Research on Information Dissemination Model of Social Network Services Based on Probabilistic Hyper-graph

Min Du

Institute of Information and Engineering, Hunan University of Science and Engineering, 425100, Peoples Republic of China
duminemail@126.com

Abstract

The analysis of dissemination information and sources of information in social network services creates an urgent need with the explosive growth of the number of users and serious situation for false information and privacy leakage detection. Observe that the characteristics of randomness and uncontrollability makes malicious users manufacture and spread false malicious information, it is very difficult to find sources of information. A new information dissemination model based on probabilistic hyper-graph (IDMPH) is proposed. The method focuses on randomness for information dissemination process, introduces probabilistic hyper-graph to calculate the value of probability, find the most likely path and the leadership qualities for source nodes. Experimental results show that the conclusions are consistent with actual complex networks, have better ability to identify the path and leader nodes, can be used as basis for analysis of information control and information sources.

Keywords: social network services, probabilistic hyper-graph, information dissemination, leader nodes

1. Introduction

There forms a variety of social network services (referred to SNS) based on a large number of social relations' data in the network, in which information comes from different friends, family and colleagues [1]. It relies on relationship between friends, expands pervasive impacts of information technology, changes the main mode of information transmission compared with the mode of traditional information content in SNS. Therefore multiple social networks accelerate the dissemination of information, facilitate the information search and contribute to the promotion of interest. At the same time, with the increase spread of emerging internet applications and the addition of subscribers, public opinions in SNS show significantly difference from traditional features and reflect the characteristics of Six Degrees of Separation in human society plenty that easily cause malicious users spread false information between acquaintances [2], influence even damage individuals' economic interests and reputation interests of the people in real life. It has the great academic value and social economic significance to study the propagation characteristics of information in SNS, further analyze the evolution between groups, explore the influence of information dissemination model, and establish effective interaction model.

In recent years, domestic and foreign scholars had studied information dissemination process and information control in SNS [2-5]. In the works, David [3]proposed cascade model for information dissemination, he analyzed and found that information spread from one user to his neighbor based on a certain probability, which is just like waterfall falling. Zhang [4] approached a new information dissemination mechanism of online social network based on complex systems theory, social physics, infectious disease dynamics and sociology. Wang [2] proposed a fake information diffusion control method to reduce

the spread of fake information by controlling the diffusion behavior of users of high rankings, therefore eventually curb or even eliminate the diffusion of fake information. All the works rely on characteristics, principles and fake control of information dissemination, but do not calculate qualitatively the value of randomness, do not consider how to implement controllability for information dissemination.

We present a new information dissemination model based on probabilistic hyper-graph (IDMPH). The method focuses on randomness for information dissemination process, introduces probabilistic hyper-graph to calculate the value of probability, find the most likely path and the leadership qualities for source nodes.

2. Model-based Definition

Definition 1 (hyper-graph). Hyper-graph is a graph in generalization, it is a non-linear story structure, and marked by $G = (V, E)$. V is a vertex set, finite and non-empty, marked by $V = \{v_1, v_2, ..., v_n\}$. E is an edge set, may be empty, remarked by $E = \{e_1, e_2, ..., e_n\}$. It is includes vertices at e_m, $e_m \subseteq V$. When a node contains a sequential edge, it is marked that $e_m = <v_i, v_j>$, $v_i, v_j \subseteq V$, $v_i and v_j$ are directed edges. An edge in hyper-graph can connect any number of vertices, which can be used to describe nodes' multiple relationships.

Definition 2 (Weighted hyper-graph). e is a hyper-edge that is given a positive weight in weighted hyper-graph, marked by $w(e)$. The value of $w(e)$ shows the degree of subjection belong to the hyper-edge.

Definition 3 (Directed probability hyper- graph). Directed probability hyper-graph adds a conditional probability parameter based on hyper-graph marked by $p(v_j | v_i)$, which indicates the possibility from node v_i to node v_j in information dissemination. $w(e_m)$ is the information of probability weights, it is calculated below, as in Equation(1) and Equation(2):

$$w(e_m) = p(v_j | v_i) \tag{1}$$

$$sum(v_i) = \sum_{e_m \in E} w(e_m) = \sum_{v_j \in V} p(v_j | v_i) \tag{2}$$

$sum(v_i)$ is the sum of the possibility as v_i is a starting point and v_j is a receiving point.

Definition 4 (Directed probability adjacency matrix). Directed probability adjacency matrix records the scores with all nodes of the probability of passing. It is given below, as in Equation (3):

$$A[i][j] = \begin{cases} 1 - p(v_j | v_i) & i \neq j, \quad <i, j> \in E(G) \\ 0 & i = j \\ \infty & else \end{cases} \tag{3}$$

The greater value of $p(v_j | v_i)$ means the greater possibility for transmission in unknown information, and the value of $1 - p(v_j | v_i)$ can be more smaller, which means the shorter path from node v_i to node v_j.

3. Method and Implementation

3.1. Calculate the Most Likely Path

In IDMPH, we propose an algorithm for the minimum probability path from sourcing point to receiving point as the most likely path. $A[i][j]$ records the evaluation of vertex v_i to vertex v_j directly in the directed probability adjacency matrix at time *t*. Recursive

methods can be used to get matrix sequences (marked $A_0, A_1, ..., A_n$). $A_k[i][j]$ is the shortest path from vertex v_i to vertex v_j in less than k steps, as in Equation(4):

$$A_{-1}[i][j] = cos t[i][j]$$
$$A_{k+1}[i][j] = \min\{A_k[i][j], A_k[i][k+1] + A_k[k+1][j]\}$$
$$(-1 \le k \le n-2, 0 \le k \le 6)$$

(4)

3.2. Normalization

The processing method processes the hyper-edge (marked e_m) which receives information from vertex v_j as a receiving point, the result is below, as in Equation (5):

$$p(v_j \mid e_m) = \frac{A[i][j]}{\sum\limits_{j=1}^{n} A[i][j]}$$

(5)

$p(v_j \mid e_m)$ indicates the overall probability of the super-edge e_m.

3.3. Calculate the Average Value and Variance for Overall Probability

The average value and variance for overall probability is dealt with as follows, as in Equation (6) and Equation (7):

$$\overline{p}_r = \sum_{v_i, v_j \in V, <v_i, v_j> \in e_m} w(e_m) \bullet p(v_j \mid e_m)/k$$

(6)

$$s_r^2 = \frac{1}{k-1} \bullet \sum_{v_i, v_j \in V, <v_i, v_j> \in e_m} [w(e_m) \bullet p(v_j \mid e_m) - \overline{p}_r]^2$$

(7)

$\overline{p}_r$ is the average value of the receiving point v_j which has super edges with the number of k, s_r^2 describes the variance.

3.4. Analyze Leader Nodes

Each source node has an independent requirement and has consistent characteristics about information dissemination in SNS, such as consistent interest and consistent theme. Therefore, groups of nodes are formed easily based on characteristics and it products a fundamental impact on information dissemination in their group. Although each node has effect by multi-factors, an independent node plays a small role throughout the whole communication process and the impact is determined by the probability of super-node edge. It can be considered that the combined effect of all the nodes is identical to the spread of information dissemination in the model. When the influence of a node is far greater than an average impact, we can judge it is a leader node, leader nodes have greater influence than general nodes. According to current research, leader nodes behave overall unity, behavioral consistency and typical abnormality because of their behavior. Leader nodes have a large value of out-degree, a firm attitude and a strong stability that makes us to distinguish them.

We set a parameter marked d to analyze the overall impact of the standard deviation in IDMPH, as in Equation (8):

$$d = \frac{\sum\limits_{j=1}^{n} p(v_j \mid e_m) - \overline{p}_r}{s_r / \sqrt{n}}$$

(8)

We set a threshold marked α and let $\alpha = 0.9$. When the value of d is greater than α, it means that the impact of the individual node is much larger than the average impact and it can be determined as a leader node.

4. Experiments and Analysis

A message can be received as a certain probability to share his friends in SNS in the course of our study. Therefore, the nodes in experiments are divided into three categories: communication nodes (I-state), susceptible nodes (S-state) and immune nodes (R-state). Communication nodes can accept the information from his friends and have an ability to disseminate. Susceptible nodes can change to communication nodes as they do not know the message and receive information in a certain probability. Immune nodes can represent that message, but no longer have the capacity of spread or interest. The nodes can be transferred from one state to another which not only depends on the state of itself, but also depends on the state of its neighbor nodes [6-10].

4.1. Set Parameters for Source Nodes in IDMPH

Table1. Parameters for Source Nodes in Initial State I

name	out-degree	Point of strength	attitude
Node 1	40	0.95	support
Node 2	40	0.5	support
Node 3	15	0.95	opposition
Node 4	15	0.5	opposition
Node 5	15	0.2	support

Table 2. Parameters for Source Nodes in Initial State II

name	out-degree	Point of strength	attitude
Node 1	15	0.95	support
Node 2	15	0.5	support
Node 3	15	0.95	opposition
Node 4	15	0.5	opposition
Node 5	15	0.2	support

In IDMPH, we provide five source nodes to analyze the quantization process and set a threshold (marked α, $\alpha = 0.1$). We assume that all source nodes are in I-state, the remaining nodes are in S-state, parameters for comparison is shown in table 1 or table 2.

4.2. Test and Analysis of Information Dissemination Process

We set 3500 nodes by 300 time steps (to reach stabilized) based on a micro-blog to simulate the spread from a life cycle perspective. It select 5 nodes as source nodes, three nodes support for views, two nodes oppose, the other nodes remain neutral and are in S-state, that is listed in Table 1.

The percentage of nodes reflects the process of information dissemination with time diagram in Figure 1. The percentage of susceptible nodes (S-state) decreases rapidly to 0 in SNS, which indicates that a message spreads fast and the user get the message in a short time. The percentage of communication nodes (I-state) decreases after increases in 5 time steps，which shows a large number of nodes gets the message in the early of information dissemination, and the percentage of communication nodes increases until the maximum value. After the users lose interesting, the percentage of communication nodes attenuates to 0. The percentage of immune nodes (R-state) presents from fast to slow

growth, and finally reaches 1. The experiments test and verify the characteristics of Six Degrees of Separation, communication nodes get message in less than or equal to 5 nodes after 5 time steps.
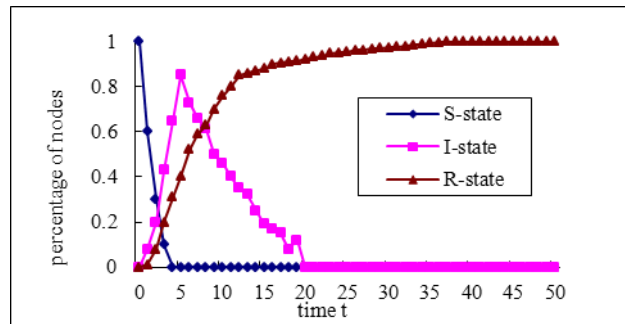

Figure 1. Comparison for Percentage of Nodes with Time in IDMPH

4.3. Comparative of Process and Characteristics for Information Dissemination

In order to analyze process and characteristics for information dissemination, parameters of source nodes are set according to tables 1 and tables 2 at different experiments. The change percentage of nodes about different attitudes with time is shown in Figure 2.
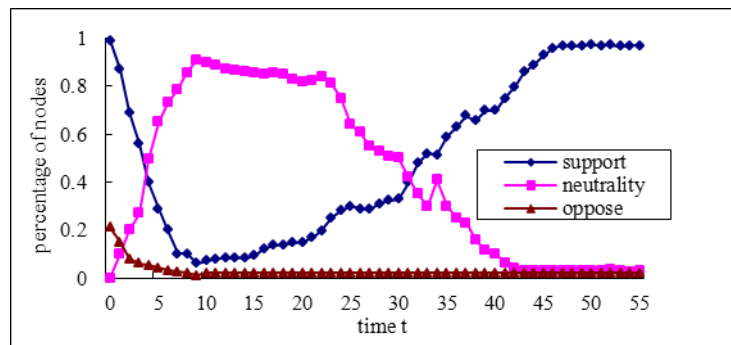
(a)Parameters of Source Nodes Set in Table 1
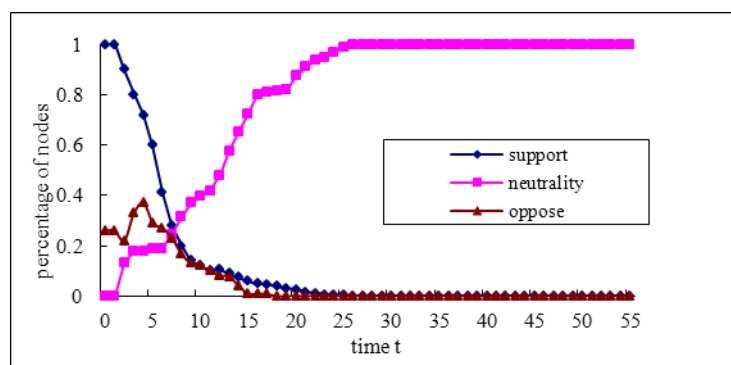
(b) Parameters of Source Nodes Set in Table 2

Figure 2. Comparison of Change Percentage for Nodes about Different Attitudes with Time in IDMPH

That shows comparison of change percentage for nodes about different attitudes with time in Figure 2(a) which set parameters according to Tables 1. At early spread of information dissemination, on the one hand neutral nodes and opposition nodes get message that causes the neutral curve and the opposition curve decline, but on the other

the support curve rises. With the pros and cons of dual heated discussion, many nodes which have not a firm attitude change into neutral. When the number of neutral nodes reaches the maximum,and the other nodes fall to the lowest point, strong positive nodes belongs to supporters and strong negative nodes belongs to opponents stably at this time. In this case, neutral nodes turn into a support and the value of the neutral curve weakens to 0.At the same time, the value of the support curve incrementally adds. Finally individual opposition nodes are isolated because of different opinions with their neighbors.

That shows comparison of change percentage for nodes about different attitudes with time in Figure 2(b) which set parameters according to tables 2. The percentage of support nodes gradually declines to 0 when the value of out-degree for source nodes lessens. And the percentage of opposition nodes eventually declines to 0 even though it improves in the beginning. But the percentage of neutral nodes increases to 1 rapidly because the nodes with view of small intensity (which have small value for out-degree and point of strength) cannot influence other nodes and are difficult to propagate in SNS, that makes nodes with great view of intensity difference belongs to intermediate state steadily and the network eventually becomes steady at the intermediate state.

4.4. Analyze Leader Nodes

Nodes changes characteristic related to their father nodes, so the view of source nodes affects their adjacent nodes easily, even though extends to the whole network. Therefore leader nodes have a greater social influence and make their neighbore nodes accept their views in a short time by high connectivity in SNS. That shows nodes with a small out-degree and influence are difficult to get recognition for the other nodes from Figure 2 (b), therefore leader nodes must have a large out-degree, have a tremendous influence and have hardline views that makes adjacent nodes behave for blind obedience greatly. That shows comparison of probability for source nodes at different times in Figure 3 and parameters of source nodes are set according to tables 1. From the experimental results, transmission probability of node 1 is larger than the value of the other source nodes and threshold limit, we can determine node 1 is a leader node. It has a large out-degree and a firm attitude. This shows analysis in the model is consistent with realistic network.
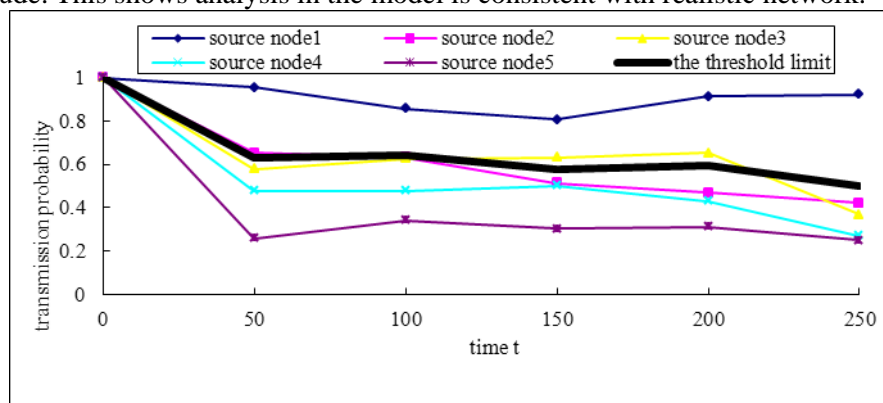

Figure 3. Comparison of Transmission Probability of Source Nodes Changeing with Times in IDMPH

5. Conclusion

We have proposed an information dissemination model of social network services based on probabilistic hyper-graph（IDMPH）that has good features for analyzing the most likely path and leader nodes. The proposed scheme can quantify the analysis for the most likely path,also has good discriminative capabilities and can identify leader nodes.

Besides the results are consistent in actual complex networks, it can provide a basis for analysis of information control and information sources.

But then, our technique parameters approaches to informatioan sources and information control. Besides out-degree of source nodes, viewpoints of attitudes, strength and the value of spread probability for nodes, there are affected by other factors, such as effect of span-depth ratio and supplying efficient algorithm. More effort is needed to address these problems.

Acknowledgment

References

[1] [1] Danah m Boyd and Nicole B. Ellison. Social network sites:Definition, history and scholarship. Journal of Computer Mediated communication,vol 13,no 1,pp.201-230**(2008)**

[2] [2] Wang Y.G., F.Z. Cai and E.K. Lua. A diffusion control method of fake information in social networks. Journal of computer research and Development, vol 49(suppl), pp. 131-137**(2012)**

[3] [3] David K., K. Jon, and T. Eva. Maximizing the spread of influence through a social network. Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, **(2003)** October 15-19; New York, USA

[4] [4] Zhang Yan-chao,Liu Yun and Zhang Hai-feng. The research of information dissemination model on online social network. Journal of Acta Physics Sinica, vol 60, no 5,pp. 050501-050507**(2011)**

[5] [5] Mo T., W.J. Chu, and W.P. Li. A micro blog user community aware method based on hypergraph. Journal of xi'an Jiaotong University xuebao,vol 11,no 46, pp.120-126**(2011)**

[6] [6] Lancieri L and N. Durand. Internet user behavior:compared study of the access traces and application to the discovery of communities. Journal of IEEE Trans. Systems, Man and Cybernetics:Part A Systems and Humans, vol 36,no 1, pp. 208-219**(2006)**

[7] [7] Ou G.Y., W. Chen, and T.J. Wang. Sentiment influence maximization model for microblogging system . Journal of Frontiers of Computer Science and Technology,vol 6,no 9, pp. 769-778**(2012)**

[8] [8] Wen Z.H., W. Huang, and R.R. Liu. Research on state reachability in planning based on model checking. Journal of computers, vol 8, no 35,pp.1634-1643**(2012)**

[9] [9] Huang Y., Z.Y. Qin, Y.M. Zhuo. Social relationship discovery based on spatio-temporal data. Application research of computers, vol 8,no 28,pp.3004 -3006**(2011)**

[10] [10] Ahn G, M. Shehab, and A. Syuicciarini. Security and privacy in social networks. IEEE Internet computing, vol 15,no 13,pp.10-12**(2011)**

Author

Min Du, she received her M.E. in Computer application (2008) from Central South University. Now she is associate professor of computer at information and engineering Department, Hunan University of Science and Engineering. Her current research interests include different aspects of information security and network security.