An Effective Measurement Scheme of Node Influence in Aviation Network

Xiao Luo, Jianbin Li, Xun Zhang, Tiedan Xu,
Yan Lu, Qian Luo
The Second Research Institute of CAAC
Chengdu, China

Shina Xie, Chuan Li
College of Computer Science
Sichuan University
Chengdu, China

Abstract—**Aviation network is the typical information network, where the airports can be regarded as nodes, and the airlines between airports can be regarded as links. But with the constraints with the Degree-Centrality, Closeness-Centrality and the Centrality among nodes, traditional researches on the influence of the information network developed slowly. Thus, we do some experiments on the traditional methods that used in the researching on the influence of the information network. And during experiments, we find that traditional methods have two drawbacks: (1) Universality is an unpredictable factor but plays an important role in research because no indicator can be suitable for any occasion. For instance, when the size of information network extends very large, it may take up a lot of time to compute the Closeness-Centrality and the Centrality among nodes. (2) Node is too difficult to select correctly. For example, it is an NP-hard problem to select K as the most influential node because it has a very high computational complexity. This paper proposed a new perspective to depict the node influence in information network: By doing research on susceptible sides of the node depicts the node influence in aviation network. Simultaneously, this scheme includes the measurement method of non-directional and directional network. And we use the Independent-Cascade-Model to evaluate the feasibility and effectiveness of our method. Because of the lack of real aviation network data, we apply our method on three similar real network data-sets (Hamsterster full, Ca-HepTh etc.). And the result refers to the experiments show that our method exceeds the greedy algorithm of influence maximization on the spread effect. Moreover, this also shows that our method not only can be suited to the aviation networks, but also can be used in other information networks. Therefore, our scheme can depict the node influence in information network accurately.**

Keywords-Aviation network; influence; influential node; measurement method

I. INTRODUCTION

Information network usually refers to the complex network which is consists of entities and the ties of those entities, that is composed of nodes and edges. Node can be said person or airport. Edges represent the relationships between entities. Such as airline, cooperation, friend relationship, recognition, etc. For example, nodes represent the airports and edges represent the airlines between airports in aviation networks.

When studying the diffusion of information, we found that the structure of information network plays an important role. For instance, when a man accepted some new, he is very likely to recommend those to his neighbors and friends. And some of them will be affected by the acceptance of the new things, and further recommendations to their neighbors, friends. In such a diffusion process, the social relationships of a man and behaviors of friends would affect his decision [1].

The research of node's influence on information network has a long history [2, 4, 7]. In recent years, with the prevalence of information networks, such as Facebook, twitter etc. makes this become a hot spot once again. While the huge number of users and the complex relationships between users bring the related research enormous challenges.

Example1 (Viral Marketing). Viral Marketing (or word-of-mouth marketing), the goal of this is to select a small set of most influential users in a large-scale information network, and to obtain the greatest possible benefits through this relationship to achieve minimal marketing costs.

In addition, the analysis of node influence can also promote the depth understanding of the information network, found opinion leaders and improved the personalized recommendation, etc. [5-8]. Intuitively, node has higher degree, which is the basis of many others related assuming studies and can make a greater influence. Other measurements of node influence are the closeness centrality, the centrality among nodes, etc. [9]

The shortcoming of this work is that these indicators are not suitable for any occasion. When the network size is very large scale, it may take up a lot of time to computing the Closeness Centrality and the Centrality among nodes. Although calculating degree centrality is simple, some researches show that selecting the top k node with biggest degree cannot get the widest scope of influence. And how to select K as the most influential node to maximize the expected number of influenced nodes also is an NP-hard problem. But the greedy method to solve this problem can only get 63% of the solutions of all optimal solutions for this NP-hard problem [1]. And it has high computational complexity. So it is not suitable for large-scale networks.

The analysis of node influence in information network usually need information propagation model. And the most Frequently used model is Independent-Cascade-Model. This detail content will show in section 1.

There is some researches show that [10]: a large-scale change in public thought (the spread of new things) is promoted by susceptible users. Based on this idea and the shortage of existing methods, and this cannot generally accurately depicted the influence of node in network. This paper proposed a new method, named SIM (Susceptible-Influence-Measurement), which is from the node susceptible side depicts the node influence in aviation network. Our method not only can be suited to the aviation networks, but also can be fit for other information networks. Moreover, experiments show that our method exceeds the greedy algorithm of influence maximization on the spread influence. And our scheme can accurately depict the node influence on information network including the aviation network.

Four sections structure the rest of this paper. The first section describes related work and Independent-Cascade-Model. The second section describes the new method proposed by author, SIM, and its extension on directional network. The third section describes experiment on different data sets and the analysis of the results. The fourth section summarizes the full text.

II. RELATED WORK

A. *Information propagation model*

A Frequently used model is Independent-Cascade-Model (IC). It is based on three common assumptions:

(1) The state of a node is either active or inactive

(2) A node is active, it will active its neighbor nodes;

(3) Once a node is activated, the state of the node cannot be changed.

In the IC model, if node u is activated in step t, it is only one chance to activate its neighbor nodes. The neighbor node v is activated by u with probability p. In other words, if there are l neighbors of v that are active, v will be activated with $1-(1-p)^l$. If v is activated successfully, it will be the activated node in step t+1. If v has more than a neighbor, they will active v in random order. Under the IC model, the activation processes begin with a set of initial active nodes, and repeated this process until there is no node can be activated. Detail activation process shown in Figure 1.

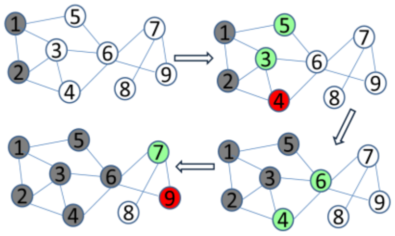


Figure 1. IC model activation process

In the figure 1, the gray represents active state; green represents the new activated state; red represents activation failed node; blank means inactive. The initial nodes are 1, 2. At first step, 3, 5 are activated successfully, and node 4 is activated failed. At second step, 3, 5 have the ability to activate other neighbor node with probability p. 4, 6 are activated successfully. At third step, 4, 6 are the given active node, and node 7 is activated successfully, node 9 is activated failed.

B. *Problem in the traditional method*

With the prevalence of complex networks, such as Facebook, twitter, etc. makes academic communities pay close attention to the research of node influence in information network [6-8]. Traditional measurements of node influence are the degree centrality, the closeness centrality and the centrality between nodes etc. But these indicators are not suitable for any occasion [9]. Calculating degree centrality is easy, but cannot accurately depict the influence of node. When the network size is very large scale, it is required a lot of time to computing the Closeness Centrality and the Centrality between nodes. Chen et.al [11] proposed a local centrality measure to describe Spread influence of node. This indicator using SIR model and considering single infection sources, cannot be suitable for any occasion.

Recent years, there are some improved algorithms that are about the greedy algorithm proposed by Kempe [4] et.al. These algorithms improve the computational efficiency of greedy algorithm, however, because of its high computational complexity are incapable to be applied in large-scale information network. It is worth noting that Leskovec [12] et.al propose CELF algorithm. This algorithm removed a lot of double counting, and it improved the computing speed of greedy algorithm, but did not expand the scope of influence. In the experimental section, we compare our algorithm with the CELF. But these existing algorithms are difficult to achieve the ideal effect, even the greedy algorithm is to find a set of influential nodes, and this does not give the node influence measurement.

III. INFLUENCE MEASUREMENT BASED ON SUSCEPTIBLE NODES

When studying the diffusion of node influence, we found the Two following details:

(1) Information spread widely benefit from the users who are susceptible.

(2) The more similar the user more easily influences each other.

Based on the above findings, this paper proposed the new thought of mining the most influential nodes in the information network. This scheme considers this kind of node had large influence, which connected with the more susceptible nodes. We use the similarity between nodes to help to describe the degree of node easily affected. Simultaneously, this scheme includes the measurement method of non-directional and directional network. This detail will show in section 2.1and section 2.2.

Information network can be represented as an non-directional (directional) graph. Node represents user. Edge between nodes represents the relationship between users. $G(V,E)$ is an non-directional (directional) graph, in which, V is the set of nodes, E is the set of edges between nodes.

A. *Influence measurement scheme in non-directional network*

Benefiting from the classic conclusion, which is "the more similar user more easily influences each other". This paper put forward utilizing the similarity of nodes to measure the degree of node easily influenced.

Definition 1.(the degree of node easily influenced) $S_{v,u}$ represents the similarity between v and u ; N_v represents the set of neighbors of node v ; E_v is the degree of node easily influenced.

$$E_v = \sum_{u\in N_v} S_{v,u} \qquad (1)$$

Definition 2.(the influence degree of node v influenced by u) $e_{v,u}$ is the influence degree of node v influenced by u .

$$e_{v,u} = S_{v,u} / E_v \qquad (2)$$

We consider this kind of user had bigger influence, who connected with the more susceptible users. The influence of a user is definition as fallow.

Definition 3.(the influence of node v) we use I_v to represent the influence of node v . $e_{u,v}$ Means the influence degree of node u influenced by v .

$$I_v = \sum_{u\in N_v} e_{u,v} \qquad (3)$$

B. *Influence measurement scheme in directional network*

Considering the direction of information spread, we proposed the influence measurement scheme in directional network.

Definition 4.(the similarity between v and u in directional network) $S_{v,uOUT}$ represents the similarity of out-degree direction; $S_{v,uIN}$ represents the similarity of in-degree direction. $S_{v,u}$ Represents the similarity between v and u in directional network.

$$E_v = \sum_{u\in N_{vOUT}} S_{v,u} \quad S_{v,u} = (S_{v,uOUT} + S_{v,uIN}) / 2 \qquad (4)$$

Definition 5.(the degree of node easily influenced in directional network) N_{vOUT} represents the set of neighbors of node v in out-degree direction, E_v is the degree of node easily influenced in directional network.

Definition 6.(the influence of node v in directional network) N_{vIN} represents the set of neighbors of node v in in-degree direction. Then, I_v is the influence of node v in the directional network.

$$I_v = \sum_{u\in N_{vIN}} e_{u,v} \qquad (5)$$

C. *The framework of algorithm SIM*

In this paper, we implement the measurement scheme showed in section 2.1 and 2.2. The pseudo code is illustrated in Algorithm1. Algorithm 1 is used for computing the influence of node in information network. Line 1 to 7 is used for computing the degree of node easily influenced. If the network is non-directional use the definition 1to computing. If the network is directional use the definition 5 to computing. Line 8 to 12 is used for computing the influence degree of node v influenced by u . Line 13 to 19 is used for computing the influence of node v . If the network is non-directional use the definition 3to computing. If the network is directional use the definition 6 to computing.

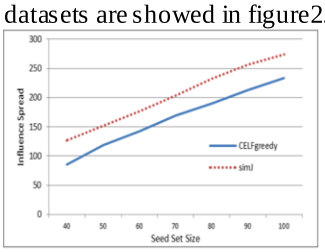
Algorithm1. framework of algorithm SIM
Input: graph $G(V, E)$, similarity $S_{v,u}$
Output: influence of node I

1. **for** $(v \in V)$ **do**
2. **if (** $G(V, E)$ is undirected**) then**
3. $E_v \leftarrow \sum_{u\in N_v} S_{v,u}$
4. **else**
5. $E_v \leftarrow \sum_{u\in N_{vOUT}} S_{v,u}$
6. **end if**
7. **end for**
8. **for** $(v \in V)$ **do**
9. **for** $(u \in N_v)$ **do**
10. $e_{v,u} \leftarrow S_{v,u} / E_v$
11. **end for**
12. **end for**

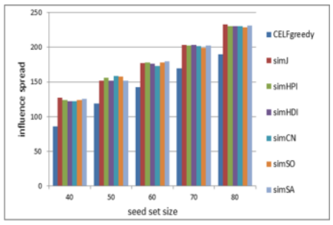
IV. EXPERIMENTAL ANALYSIS

Because of the lack of real aviation network data, we apply our method on three similar real network data-sets. In this paper, we use three real datasets (Ca-HepTh, Hamsterster full and Wiki-Vote), which represent different social semantics. Ca-HepTh represents cooperation relationship; Hamsterster full represents friend relationship; Wiki-Vote represents recognition relationship that is a directional network.

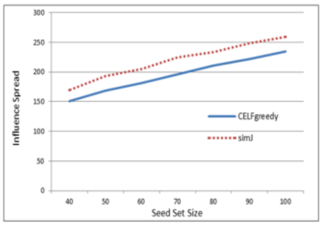
We use the IC Model to evaluate the feasibility and effectiveness of our method and set $p = 0.01$. We compare our algorithm with the CELF. When computing the similarity, we used six kind of local similarity of node [13]. Which include fallowing indicators: Sorenson, Salton, CN, Jaccard, HPI and HDI. The results of experiment on three real networks datasets are showed in figure2.



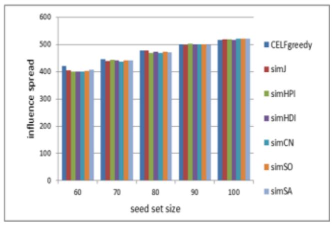
(a). result for data set Ca-HepTh



(b). result for data set Ca-HepTh



(c). result for data set Hamsterster



(d). result for data set Wiki-Vote

Figure 2. comparison of the CELF greedy algorithm and SIM on three real data sets

Figure 2 compares the spreading effects of CELF and SIM algorithm on three real datasets. The horizontal coordinate is seed set size, which is referred to the size of initial active nodes. Vertical coordinate is influence spread of initial seed set. That is means the scope of influence caused by the initial seed set. In the figure, simJ represents the result of Jaccard indicator; simSO represents the result of Sorenson indicator; simSA represents the result of Salton indicator, and so on. (a) and (b) are the result for data set Ca-HepTh. (c) is the result for data set Hamsterster full.(d) is the result for data set Wiki-Vote . From this figure, we can find that our algorithm is better than the CELF greedy algorithm on the scope of influence. At the same time, our algorithm can give a specific size of nodes influence, which is the greedy algorithm cannot come true. Our scheme can accurately depict the influence in information network. This scheme is based on the structure of network, and does not need more other information. So it is pervasive.

V. CONCLUSION

In this paper, aviation networks are viewed as Information-Networks. The Degree Centrality, the Closeness Centrality and the Centrality between nodes limit traditional research on information networks influence. In experiments, we find that traditional methods have some shortcoming: first, these indicators are not suitable for any occasion. When the network size is very large scale, it may cost a lot of time to computing the Closeness Centrality and the Centrality between nodes. Second, it is an NP-hard problem to selecting K as the most influential node. The greedy method to solve this problem can only get 63% of the solutions of all optimal solutions for this NP-hard problem. And it has high computational complexity. So it is not suitable for large-scale network.

This paper proposes the new thought of mining the most influential nodes in the aviation network. SIM, which considers this kind of node had bigger influence, which connected with the more susceptible nodes. Simultaneously, this scheme includes the measurement method of non-directional and directional network. Our method not only can be suited to the aviation networks, but also can be fit for other Information-Networks. Due to the lack of real aviation network data, we test our method on three similar real network datasets. The results of experiment show that our algorithm is better than the greedy algorithm, and can accurately depict the node influence in information network including the aviation network.

ACKNOWLEDGMENT

This research was financially supported by the Natural Science Foundation of China under Grant No. 61103043, U1233118, U1333122) the National “12th-Five-Year Plan”, Key Technology R&D Program of China under Grant No. 2012BAG04B02.

REFERENCES

[1] Tang L, Liu H. Community detection and mining in social media[J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2010, 2(1): 1-137.

[2] Serrat O. Social network analysis[J]. 2009.

[3] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 1029-1038.

[4] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003: 137-146.

[5] Song X, Chi Y, Hino K, et al. Identifying opinion leaders in the blogosphere[C]//Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007: 971-974.

[6] Centola D. The spread of behavior in an online social network experiment[J]. science, 2010, 329(5996): 1194-1197.

[7] Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network[C]//AAAI. 2007, 7: 1371-1376.

[8] Guille A. Information diffusion in online social networks[C]//Proceedings of the 2013 Sigmod/PODS Ph. D. symposium on PhD symposium. ACM, 2013: 31-36.

[9] WANG X, LI X, Guanrong C. Network Science: An Introduction [J]. 2012.

[10] Watts D J, Dodds P S. Influentials, networks, and public opinion formation[J]. Journal of consumer research, 2007, 34(4): 441-458.

[11] Chen D, Lü L, Shang M S, et al. Identifying influential nodes in complex networks[J]. Physica A: Statistical Mechanics and its Applications, 2012, 391(4): 1777-1787.

[12] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007: 420-429.

[13] Lin-yuan L. Link prediction on complex networks[J]. Journal of University of Electronic Science and Technology of China, 2010, 9: 5-39.