack is always difficult. This challenge creates an opportunity to explore different methodologies, concepts and algorithms used to identify these kinds of community on the basis of certain pattern, properties, structure and trend in their linkage. This paper tries to find the hidden information in huge social network by compressing it in small networks through apriori algorithm and then diagnosed using viterbi algorithm to predict the most probable pattern of conversation to be followed in the network and if this pattern matches with the existing pattern of criminals, terrorists and hijackers then it may be helpful to generate some kind of alert before crime.**

Keywords—*Social Network Analysis; Socio-Technical Attacks*

I. Introduction

Socio-technical attack is an organized approach which is defined by the interaction among people through maltreatment of technology with some of the malicious intent to attack the social structure based on trust and faith.

Awful advertisement over internet and mobile phones may defame a person, organization, group and brand value in society which may be proved to be fatal. People are always very sensitive towards their religion therefore mass spread of manipulated information against their religious belief may create pandemonium in the society and can be one of the reasons for social riots, political misbalance etc. Cyber attack on water, electricity, finance, healthcare, food and transportation system are may create chaos in society within few minutes and may prove even more destructive than that of a bomb as it does not attack physically but it attacks on the faith and trust which is the basic pillar of our social structure.

Trust is a belief that the person who is being trusted will do what is being expected for and it starts from the family which grows to build a society. Trust for information may be established if it either comes from genuine source or information is validated by authentic body so that there is always a feeling of security and optimism. There is always some sense of insecurity during online transaction of money that a site or portal is being used is authentic or not or whether SMSs are being received is coming from right source. Telecommunication network and cyberspace may be used interchangeably as both are very widespread and almost exhaustive and the most easily accessible platform for masses. If we want to express something to common people it is one of the very comfortable areas to explore with but it is vulnerable if it is being used maliciously. This illegal use of cyber space comes under socio-technical attack such as spam attack, social phishing attack, automated social engineering, bulk SMSs and different malwares. Now, even the hardware and software developers are trying to develop components by keeping in the mind all the changes they are going to bring about in the information flow.

II. Literature Review

After the very exhaustive survey about the socio technical attacks, it is being found that there are lots of paper dealing with relationship between society and technology, different kind of cyber attacks, social informatics based on principle, theory and practices but there are very few papers focusing on the prior detection of viral attacks on our social structure based on trust and faith through telecom and internet technology. In 2012, Katharine et al. proposed a framework based on perception and applied a theory of resilience for business security which may predict and prevent organizational threats. They projected six resilient factors as culture, procedures, goals/visions/values, people, infrastructure and technology and tested on cyber system of 2012 London Olympic and poisoning of the UK's water supply in order to anticipate the key contributors to these threats [1]. A paper by Robin Gandhi et al. explained about few cases related cyber attacks which are politically and economically motivated due to dissatisfaction from policies or socio-cutural disputes and also briefed about different attack agents, origin, coordinators, timing, victims as well as consequences [2]. A paper published in 2011 focused on the trustworthiness and effectiveness of communications on cyber-security risks. Trustworthiness depends on source, having objectivity, identity, ethics, recommendation, reputation and some motivational value, information having accuracy, believability, consistency, relevancy and citation, end user should be contextual, familiar and propensity to take risk [3]. Paper published in Springer 2005, revealed that open source software developer are now considering both social and material

978-1-4799-7012-4/14 $31.00 © 2014 IEEE
DOI 10.1109/EnT.2014.11

aspect associated with the project. Software is being designed on the basis of usability and perspective of particular group of people in the society [4]. Markus Huber et al. discussed about several socio-technical attacks in social networking sites such as social-phishing and social engineering bots which used personal information to launch the attack [5]. There are several examples of socio-technical disaster and public place failure such as Bhopal, Tylenol, Chernobly, Challenger and Exxon which collapsed due to breakdown of computer system involved production, transportation and maintenance of infrastructure. Most of the article responded to the theoretical aspect of socio-technical but very few papers discussed about the mathematical model to solve the problem to some extent. This paper tried to tackle this issue by simulating the problem through CDR as telephonic interaction form stronger network than that of online social network, therefore, researchers are trying to analyze this telephonic social network to quantify reciprocity in social network, to predict the social tie strength and calculate willingness level of receiver. Some people are trying to carve a way from CDR to identify the cases of fraudulent usage, abnormal behavior patterns but others are interested to find the outliers on the basis of characteristics of the shape of the time series.

After a very exhaustive literature survey, it has been observed that very little has been done to develop a model for the prior identification of socio- technical attacks in any social network. Here, the sequence of communication has been considered as input and most probable sequence of communication has been identified to trace the attack.

III. METHODOLOGY

A. Implementation of Apriori Algorithm to generate the network

Apriori is a classical algorithm for frequent item set mining and association rule learning over transactional databases where each transaction is seen as a set of items. Apriori uses “bottom up” approach where frequent subsets are extended with one item at a time and groups of candidates are tested against the data and algorithm terminates when no further successful extensions are found. There are different logics to generate a social network based on attributes of a node to qualify the requirements to solve a particular problem. The concept of apriori algorithm has been used to form a network in which the links depends on the threshold frequency compared with the frequency of two candidate numbers.

B. Implementation of HMM Viterbi Algorithm over the network

Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the Viterbi path – especially in the context of Markov information source and hidden Markov models. A Hidden Markov model is being implemented to observe the sequence of observation irrespective of the sequence of states the model went through to generate the sequence of observations. The Viterbi algorithm was proposed by Andrew Viterbi in 1967 as a decoding algorithm for convolutional codes over noisy digital communication links such as codes used in both CDMA and digital cellular, dial up modems, satellite, deep-space communication links, speech recognition, keyword spotting, computational linguistics and bioinformatics.

IV. SOLUTION

Call details record reveals certain facts about their behavior of caller and receiver and may be analyzed to demonstrate that every cell phone user is bound with certain network which is highly dynamic in relation to; the calling person, time of the call, duration of the call and frequency of the call. This above fact has been used to create a social network by applying Apriori algorithm over the CDR as follows:

Algorithm to generate the network

```
INPUT: Threshold frequency $T_h$ to generate the network of two numbers
OUTPUT: The network of callers and receivers
Function NetworkForm($T_h$)
Step1: Scan the CDR to identify the different callers and receiver numbers
Step2: Generate a set of candidate of two different numbers
Step3: Find the frequency f of each candidate of two different numbers
Step4: While f>=$T_h$
           Generate the link between two numbers
       end while
```

.

The network among callers and receivers has been generated using CDR is as follows:

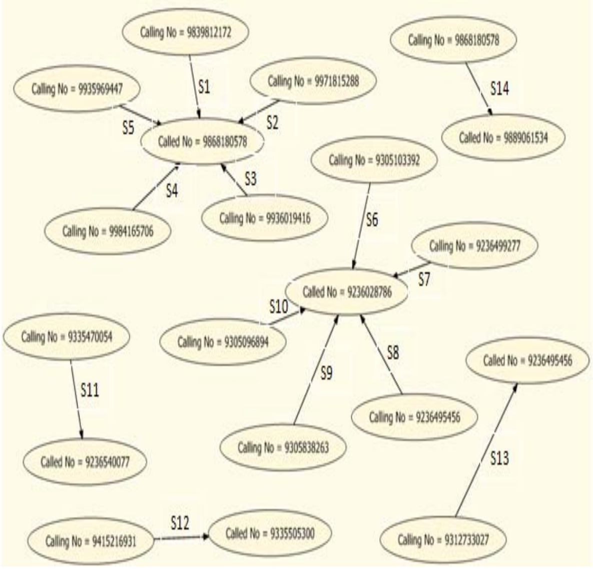
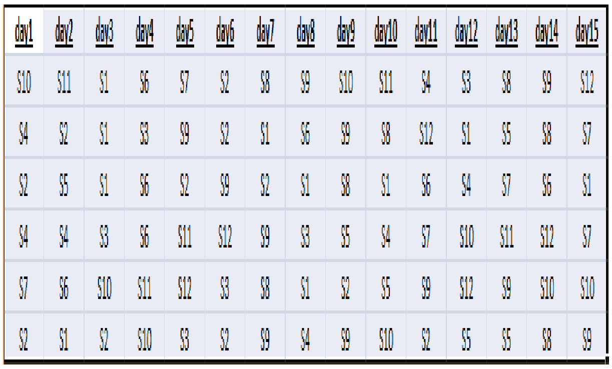


Figure.1 Network with link marked with different states from S1 to S14

Each edge of the network has been marked as different states of the network which represents communication between particular caller and receiver. Links assumed to be states have been marked from S1 to S14 and all these links have been observed till 15 days. Following data sheet represents the sequence of state per day at different point of time.

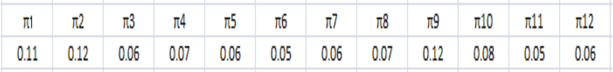


day1	day2	day3	day4	day5	day6	day7	day8	day9	day10	day11	day12	day13	day14	day15
S10	S11	S1	S6	S7	S2	S8	S9	S10	S11	S4	S3	S8	S9	S12
S4	S2	S1	S3	S9	S2	S1	S6	S9	S8	S12	S1	S5	S8	S7
S2	S5	S1	S6	S2	S9	S2	S1	S8	S1	S6	S4	S7	S6	S1
S4	S4	S3	S6	S11	S12	S9	S3	S5	S4	S7	S10	S11	S12	S7
S7	S6	S10	S11	S12	S3	S8	S1	S2	S5	S9	S12	S9	S10	S10
S2	S1	S2	S10	S3	S2	S9	S4	S9	S10	S2	S5	S5	S8	S9

Figure.2. Different states in different days

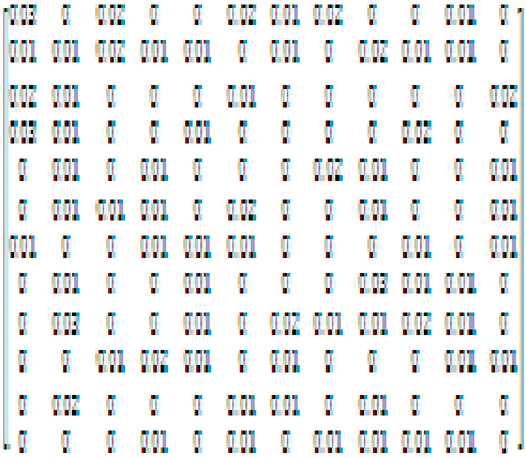
The initial probability for a state = Frequency of state / Total number of states

The initial probability for S1 = 10/90 = 0.11
The initial probability for S2 = 11/90 = 0.12
The initial probability for S3 = 6/90 = 0.06
…………



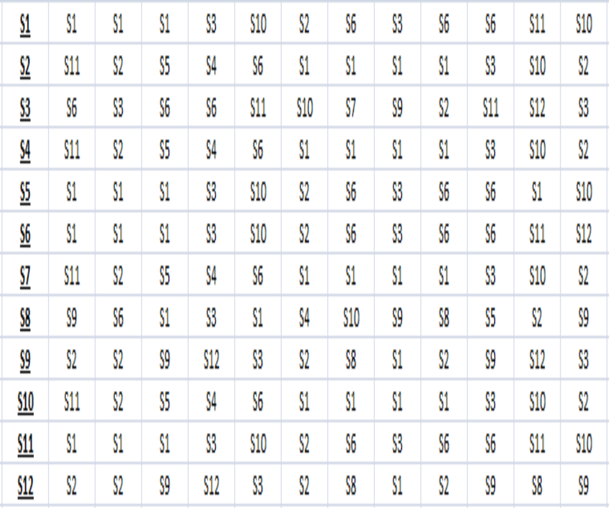
π1	π2	π3	π4	π5	π6	π7	π8	π9	π10	π11	π12
0.11	0.12	0.06	0.07	0.06	0.05	0.06	0.07	0.12	0.08	0.05	0.06

In the table shown in Figure.2 have columns from day1 to day15 in which column day1 has sequence of states as S10-S4-S2-S4-S7-S2, similarly column day2 has S11-S2-S5-S4-S6-S1 and so on. The transition matrix T of size 12x12 is being formed where T_{ij} = Total number of transitions from state S_i to S_j / Total number of transitions and 12 is total number of states being considered, hence,
$T_{11} = 3/90 = 0.033$
$T_{12} = 0/90 = 0$
$T_{13} = 0/90 = 0$
…



$$T = \begin{bmatrix} 0.03 & 0 & 0.02 & 0 & 0 & 0.02 & 0.01 & 0.02 & 0 & 0 & 0.01 & 0 \\ 0.01 & 0.01 & 0.02 & 0.01 & 0.01 & 0 & 0.01 & 0 & 0.02 & 0.01 & 0.01 & 0 \\ 0.02 & 0.01 & 0 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0.02 \\ 0.03 & 0.01 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0.01 & 0 & 0.01 & 0 & 0 & 0 & 0.02 & 0.01 & 0 & 0 & 0.01 \\ 0 & 0.01 & 0.01 & 0.01 & 0 & 0.02 & 0 & 0 & 0.01 & 0 & 0 & 0.01 \\ 0.01 & 0 & 0 & 0.01 & 0.01 & 0.01 & 0 & 0 & 0 & 0.01 & 0 & 0.01 \\ 0 & 0.01 & 0 & 0 & 0.01 & 0 & 0 & 0 & 0.03 & 0.01 & 0.01 & 0 \\ 0 & 0.03 & 0 & 0 & 0.01 & 0 & 0.02 & 0.01 & 0.01 & 0.02 & 0.01 & 0 \\ 0 & 0 & 0.01 & 0.02 & 0.01 & 0 & 0.01 & 0 & 0 & 0 & 0.01 & 0.01 \\ 0 & 0.02 & 0 & 0 & 0 & 0.01 & 0.01 & 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.01 & 0 & 0.01 & 0 & 0.01 & 0.01 & 0.01 & 0.01 & 0 \end{bmatrix}$$

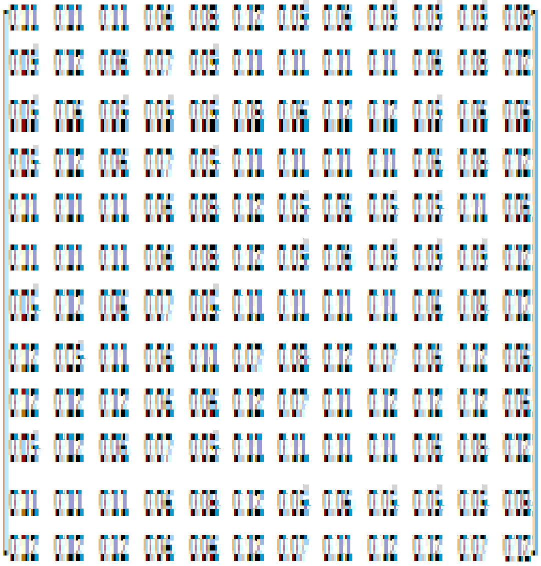
Now, the sequence of states after a particular state has been observed from the Figure.2 as follows:
S1-S1-S1-S3-S10-S2-S6-S3-S6-S6-S11-S10 is the sequence after the state S1
S4-S7-S2-S11-S2-S5-S4-S6-S1-S1-S1-S1 is the sequence after the state S2
S10-S2-S6-S3-S6-S6-S11-S10-S7-S9-S2-S11 is the sequence after the state S3
…………….



S1	S1	S1	S1	S3	S10	S2	S6	S3	S6	S6	S11	S10
S2	S11	S2	S5	S4	S6	S1	S1	S1	S1	S3	S10	S2
S3	S6	S3	S6	S6	S11	S10	S7	S9	S2	S11	S12	S3
S4	S11	S2	S5	S4	S6	S1	S1	S1	S1	S3	S10	S2
S5	S1	S1	S1	S3	S10	S2	S6	S3	S6	S6	S1	S10
S6	S1	S1	S1	S3	S10	S2	S6	S3	S6	S6	S11	S12
S7	S11	S2	S5	S4	S6	S1	S1	S1	S1	S3	S10	S2
S8	S9	S6	S1	S3	S1	S4	S10	S9	S8	S5	S2	S9
S9	S2	S2	S9	S12	S3	S2	S8	S1	S2	S9	S12	S3
S10	S11	S2	S5	S4	S6	S1	S1	S1	S1	S3	S10	S2
S11	S1	S1	S1	S3	S10	S2	S6	S3	S6	S6	S11	S10
S12	S2	S2	S9	S12	S3	S2	S8	S1	S2	S9	S8	S9

Figure.3 represents all the sequences after the particular state.

Emission matrix E of size 12x12 is being configured in which each element contains the initial probability of state present in Figure.3,hence,



$$E = \begin{bmatrix} 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 & 0.05 & 0.06 & 0.05 & 0.05 & 0.05 & 0.08 \\ 0.05 & 0.12 & 0.06 & 0.07 & 0.05 & 0.11 & 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 \\ 0.05 & 0.06 & 0.05 & 0.05 & 0.05 & 0.08 & 0.06 & 0.12 & 0.12 & 0.05 & 0.06 & 0.06 \\ 0.05 & 0.12 & 0.06 & 0.07 & 0.05 & 0.11 & 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 \\ 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 & 0.05 & 0.06 & 0.05 & 0.05 & 0.11 & 0.08 \\ 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 & 0.05 & 0.06 & 0.05 & 0.05 & 0.05 & 0.06 \\ 0.05 & 0.12 & 0.06 & 0.07 & 0.05 & 0.11 & 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 \\ 0.12 & 0.05 & 0.11 & 0.06 & 0.11 & 0.07 & 0.08 & 0.12 & 0.07 & 0.06 & 0.12 & 0.06 \\ 0.12 & 0.12 & 0.12 & 0.06 & 0.06 & 0.12 & 0.07 & 0.11 & 0.12 & 0.12 & 0.12 & 0.06 \\ 0.05 & 0.12 & 0.06 & 0.07 & 0.05 & 0.11 & 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 \\ 0.11 & 0.11 & 0.11 & 0.06 & 0.08 & 0.12 & 0.05 & 0.06 & 0.05 & 0.05 & 0.05 & 0.08 \\ 0.12 & 0.12 & 0.12 & 0.06 & 0.06 & 0.12 & 0.07 & 0.11 & 0.12 & 0.12 & 0.07 & 0.12 \end{bmatrix}$$

The data furnished above is being applied to following algorithm:
Modified Viterbi algorithm
INPUT:

The state space $S = \{S_1, S_2, \ldots, S_k\}$, observed states at different point of time $Y = \{y_1, y_2, \ldots y_T\}$ such that if at time t the observed state is S_i then $y_t = i$

Transition matrix A of size K x K where K is total number of states and A_{ij} stores the transition probability from S_i to S_j

Emission matrix B of size K x K such that B_{ij} stores the probability of N the states given that particular state

A array of initial probabilities π of size K such that π_i stores the probability that $x_1 = S_i$

OUTPUT:

The most likely hidden state sequence $X = \{x_1, x_2, \ldots, x_T\}$

```
Function VITERBI (S, π, Y, A, B ) : X
        for each state $S_i$ do
                $T_1[i,1] = \pi_i * B_{iy1}$
                $T_2[i,1] = 0$
        end for
for i= 2 to T do
        for each state $S_j$ do
                $T_1[j,i] = \max_k(T_1[k,i-1]*A_{kj}*B_{jyi})$
                $T_2[j,i] = \arg\max_k(T_1[k,i-1]*A_{kj}*B_{jyi})$
        end for
end for
$Z_T = \arg\max_k(T_1[k,T])$
$X_T = S_{zT}$
for i=T, T-1,…..2 do
        $Z_{i-1} = T_2[Z_i,i]$
        $X_{i-1} = S_{zi-1}$
end for
return X
end function
```

V. RESULTS

Let the sequence of states that has been observed during day5 and day6 is S7-S9-S2-S11-S12-S3-S2-S2-S9-S12-S3-S2. The sequence of states as output is S2-S9-S2-S9-S2-S9-S2-S9-S10-S4-S1-S1 which implies that output sequence will be followed most in next point of time as per given input . Hence, output as a sequence implies that for a given transition matrix, emission matrix and given sequence of communication between two subscribers as input the most probable sequence of communications between different mobile numbers may be predicted.

These algorithms may be used under different perspectives and circumstances such as the safest network structure for terrorists to communicate will be a network in which all the nodes exchange information with only two neighboring nodes so that there is less chance for the information to get compromised. This kind of situation may be identified and sequence of exchange of information among different users may be predicted by passing the value of $T_h = 1$ to the first algorithm to generate the network. Once the network has been formed then by evaluating the TRANS and EMIS matrices for that network and passing both as a parameter we may get the sequence of links formed in the network. This sequence of links may be network used by the terrorists to pass the information from one to another which is very difficult to get detected. After getting the CDR of 19 hijackers of 9/11 attack, a network with some pattern may be mapped in to matrix of state transition probabilities which may be used to identify other such network and predict other future moves. A large social network may be a collection some small networks which may be connected or disconnected but if want to know the effect of one network on other network and want to predict which sub-network initiates the other sub-network then these two algorithms may be implemented to find which network triggers the other network to become active. This type of study will be very fruitful and beneficial in identifying potential "criminals" and their network of criminal contacts by detecting any unusual pattern of communication or matching this pattern with pre-observed pattern of criminal activities.

VI. CONCLUSION

Due to emergence of internet on mobile phone, the different social networks such as on social networking sites, blogs, opinion, ratings, review, serial bookmarking, social news, media sharing, Wikipedia led the people to disperse any kind of information very easily. Rigorous analysis of these patterns can reveal some very undisclosed and important information explicitly whether that person is conducting malignant or harmless communications with a particular user and may be a reason for any kind of socio technical attacks. From the above simulation done on CDR, it may be concluded that if this kind of simulation applied on networks based on the internet and if we are in the position to get the data which could be transformed in transition and emission matrix then several kind of prediction may be drawn which will be helpful to take our decisions.

REFERENCES

[1] Katharine E Worton, "Using Socio-Technical and Resilience frameworks to Anticipate Threat", Workshop on Socio-Technical Aspects in Security and Trust, Socio-Technical Center, University of Leeds, UK 2012.

[2] Robin Gandhi, Anup Sharma, William Mahoney, William Sousan, Qiuming Zhu and Phillip Laplante, "Dimension of Cyber-Attacks Social, Political, Economic and Cultural", IEEE Technology and Society Magazine,2011.

[3] Jason R. C. Nurse, Sadie Creese, Michael Goldsmith, Koen Lamberts, "Trustworthy and Effective Communication of Cybersecurity Risks: A Review", University of Warwick, Coventry, CV4 7AL, UK.

[4] Nicolas Ducheneaut, "Socialization in an Open Source Software Community: A Socio-Technical Analysis", Computer Supported Cooperative Work, Springer, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA, 94304, USA, 2005.

[5] Markus Huber, Martin Mulazzani, Sebasttian Schrittwieser, Edgar Weippl, "Cheap and Automated Socio-Technical Attacks based on Social Networking Sites", SBA Research, Favoritenstrasse 16, AT-1040 Vienna, Austria, ACM 2010.

[6] Huiqi Zhang and Ram Dantu, "Predicting Social Ties in Mobile phone Networks", IEEE International Conference on Intelligence and Security Informatics (ISI), Vancouver, Canada, pp. 25 - 30, 2010.

[7] Xu Ye, “The Application of ARIMA Model in Chinese Mobile User Prediction”, IEEE International Conference on Granular Computing , San Jose, CA , pp. 586 – 591, 2010.

[8] S. Phithakkitnukoon, H. Husna and R. Dantu, “Behavioral Entropy of Cellular Phone User”, Social Computing, Behavioral Modeling, and Prediction, pp. 160–167, 2008.

[9] Hui Zang, Jean C. Bolot, Sprint, California, USA, “Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Network”, ACM 2007.

[10] ZhaoHui Tang and Jamie MacLennan, Data Mining with SQL Server 2005, pp. 230 – 234, 2005.

[11] Wang Mingsong, “Mobile Users Predict Method Analysis”, Telecom Engineering Technics and Standardization, Beijing Consulting and Design Institute of P & T, China, pp. 19 – 22 2004.

[12] Ben Congleton, Satyendra Nainwal, “Mining the Mine Exploratory Social Network Analysis of the Reality Mining Dataset”, School of Information, University of Michigan, Ann Arbor, USA.

[13] N. Eagel and A. Pentland (2006), “Reality Mining: Sensing Complex Social Systems”, Personal and Ubiquitous Computing, Vol 10(4), 2006.

[14] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabsi, D. Brewer,N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King,M. Macy, D. Roy, and M. Van Alstyne, “Computational Social Science,”Science, vol 323, pp. 721–723, 2009.

[15] K. Farrahi and D. Gatica-Perez, “What did you do today? Discovering daily routines from large-scale mobile data,” in Proceeding of the 16th ACM international conference on Multimedia. ACM, pp. 849–852, 2008.

[16] Campbell, S. Eisenman, N. Lane, E. Miluzzo, and R. Peterson, “People-centric urban sensing”,2nd annual international workshop on Wireless internet. ACM, pp. 18, 2006.

[17] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. Barabsi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne, “Computational Social Science”, Science, vol. 323, pp. 721–723, 2009.