International Journal of Database Theory and Application
Vol.8, No.5 (2015), pp.285-292
http://dx.doi.org/10.14257/ijdta.2015.8.5.25

The Dynamic Influence Graph Model on Mobile Datasets

Zhipeng Liu[1], Dechang Pi[2] and Yehong Wu[3]

[1,3]Zhipeng Liu Yehong Wu College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Wenyuan Street 9, Nanjing, Jiangsu, 210023, P.R. China

[2]Dechang Pi College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Yudao Street 29, Nanjing, Jiangsu, 210016, P.R. China

[1]liuzhipengcs@139.com, [2]dc.pi@nuaa.edu.cn, [3]wuyh@njupt.edu.cn

Abstract

With the rapid development of mobile technologies, more and more people are equipped with smartphones. It is possible for scientists to collect and analyze mobile data efficiently. Mobile data contain rich semantic as well as topological information. Rich information can be inferred from these data such as social influence among different nodes in mobile social network. However, it is difficult to estimate the strength of social influence due to the characteristics of inherent dynamic and large scale of mobile social network. In this paper, a Dynamic Influence Graph (DIG) model is proposed which utilizes temporal information in a topological perspective, and an efficient algorithm is proposed based on the DIG model. The proposed algorithm can calculate social influence between any two nodes in a given mobile social network stream segment, and takes edge weights, node connectivity and temporal information into consideration. Experimental results with a real mobile social network dataset show that the proposed approach can infer social influence and achieve a-state-of-the-art accuracy (82-86%) efficiently and automatically.

Keywords: *mobile social network; social influence; influence graph; path strength*

1. Introduction

With the rapid development of mobile technologies [1-2], more and more people are equipped with mobile phones. It is reported that mobile data traffic reached 885PBs per month at the end of 2012 [3]. With the widely deployment of smartphones, it is possible for scientists to collect and analyze mobile data efficiently. Two famous mobile dataset collections are MIT reality mining project [4-5] and the Lausanne data collection campaign sponsored by Nokia [6-7]. Mobile data contain rich semantic as well as topological information. For example, the Lausanne dataset contains data of GPS, Bluetooth, WLAN, calls, SMS, pictures, videos, audio, calendar and phone book entries. The quantity of each type of mobile data is massive. For example, the Lausanne dataset contains 4,527,539 GPS points, 12,568,788 WLAN scans, 15,362,182 Bluetooth scans, 132,109 3,907h calls and 88,225 SMS. Various data types and massive data provide great opportunities as well as challenges for data scientists. Rich information can be inferred from mobile data, such as social diffusion [8], contextualized recommendation [9], online choice [10] and inferring mobile network structures [11]. All these social applications are based on the computation of social influence among different nodes based on mobile data [12-15]. Social influences are reflected by a sequence of social reactions and interactions, and mobile data provides valuable resources for characterizing social influence between different people. It is difficult to compute social influence by

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2015 SERSC

mobile data due to the characteristics of inherent dynamic and large scale of mobile social network. Most of existing social influence computation methods uses statistical methods, which relies heavily on the experience of experts. Eagle *et al*. [16-17] compare observational data collected from smartphones with standard self-report survey data, and the proposed statistical method can infer friendship network structure by an accuracy of 95%. However, the computation process is heavily depended on user experience, thus it cannot be applied to large scale mobile network. Many studies have been performed to discover social influence on other kinds of social network which might facilitate the social influence computation of mobile social network. However, these algorithms have some defects. Node-based centrality measure such as node between [18] requires $O(n^3)$ time and $O(n^2)$ storage which is computational expensive. Brandes proposed a faster between computation algorithm [19], Freeman proposed an algorithm based on network flow [20], Newman proposed an algorithm based on random walks [21]. Java *et al*. [22] employed a link analysis method to deal with blog graph influence, which is similar to the concept of random walks. However, these influence computation algorithms do not take temporal information into consideration, which plays an essential role in mobile social network analysis. In this paper, a Dynamic Influence Graph (DIG) model is proposed to compute the social influence of mobile networks. The DIG model is based on blog influence graph [22], and computes the influence between a source node and a destination node, and takes edge weights, node connectivity and temporal information into consideration. Experimental results with a real mobile social network dataset show that the proposed approach infers social influence and achieves a-state-of-the-art accuracy (82-86%) efficiently and automatically.

2. Problem Formulation and Motivation

2.1. Concepts about Social Influence

Definition 1 Mobile *Social* network stream A mobile *social* network stream $\mathbb{G}$ is a series of directed graphs G^t. $\mathbb{G} = (G^1, G^2, \ldots, G^T)$, which evolves infinitely over time. Each static directed graph G^t represents the relationships of communications among n^t nodes and e^t edges. Here, $t \in [0..T]$. In mobile social network, node v_i^t denotes the person *i* with temporal information *t*, edge e_j^t are made up of two kinds of links: call logs and SMS information, which can be used to model social influence in mobile network. Here, $i, j \in [0, n^t - 1]$. In this paper, SMS information is used to calculate social influence.

Definition 2 Mobile *Social* network stream segment The series of static graphs between time interval $[t_s, t_{s+1} - 1]$ compose the *s*th segment $\mathbb{G}^s$, $s \geq 1$. $\mathbb{G}^s = (G^{t_s}, G^{t_s+1}, \ldots, G^{t_{s+1}-1})$.

The definitions of influence graph and path strength are presented as follows.

Definition 3 Influence Graph For a specific static graph G^t, the presence of a link from node v_i^t to v_j^t is considered as the fact that the node v_i^t is influenced by v_j^t. An influence graph $IG^{(t)}$ is a weighted, directed graph for G^t with edge weights representing social influence. The series of static influence graphs between time interval $[t_s, t_{s+1} - 1]$ are denoted by $\mathbb{IG}^s = (IG^{t_s}, IG^{t_s+1}, \ldots, IG^{t_{s+1}-1})$.

Our goal is to compute social influence from node v_i^t to v_j^t on a specified influence graph IG^t in $\mathbb{IG}^s$. Definition 3 indicates that the edges in IG^t are the reverse of G^t to reflect this influence, so the direction in IG^t is opposite to G^t.

 Copyright ⓒ 2015 SERSC

Multiple edges between two nodes indicate stronger influence and have higher weights.

Definition 4 Path Strength Given a specified mobile social network stream segment G^s and two nodes v_i^t and v_j^t. If there exists a path p between node v_i^t and v_j^t, where $p = < v_i^t, v_{i+1}^t, ..., v_j^t >$, the path strength S is defined as

$$S(p) = \prod_{k=i}^{j-1} \frac{w_{k,k+1}}{d_{v_k}} \tag{1}$$

Here, $w_{k,k+1}$ denotes the weight of the edge between node v_k^t and v_{k+1}^t. d_{v_k} denotes the sum of the edge weights between v_k^t and its neighbors.

Definition 5 Social influence Social influence is defined as the cumulative path strength of all possible paths from node v_i^t to v_j^t in IG^t. Suppose there are *N* paths between them, the social influence $F(v_i^t, v_j^t)$ is defined as

$$F(v_i^t, v_j^t) = \sum_{m=1}^{N} S(p_m^t) \tag{2}$$

The social influence model considers all possible paths between two nodes including cycles, and has a directional effect.

2.2. Time Sensitive Rank (TS-Rank)

PageRank [23] is motivated by the observation that a hyperlink from one web page to another is an indication of authority transformation to the destination page. PageRank computes a score for each web page by link analysis. Given a page p with inlinks I_p and outlinks O_p, the PageRank score PR is obtained by the following equation:

$$PR_p = d \cdot \sum_{q \in I_p} \frac{PR_q}{\|O_q\|} + (1-d) \cdot E_p \tag{3}$$

Here, d is the damping factor. E denotes the random page vector selected in the web graph. PageRank is simple, effective, and robust for malicious attacks. It is particularly suitable to rank a set of web pages or people. However, PageRank does not take temporal information of links into consideration.

Yu *et al*. [24] proposed TS-Rank algorithm to integrate time factor into the ranking model. TS-Rank uses the Markov chain model and a random reader to formulate the problem. Suppose the reader will follow each reference uniformly at random, the probability of following each reference is $1/O_i$. Here, O_i is the number of out-links or references of the page. A transition probability matrix of the chain is denoted by A. This matrix is the stochastic transition matrix of a Markov chain, so

$$A_{ij} = \begin{cases} \frac{1}{O_i} & if\ (i,j) \in E \\ 0 & otherwise. \end{cases} \quad \text{and} \sum_{j=1}^{n} A_{ij} = 1 \tag{4}$$

Since a finite Markov chain defined by A has a unique stationary probability distribution if A is irreducible and aperiodic. TS-Rank uses the same technique as in PageRank, an artificial link from each state to every state with time function $f(t)$ ($0 < f(t) < 1$) is added, where t is the difference between the current time and the

Copyright ⓒ 2015 SERSC

established time of the connection. Thus $P_T = (F + H)^T P_T$, where F and H are both $n \times n$ square matrix defined by

$$F_{ij} = \frac{1 - f(t_i)}{n} \text{ and } H_{ij} = \begin{cases} \frac{f(t_i)}{O_i} & if\ (i, j) \in E \\ 0 & otherwise \end{cases} \quad (5)$$

TS-Rank uses an exponential function where $f(x) = 0.5^{t_d / x}$, t_d is the difference in month between the current time and the time when the link is established. Parameter x is set to 3 as decay parameter based on user experience in the original model.

3. Dynamic Influence Graph (DIG)

DIG model is consisted of the following two stages, which are conversion stage and computation stage.

1. The conversion stage: For a specified mobile social network stream segment $\mathbb{G}^s = (G^{t_s}, G^{t_s+1}, \ldots, G^{t_{s+1}-1})$. The corresponding influence graph $\mathbb{IG}^s = (IG^{t_s}, IG^{t_s+1}, \ldots, IG^{t_{s+1}-1})$ is constructed. Each influence graph in $\mathbb{IG}^s$ is computed by TS-Rank algorithm.

2. The computation stage: For each influence graph IG^t in $\mathbb{IG}^s$, the social influences between two nodes v_i^t and v_j^t are computed. The computation of influence is based on definition 4 and 5. From definition 4 it can be inferred that $S(p)$ increases with a higher weights of the edges on p, and decreases with a higher degrees of the vertices on p. $S(p)$ has a negative relationship with the length of p, since $0 < w_{i,i+1}^t / d_{v_i^t} < 1$. If $F(v_i^t, v_j^t)$ is less than a predefined user-independent minimum threshold σ, the influence of the two nodes is omitted.

The pseudo-code of DIG algorithm is depicted in Figure 1 based on descriptions and definitions above.

Algorithm DIG (Dynamic Influence Graph)

Input: Mobile social network stream segment $\mathbb{G}^s$, node v_i^t and v_j^t, time of influence t, a predefined user-independent minimum threshold σ.

Output: Social influence $F(v_i^t, v_j^t)$, which from node v_i^t to v_j^t on the corresponding influence graph IG^t in $\mathbb{IG}^s$.

/*Influence graph conversion phase*/

1 For each $G^{t_s} \in \mathbb{G}^s$ do

2 $IG^{t_s} \leftarrow Reverse(G^{t_s})$

3 //Convert G^{t_s} to IG^{t_s} by invert the edge direction of G^{t_s}.

4 End For

4 For each $IG^{t_s} \in \mathbb{IG}^s$

5 TS-Rank($V_i^{t_s}, V_j^{t_s}$) //Calculate influence of IG^{t_s} by TS-Rank algorithm

6 End For

/*Social influence computation phase*/

 Copyright ⓒ 2015 SERSC

7 For influence graph $IG^t \in \mathbb{G}^s$

8 *exist* := p (v_i^t, v_j^t)

//Decide if there exists a path p between node v_i and node v_j

9 If *exist*

10 Calculate $S(p)$ by definition 4

11 Calculate $F(v_i^t, v_j^t)$ by definition 5

12 End If

13 End For

14 If $F(v_i^t, v_j^t) > \sigma$

15 Output $F(v_i^t, v_j^t)$.

Figure 1. The DIG Algorithm

The uncertain two parts of time complexity of the DIG algorithm are caused by Line 2 and 8 in Figure 1. DIG uses compressed sparse row (*CSR*) structure to store $\mathbb{G}^s$ on disk, and assumes that each G^{t_s} can be loaded into main memory as a whole. So the time complexity of Line 2 is $O((n^t + e^t))$. The operation of Line 8 can be computed by breadth first search, so the time complexity is also $O((n^t + e^t))$.

4. Experimental Evaluation

The experiments and the performance results of DIG algorithm are presented in this section.

4.1. Experimental Setting

The SMS in MIT Reality dataset [4] is used for experimental analysis.

The MIT Reality project has been conducted since 2005. One hundred teachers and students have been involved in this project during one year of data collection. The number of people remained stable during the experiment. It records almost 450,000 hours of information. Reality Mining includes 18760 calls, 3590 SMS messages, and 285512 proximity interactions.

Our experiments were conducted on a Pentium(R) D 3.0 GHz PC with 2 GBytes of main memory, running on CentOS 4.5 operating system. The algorithm is implemented in C++ using GraphChi [25]. It can perform large-scale graph computation on one PC. It is based on vertex-centric model of computation.

Methods. The proposed DIG algorithm is compared with the following baseline influence graph computation methods.

1. Influence graph with PageRank [23] (IP). It employs an algorithm that a cell phone with more incoming connections is in general higher quality than a cell phone with fewer incoming links, with completely ignorance of temporal aspects.

2. Influence graph with TimedPageRank [24] (ITP). It generalizes PageRank with temporal aspects, weighting each communication considers recent communications more important.

Evaluation aspects. Six months of SMS information of MIT Reality dataset is used to evaluate the proposed algorithm. We mainly consider two performance measurements: accuracy and sensitivity. For accuracy, we compare the social influence obtained from these algorithms with self-report data. This is an objective measure. For sensitivity, we

Copyright ⓒ 2015 SERSC

analyze different parameters in DIG. Examples are listed to demonstrate how discovered social influence can benefit other applications.

4.2. Accuracy

In this set of experiments, six months of SMS data is used to perform various evaluations for the proposed methods. The social influences are computed by different algorithms and the results are compared with self-report data. The parameter σ is set to 0.1 during experiments.

Table 1. Comparison Results of Different Algorithms

Months / Algorithms	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
IP	61%	58%	56%	63%	59%	62%
ITP	74%	76%	74%	76%	77%	75%
DIG	83%	85%	82%	84%	86%	86%

Table 1 presents the experiment results. Each row gives the different approaches for a group of communication data. Each column lists the monthly time intervals. The results are explained below:

Column 1: It lists various kinds of algorithms for comparison. The original PageRank algorithm with damping factor of 0.85 is used. The original exponentially decay of weights with $f(t) = 0.5^{(y-t_i)/12}$ is used for TimedPageRank, in which y is the current time, t_i is the start time of the communication event and (y-t_i) is the time gap in months. In TS-Rank algorithm, the decay parameter x of $f(x) = 0.5^{t_d/x}$ is selected empirically, and parameter x is set to 3 throughout experiments.

Column 2-7: They list the percentage of correct prediction with self-report data month by month. Since PageRank in IP does not take temporal aspects into consideration, the prediction rate is low. The results of DIG algorithm are in generally better than ITP. With smaller decay parameter, the weights of temporal aspects decrease rapidly with time. The results of DIG are better than IP; the reason behind this phenomenon is that social influences of mobile communication are dynamic. The social influence decreases dramatically with elapsed time.

4.3. Sensitivity

In this set of experiments, six months of SMS information of MIT Reality dataset is used to perform various evaluations for the proposed methods. Social influences are computed by DIG algorithm with different value of parameter x in TS-Rank and the results are compared with self-report data.

Table 2. Comparison Results of different Parameter x in DIG

Months / Components	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6
x =1	77%	76%	78%	75%	77%	79%
x =2	83%	80%	82%	83%	85%	82%
x =3	84%	83%	88%	86%	89%	88%
x =6	80%	80%	82%	83%	83%	82%
x =12	74%	72%	76%	73%	74%	75%

 Copyright ⓒ 2015 SERSC

Table 2 presents the experiment results. Each row gives the different parameter x for DIG algorithm. Each column lists the monthly time intervals. The results are explained below:

Column 1: It lists various kinds of parameter x for comparison. There is a lot of x which can be chosen to compute social influence. The results of other parameter x are omitted for simplicity.

Column 2-8: They list the percentage of correct prediction with self-report data month by month. Experimental results show that better performance are achieved when x =3.

5. Conclusions

In this paper, a novel problem of mining social influence on dynamic mobile social networks is proposed. The proposed DIG algorithm is composed of two steps: the social influence graph conversion step and the social influence computation step. Compared with previous methods with blog influence, our algorithm takes the temporal aspects of social network into consideration, and combines influence graph generation and computation into a unified generative process. Experimental results with a real mobile social network dataset show that the proposed approach infers social influence and achieves a-state-of-the-art accuracy (82-86%) efficiently and automatically.

Acknowledgements

This research is sponsored by NUPTSF (Grant No. 49) of Nanjing University of Posts and Telecommunications.

References

[1] Z. Cai, S. Wen and L. Liu, "Dynamic cluster member selection method for multi-target tracking in wireless sensor network", Journal of Central South University: Science and Technology, vol. 21, no. 1, **(2014)**.
[2] X. Zhao, Y. Zhuang and J. Wang, "Local adaptive transmit power assignment strategy for wireless sensor networks", Journal of Central South University: Science and Technology, vol. 19, no. 2, **(2012)**.
[3] Cisco Technology, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update", Cisco Public Information, US, Los Angeles, **(2013)**.
[4] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems", Personal and ubiquitous computing, vol. 10, no. 4, **(2006)**.
[5] A. Pentland, "Reality mining of mobile communications: Toward a new deal on data", The Global Information Technology, vol. 3, no. 11, **(2009)**.
[6] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, O. Dousse, J. Eberle and M. Miettinen, "The mobile data challenge: Big data for mobile computing research", Proceedings of International Conference on Pervasive Services; Zurich, Switzerland, July 29th-August 3rd, **(2012)**.
[7] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez and J. Laurila, "Towards rich mobile phone datasets: Lausanne data collection campaign", Proceedings of International Conference on Pervasive Services; Graz, Austria, August 17-23, **(2010)**.
[8] A. Madan and A. Pentland, "Modeling Social Diffusion Phenomena using Reality Mining", Human Behavior Modeling, vol. 11, no. 2, **(2009)**.
[9] J. J. Jung, H. Lee and K. S. Choi, "Contextualized recommendation based on reality mining from mobile subscribers", Cybernetics and Systems: An International Journal, vol. 40, no. 2, **(2009)**.
[10] H. Zhu and B. A. Huberman, "To Switch or Not To Switch Understanding Social Influence in Online Choices", American Behavioral Scientist, vol. 7, no. 3, **(2014)**.
[11] N. Eagle and A. S. Pentland, "Eigen behaviors: Identifying structure in routine", Behavioral Ecology and Sociobiology, vol. 63, no. 7, **(2009)**.
[12] R. B. Cialdini and M. R. Trost, "Social influence: Social norms, conformity and compliance", vol. 1, no. 2, **(1998)**.
[13] J. C. Turner, "Social influence", Thomson Brooks Publishing Co, US, **(1991)**, pp. 13-31.
[14] D. Dahl, "Social Influence and Consumer Behavior", Journal of Consumer Research, vol. 40, no. 2, **(2013)**.

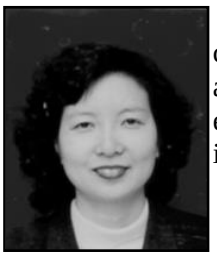
Copyright ⓒ 2015 SERSC

[15] S. Dewan, Y. J. Ho and J. Ramaprasad, "Quantifying Social Influence in an Online Music Community", Proceedings of the International Conference on Information Systems, Hong Kong, China, July 7-9, **(2013)**.
[16] N. Eagle, A. S. Pentland and D. Lazer, "Mobile phone data for inferring social network structure", Social computing, behavioral modeling, and prediction, Springer, US, **(2008)**, pp. 79-88.
[17] N. Eagle, A. S. Pentland and D. Lazer, "Inferring friendship network structure by using mobile phone data", Proceedings of the National Academy of Sciences, vol. 106, no. 36, **(2009)**.
[18] L. C. Freeman, "A set of measures of centrality based on between. Sociometry", vol. 40, no. 1, **(1977)**.
[19] U. Brandes, "A faster algorithm for between centrality", Journal of Mathematical Sociology, vol. 25, no. 2, **(2001)**.
[20] L. C. Freeman, S. P. Borgatti and D. R. White, "Centrality in valued graphs: A measure of between based on network flow", Social networks, vol. 13, no. 2, **(1991)**.
[21] M. E. Newman, "A measure of between centrality based on random walks", Social networks, vol. 27, no. 1, **(2005)**.
[22] A. Java, P. Kolari and T. Finin, "Modeling the spread of influence on the blogosphere", Proceedings of international World Wide Web, Edinburgh, Scotland, May 22-26, **(2006)**.
[23] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", Computer networks and ISDN systems, vol. 30, no. 1, **(1998)**.
[24] P. S. Yu, X. Li and B. Liu, "Adding the temporal dimension to search-A case study in publication search", Proceedings of International Conference on Web Intelligence, Compiegne, France, September 19-22, **(2005)**.
[25] A. Kyrola, G. E. Blelloch and C. Guestrin, "GraphChi: Large-Scale Graph Computation on Just a PC", Proceedings of Operating Systems Design and Implementation, Hollywood, CA, October 8-10, **(2012)**.

Authors

Zhipeng Liu, was born in 1980. He received the B.S. and M.S degrees in computer science and technology from Nanjing University of Posts and Telecommunications (NUPT), in 2002 and 2005, respectively. Now, he is a teacher in software engineering department of NUPT. His research interest is data mining.

Dechang Pi, was born in 1971. He was a Ph.D. of Nanjing University of Aeronautics and Astronautics (NUAA) of China and now he is a professor and Ph.D. supervisor in NUAA. His research interests are data mining and database systems.



Yehong Wu, was born in 1966. She received M.S degrees in computer science and technology from Nanjing University of Posts and Telecommunications (NUPT). Now, she is a teacher in software engineering department of NUPT. Her research interest is information security.

 Copyright ⓒ 2015 SERSC