Comput Math Organ Theory (2013) 19:569–579
DOI 10.1007/s10588-012-9142-x

MANUSCRIPT

Mining social networks using wave propagation

Xiaojie Wang · Hong Tao · Zheng Xie · Dongyun Yi

Published online: 31 October 2012

Abstract With the development of modern technology(communication, transportation, etc.), many new social networks have formed and influenced our life. The research of mining these new social networks has been used in many aspects. But compared with traditional networks, these new social networks are usually very large. Due to the complexity of the latter, few model can be adapted to mine them effectively. In this paper, we try to mine these new social networks using Wave Propagation process and mainly discuss two applications of our model, solving Message Broadcasting problem and Rumor Spreading problem. Our model has the following advantages: (1) We can simulate the real networks message transmitting process in time since we include a time factor in our model. (2) Our Message Broadcasting algorithm can mine the underlying relationship of real networks and represent some clustering properties. (3) We also provide an algorithm to detect social network and find the rumor makers. Complexity analysis shows our algorithms are scalable for large social network and stable analysis proofs our algorithms are stable.

Keywords Social network · Wave propagation · Message broadcasting · Rumor spreading

1 Introduction

In social network analysis (SNA) (Pinheiro 2011), the social relationships are researched by network theory. Nodes are used to represented individual actors. Connections or links are used to represented relationships between the individuals, such

X. Wang · H. Tao · Z. Xie (✉) · D. Yi
Mathematics and System Science, College of Science, National University of Defense Technology, Changsha 410073, China
e-mail: lenozhengxie@yahoo.com.cn

X. Wang
e-mail: wangxiaojie0817@yahoo.cn

as friendship, kinship, organizational position, sexual relationships, etc. Compared with traditional social network, modern social network has more form, like websites networks, blog networks, citation networks, collaboration networks, product co-purchasing networks, road networks, e-mail networks, etc. Although SNA has been studied a lot, mining the underlying relationship of the network is still difficult.

In recent years, social marketing techniques has been used to increase brands or products awareness (Hartline et al. 2008; Kempe et al. 2003; Domingos and Richardson 2001; Leskovec et al. 2007; Richardson and Domingos 2002). Comparing with traditional marketing techniques, word-of-mouth marketing is becoming more and more useful in recent year (Brown and Reinegen 1987; Goldenberg et al. 2001; Mahajan et al. 1999), which also needs to understand the social network more clearly.

In this paper, we try to model the message transmitting process in social network as wave propagation process. In a social network, each node indicates a candidate (person, animal, website etc.), and each edge is defined as the relationship between them. We let the earliest message sources act as the wave sources and model the message transmitting process as wave propagation process. The wave propagation process provides our work with the following contributions: (1) As the wave propagation process has time-dependent property, our model can simulate the message transmitting process step by step. (2) Message broadcasting algorithm can choose suitable persons to broadcast a message all over the network. (3) For there are many rumors in the world, we give a rumor transmitting algorithm to find the initial rumor makers.

The paper is organized as follow. In Sect. 2 we talk about some related work in mining social networks. In Sect. 3 we discuss the wave propagation on networks more clearly. Section 4 denotes two kinds of applications of our model, including message broadcasting problem and rumor transmitting problem. In Sect. 5 we discuss the complexity and stability of our two algorithms. Section 6 gives the empirical results of the two. Finally, Sect. 7 gives our conclusion.

2 Related work

In Rogers's book "Diffusion of Innovations" Rogers (2003), the concept of diffusion in the social network was first proposed. With successive grouping of consumers adopting the new technology, Rogers makes his well-known S curve, as early adopters adopt an innovation first, then early majority, late majority, and the laggards adopt it last.

With the rapid development of the Internet industry, the great need of searching give birth to PageRank algorithm (Brin and Page 1998), which can measure the relative importance for the element of a hyperlinked set of documents. Though PageRank has been used successfully, it is something complex.

In some recently studied, heat diffusion process (Ma et al. 2008) has been used in the study of mining social networks. They use heat diffusion process to simulate the diffusion of the message and can model the diffusion of messages well. But physicians have proved that the inverse process of heat diffusion process is unstable, it limits the use of heat diffusion method in mining social networks.

3 Wave propagation on social networks

3.1 Propagation on undirected social networks

An undirected social network can be described as a graph $G = (V, E)$, where V is the vertex set representing persons in the network, and $V = \{v_1, v_2, \ldots, v_n\}$. E is the set of edges representing the connections between persons, and $E = \{(v_i, v_j) | \text{an edge from } v_i \text{ to } v_j \text{ exists}\}$.

Here we makes an analogy, the message transmitting process is similar to the wave propagation process. Then the message sensitivity of each person in the social network can be viewed as an amplitude function $f_i(t)$ associated to each node in the graph. The value $f_i(t)$ describes the amplitude of node v_i at time t, which begins at time zero with the initial distribution of amplitudes given by $f_i(0)$. The $F(t) = [f_1(t), f_2(t), \ldots, f_n(t)]^T$ denotes the amplitude vector function with all the $f_i(t)$ as its constituents.

By using the analogy above, the message transmitting process can be viewed as an wave propagation process in some special area, i.e. the network. We build the model as follow. At time t, each node i receives an amount of amplitude from all its neighbor j during a time step Δt. To simplify the problem, we notice the received amplitude of node i should be proportional to the time step Δt and the amplitude differences between it and all its neighbors $f_j(t) - f_i(t)$. Based on this simplification, we assume the receiving amplitude of node i from node j at time t is $c^2(f_j(t) - f_i(t))\Delta t$, where c is the wave velocity and Δt is the time step. As the total amount of amplitude received by node i is from all its neighbors, we can write it as:

$$\frac{\partial^2 f_i(x,t)}{\partial t^2} = c^2 \sum_{j:(v_i,v_j)\in E} \big(f_j(t) - f_i(t)\big) \tag{1}$$

where $f_i(t)$ is the amplitude of node i at time t, and E is the set of edges.

To find a closed solution of (1), we express it in a matrix form:

$$\frac{\partial^2 F(t)}{\partial t^2} = c^2 H F(t) \tag{2}$$

where $F(t) = [f_1(t), f_2(t), \ldots, f_n(t)]^T$ and H is the discrete form of Laplace-Beltrami operator:

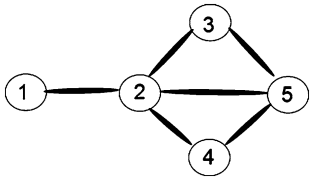
$$H_{ij} = \begin{cases} 1: & (v_i, v_j) \in E \text{ or } (v_j, v_i) \in E \\ -d_i: & i = j \\ 0: & \text{else} \end{cases} \tag{3}$$

While (2) is a simple form, it is hard to calculate. We here make an approximation method and derive an iterative form.

For the second derivative of function $f_i(t)$, we can make an approximation:

$$\frac{\partial^2 f_i(x,t)}{\partial t^2} = \frac{f_i(t+\Delta t) - 2f_i(t) + f_i(t-\Delta t)}{\Delta t^2} + O\big(\Delta t^2\big) \tag{4}$$

Fig. 1 Nodes connections



Substitute (4) into (1) and get:

$$f_i(t+\Delta t) - 2f_i(t) + f_i(t-\Delta t) = c^2\Delta t^2 \sum_{j:(v_i,v_j)\in E} \big(f_j(t) - f_i(t)\big) \tag{5}$$

Using the approximation above, we derive an iterative form:

$$F(t+\Delta t) = c^2\Delta t^2 H F(t) + 2F(t) - F(t-\Delta t) \tag{6}$$

where d_i denotes the degree of node v_i.

In order to describe the wave propagation process more clearly, we set an example here. The graph in the example includes five nodes (see Fig. 1). Initially, suppose node 1 is the source with amplitude 1, then the vector $F(0) = (1, 0, 0, 0, 0)^T$. The matrix H is:

$$H = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & -4 & 1 & 1 & 1 \\ 0 & 1 & -2 & 0 & 1 \\ 0 & 1 & 0 & -2 & 1 \\ 0 & 1 & 1 & 1 & -3 \end{pmatrix}$$

The graph is shown in Fig. 1.

Without loss of generality, we set the wave velocity $c = 1$, and the time step Δt to be 0.01. Then the amplitude of each node is shown in Fig. 2.

Now we depict Fig. 1 and Fig. 2 in the view of message transmitting process in social network. Node 1 is the source, i.e. the message publisher, node 2 is the acquaintance of node 1 and other nodes are all acquaintances of node 2. At first, the amplitude of node 2 increases quickly and that of node 1 decreases, indicating node 1 sends a message to 2. Then 2 send it to others. The figure shows the trend of message sensitivity with time among these persons.

3.2 Propagation on directed social networks

In many real social networks, the status between two persons is not equivalent. For example, in army, lieutenant can command solder, while the opposite is impossible, i.e. the connection is single-directed. Similar to the analysis at Sect. 3.1, we can adopt the wave equation to discuss the propagation on directed social networks.

For a directed graph $G(V, E)$, any edge (v_i, v_j) presents a single-directed connection from node v_i to node v_j. At time t, each node v_i can receive amplitude from all its neighbors who have a connection to it. Supposing this process has three characters: (1) the receiving should be proportional to the time step Δt; (2) the receiving should

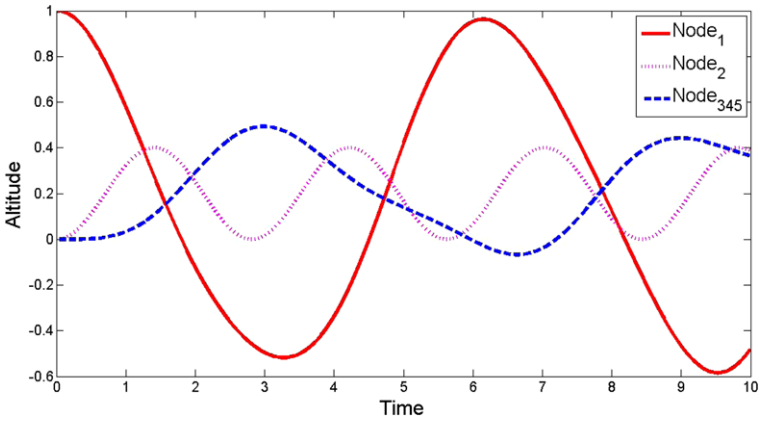


Fig. 2 Amplitude of all nodes

be proportion to the amplitude of its neighbors; (3) if there's no neighbor, the receiving is 0. According to the assumptions, the receiving is $c^2 \sum_{j:(v_j,v_i)\in E} \theta_j f_j(t)\Delta t$, where θ_j is the transmitting percentage coefficient of node v_j.

For the sending, we also suppose it satisfies the following assumptions: (1) the sending should be proportional to the time step Δt; (2) the sending should be proportional to the current amplitude of node v_i; (3) each node has the same ability to transmit which denotes $\theta_j = \frac{1}{d_j}$. According to the assumptions, the sending is $c^2 f_i(t)\Delta t$, and each of its neighbor receives $\frac{c^2 f_i(t)\Delta t}{d_i}$. Thus we get:

$$\frac{\partial^2 f_i(x,t)}{\partial t^2} = c^2\left(-\beta_i f_i(t) + \sum_{j:(v_j,v_i)\in E} \frac{f_j(t)\Delta t}{d_j}\right) \tag{7}$$

where β_i is a sign function to judge whether node v_i has any out-links: if has, $\beta_i = 1$, else $\beta_i = 0$.

Similar to the analysis of undirected social network in Sect. 3.1, we can derive the iterative form:

$$F(t+\Delta t) = c^2 \Delta t^2 H F(t) + 2F(t) - F(t-\Delta t) \tag{8}$$

where H satisfies:

$$H_{ij} = \begin{cases} \frac{1}{d_j} & (v_j, v_i) \in E \\ -\beta_i & i = j \\ 0 & \text{else} \end{cases}$$

4 Underlying relationships finding

In every social network, the underlying relationships are usually hidden but important. For example, in an e-mail network, there may be some key persons who always

send e-mails to others. In some aspects, if we can know who these persons are, we can make use of them to send e-mails and influence other company members more efficiently. Though we have mastered all the connections between each pairs of persons, we don't know who the key persons are. To finding these key persons, we must know all the underlying relationships in the network first, i.e. the underlying influencing scope of all the persons.

Seeing all the persons as nodes and the connections between each other as edges, we can build the social network. By associating influencing node sets to each node, the underlying influencing scope of persons indicates all these sets. Then our task is to find these sets, and that can be done as follow: (1) set an initial amplitude on node v_i and others 0; (2) simulate the wave propagation process in the network; (3) after a long time, statistic the number of nodes which v_i successfully influences. Where successfully influence for a node indicates the amplitude of it is higher than a given adoption threshold γ during the simulation.

4.1 Message broadcasting

The Message Broadcasting problem often comes from real situations. For example, in a company, an important decision must be broadcasted to all the members as soon as possible. This thing is usually done as follow: the CEO inform some directors, and the directors inform their managers, and then the managers inform others. For the CEO, what he should do is just selecting k well-chosen directors to broadcast his decision. And how to choose these directors is the problem.

This can be seen as an application of underlying relationship finding. First, we find the influencing node sets of all the nodes. In this process, we set an adoption threshold γ. If at some time t, the amplitude of node j is greater than γ, then we called node j has been successfully influenced. Then we choose k nodes whose influencing node sets is the biggest, and they're our "well-chose". Note that the social network we study here has finite members, so no boundary condition is necessary. We use greedy algorithm (Fig. 3) to deal with this problem.

4.2 Rumor spreading

The Rumor Spreading problem is arising more and more attention this decade (Fountoulakis and Panagiotou 2010), especially since the tragedy of 9.11. After 9.11, more and more rumors came up and disturb our normal life, which should be controlled. When a rumor come up, who made it at first is always important. That is, we want to know who is the rumor maker.

After modeling the message transmitting process as wave propagation process (see Sect. 3), this problem can be described as: given current state of amplitude in the network, we want to know what the initial state is. We can use the inverse process of wave propagation process to study this problem.

For a undirected graph, according to (6), we can solve the inverse process by:

$$F(t-\Delta t) = c^2 \Delta t^2 H F(t) + 2F(t) - F(t+\Delta t) \tag{9}$$

Given the current state of amplitude and then this iterative form can solve the inverse process. Here boundary condition is also unnecessary. But one problem is that

Input
A social network $G(V, E)$; number k; adoption threshold γ
Output
k nodes
for node i **do**
$F(0) = 0$ $f_i(0) = 1$;
Execute the wave propagation for T time period
for node $j! = i$ **do**
if $|(f_j(t))| \geq \gamma$ for any t in T **then**
Add node j into the influencing node set S_i
end if
end for
end for
$R = \emptyset$
for $i = 1 : k$ **do**
Choose S_i to maximize $\{S_i - R \cap S_i\}$
$R = R \cup S_i$
Output node i
end for

Fig. 3 Flowchart of Message Broadcasting problem

we don't know how long the wave has already propagated. In the view of calculation, we don't know when to terminate.

To decide the termination criterion, we make a natural assumption that when a rumor came at first, the number of the persons who knew it is the least. And in the iterative process, we can do this by adding an source-like threshold σ to judge the source node. If at some time t the amplitude of a node i is greater than σ, then it may be a source. Then what we should do is to see when the number of source-like nodes is least and terminate our iteration.

5 Complexity and stability

Supposing a social network has N nodes and M edges, we discuss the complexity of each algorithm:

1. Message Broadcasting
   For this problem, our algorithm has two part. First, we simulate the wave propagation process and find the influencing node sets of each node, which costs $O(N^2T)$, where T is the simulation time. Then we use the greedy algorithm to choose k nodes. In greedy algorithm, this part just cost linear time $O(kN)$. So the total complexity is $O(N^2T + kN)$.
2. Rumor Spreading
   The Rumor Spreading problem needs us to solve the inverse problem of wave propagation. For each time step, our iterative form costs $O(N^2)$ and the termination criterion judgement costs $O(N)$. In general, the total complexity is $O(N^2T + NT)$, where T is also the simulation time.

In our model, we mainly use wave propagation process and its inverse process. For the wave propagation process, the iterative form (scheme (6)) is always stable. For its inverse, we use scheme (9) to calculate. In scheme (9), function F defined on nodes of network is approximated by linear interpolation functions. Consulting the definition about accuracy of finite volume method, we can also say that scheme (9) has first order temporal accuracy. Since the length of edges equals 1, so the stable condition Courant-Friedrichs-Lewy condition (Subramani and Rajagopalan 2003) is $c\Delta t < 1$. By setting our simulation time step $\Delta t < 1$, our iterative form can get a stable solution.

6 Empirical analysis

Lots of data sets can be used in our model, and here we use an e-mails dataset of a company. In the empirical analysis, we discuss several paraments first, and then set them in our algorithm to solve the message broadcasting problem and rumor transmitting problem.

6.1 Data set

Here we use the network of Enron (Leskovec et al. 2009; Klimmt and Yang 2004). The network has 36692 nodes and 183831 undirected edges. Nodes of the network are email addresses and if an address i sent at least one email to address j, the graph contains an undirected edge from i to j. Note that non-Enron email addresses act as sinks and sources in the network as we only observe their communication with the Enron email addresses.

6.2 Results of message broadcasting problem

Given Enron network, we have to decide several parameters in our algorithm first. In our simulation, we let the wave velocity $c = 1$ and the adoption threshold $\gamma = 0.01$ while the initial amplitude of a source node is $f_s(0) = 1$. As we have discussed above, if at some time step t s node's amplitude is greater than γ, we think it has been successfully influenced. Besides, we use $\Delta t = 0.01$ as the simulation time step and we use 3 different simulation time $T = 0.1, 0.2, 0.5$. The empirical results are shown in Table 1 and Fig. 4.

Where coverage indicates the number of influenced nodes at time T.

In Table 1 we notice at early time, the message transmits quite slowly, but the transmiting speed goes up explosively with time. Take 7 nodes for example, when $T = 0.1$, it can just cover 6680 nodes (%18.21), while $T = 0.5$, it goes up to 25298 nodes (%68.94). Fig. 4 shows the coverage size at different time with different number of source nodes.

6.3 Results of rumor transmitting problem

In our algorithm, current state of all nodes must be given first. We give it by assuming some nodes to be the source (here we choose node 28521 and 28522), and then

Table 1 Nodes ID and coverage in different time period

$T = 0.10$	step	1	2	3	4	5	6	7
	ID	5038	273	458	1028	195	1139	370
	coverage	1384	2751	3901	4671	5388	6046	6680
$T = 0.20$	step	1	2	3	4	5	6	7
	ID	5038	273	458	140	1028	195	1139
	coverage	1384	2755	3915	5071	5805	6487	7099
$T = 0.50$	step	1	2	3	4	5	6	7
	ID	195	136	370	140	76	458	273
	coverage	17189	19885	21754	23454	24726	25013	25298

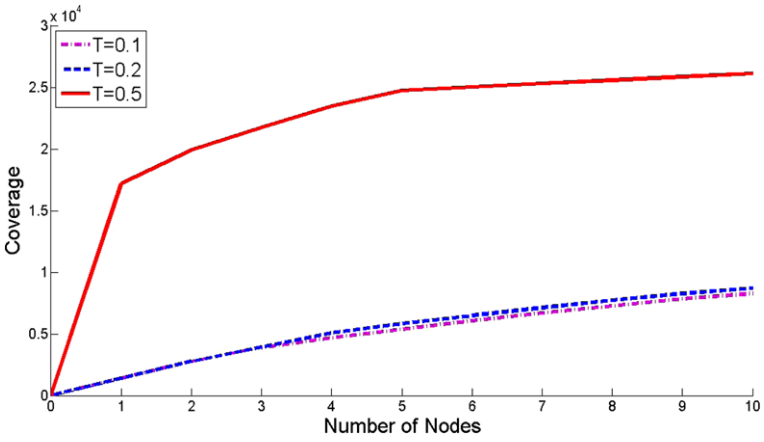


Fig. 4 Coverage with nodes number in different time period

simulate the wave propagation process for some time period and let the current state of amplitude to be the input. As for the source-like threshold σ, we let $\sigma = 0.001$. If at some time step, one node's amplitude is greater than σ, then it may be a source. Other parameters like c and Δt are the same as in Sect. 6.2. When the number of source-like nodes is least, we terminate our algorithm.

Our empirical results are shown in Fig. 5. For clarifying, we omit the edges in all graphs.

Where red nodes indicate the source-like nodes while grey not. The upper left indicates the current state (29901 red nodes), the upper right (18741 red nodes) and lower left (6017 red nodes) indicate two middle states, and the lower right (2 red nodes) indicates the terminated state. We also plot the tiny structure (red box in terminated state) in Fig. 6. Using our algorithm, we successfully find the rumor makers(node 28521 and 28522).

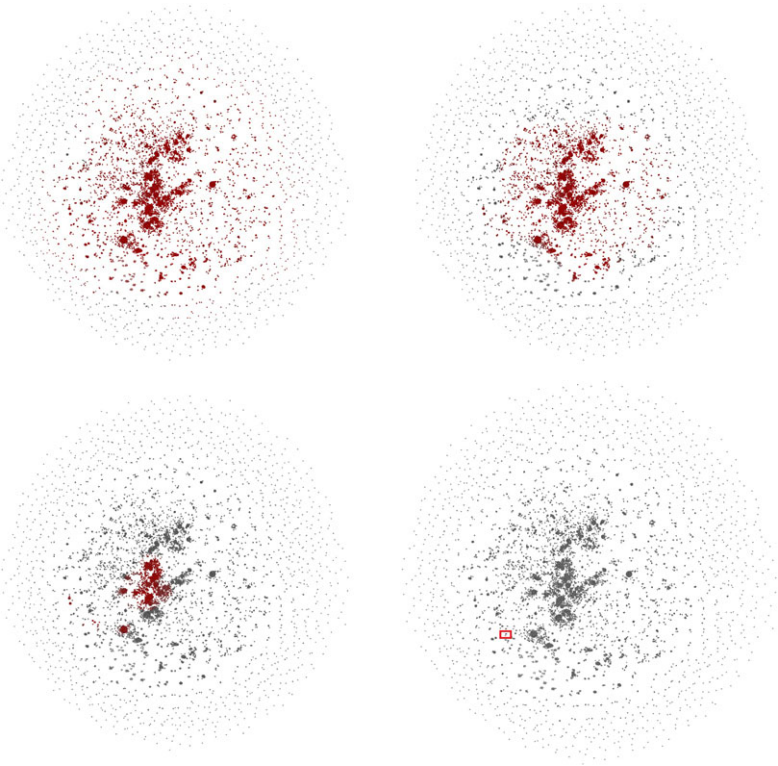
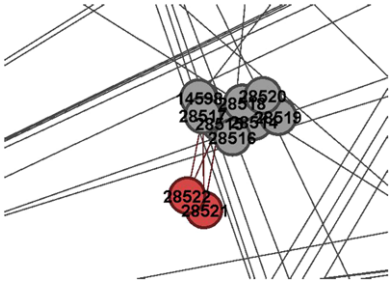


Fig. 5 Empirical results of Rumor Transmitting problem

Fig. 6 Tiny structure of terminated state



7 Conclusion

In this paper, we try to build a framework of mining social networks using Wave Propagation models. We mainly discuss two Wave Propagation models and two applications of our framework, including attempts to solving message broadcasting problem and rumor spreading problem. The complexity analysis shows our framework is scalable for large networks, and we can guarantee the stability of our algorithm by suitable setting some parameters. And the empirical analysis of Enron e-mail dataset shows that our work is promising.

Acknowledgements The work is partially supported by Natural Science Foundation of China (No. 11001237) and NUDT Preparing Research Project JC-02-01-04.

References

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Proc of the seventh international conference on the world wide web, pp 107–117

Pinheiro CAR (2011) Social network analysis in telecommunications. Wiley, New York, p 4. ISBN 978-1-118-01094-5

Hartline JD, Mirrokni VS, Sundararajan M (2008) Optimal marketing strategies over social networks. In: Proc of the ACM WWW conf, pp 189–198

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proc of the ACM SIGKDD conf, pp 137–146

Domingos P, Richardson M (2001) Mining the network value of customers. In: Proc of the ACM SIGKDD conf, pp 57–66

Leskovec MJ, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. ACM Trans Web 1(1)

Brown J, Reinegen P (1987) Social ties and word-of-mouth referral behavior. J Consum Res 14(3):350–362

Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12(3):211–223

Mahajan V, Muller E, Bass F (1999) New product diffusion models in marketing: a review and directions for research. J Mark 54(1):1–26

Subramani MR, Rajagopalan B (2003) Knowledge-sharing and influence in online social networks via viral marketing. Commun ACM 46(12):300–307

Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proc of the ACM SIGKDD conf, pp 61–70

Rogers EM (2003) Diffusion of innovations, 5th edn. Free Press, New York

Ma H, Yang H, Lyu MR, King I (2008) Mining social networks using heat diffusion processes for marketing candidates selection. In: CIKM'08, October 26–30

Fountoulakis N, Panagiotou K (2010) Rumor spreading on random regular graphs and expanders. In: 14th inter workshop on randomization and comput (RANDOM). LNCS, vol 6302, pp 560–573

Leskovec J, Lang K, Dasgupta A, Mahoney M (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. Internet Math 6(1):29–123

Klimmt B, Yang Y (2004) Introducing the Enron corpus. In: CEAS conference

Xiaojie Wang is currently postgraduate of statistics in National University of Defense Technology.

Hong Tao is currently postgraduate of statistics in National University of Defense Technology.

Zheng Xie is a Ph.D. in applied mathematics. He is currently an assistant professor of statistics in National University of Defense Technology.

Dongyun Yi is a Ph.D. in statistics. He is currently a professor of statistics, and the dean of college of science at National University of Defense Technology.