A new method for ranking the most influential node in complex networks

Zhisong Wang[1 2 3]
[1]College of Mechanical Engineering, YANSHAN University, Qinhuangdao , 066004,China
[2]Key Laboratory of Advanced Forging & Stamping Technology and Science,Yanshan University, 066004,China
[3]Hebei Provincial Key Laboratory of Parallel Robot and Mechatronic System,Yanshan University,066004,China
wzs@ysu.edu.cn

Abstract—Currently, understanding the topology structure and function of complex networks has become a hot research topic, and finding the most influential node in a complex network has great significance in marketing, public opinion analysis, disease control and so on. We often use the degree centrality and some other centralities to measure the impact of the node, but they only reflect partial nature of the network. In order to describe the key nodes of the network more accurately, in this paper, we present a new method of ranking the most influential node in a complex network, which not only takes the degree centrality into consideration but also takes the position of the nodes in the network and the nodes' important neighbors into account. Then we use the Independent Cascade Model for propagation simulation. Experiments show that the method we proposed could identify the most influential nodes more effectively. Compared with the traditional method it has better dissemination of results and lower time complexity.

Keywords- complex networks; node influence; node ranking; information dissemination

I. INTRODUCTION

SNS (Network Service Social) social network as a complex network depicts the individual members in the network and their relations. If the complex network is considered as a graph, the nodes in the graph represent the individual and the edges represent the relationship between the two individuals. In the real world, a large number of complex systems can be described by the network such as the connection between the router, the author and co-authors, protein function relations and so on[1]. Therefore, the research based on the complex networks has become a hot research topic at present. As the basis of the research of complex network, network node influence has great significance in theory and practical application. It has very good application values in the field of marketing, public opinion supervision, disease prevention and other fields. And it also receives more attention[2] and becomes one of the hot research.

Identifying the most influential nodes in network is conducive to betterly analyze the network. For example, in the field of controlling the disease transmission, finding the crucial node for the disease transmission as soon as possible can control the spread of disease. So how to identify the most influential node in the network? Literature [3] used the degree centricity to measure the influence of the node, which pointed out that in the network of the power rate that is distributed evenly, the larger the node degree is, the larger the node influence is. Literature [4] proposed that rank the node influence according to the betweenness centrality and the betweenness reflects the number of the shortest path through a certain point. Literature [5] proposed two propagation model, namely, independent cascade model and linear threshold model. And it used the greedy algorithm to solve the most influential nodes. The PageRank algorithm which is proposed by literature [6] is applied to the page rank and the author thought that the importance of the web page depends on the number of links pointing to it. Literature [7] proposed using the K - shell to decompose the network and the node with the largest K-shell value has good communication effect.

At present, many research results have considered that the node with the high degree or the larger betweenness in the network is the key node in the communication. Obviously, the node with high degree has more interpersonal relationship and the node with larger betweenness has more shortest paths. However, both of them haven't considered the overall structure of the network and the location of the node in the network. In Fig. 1[8], node 14 has the largest degree. However, its neighbor nodes are difficult to continue to spread the information, which makes the node may not be the best choice for the key nodes. At the same time, the K-shell method points out that the nodes in the central location of the network are more likely to spread information through the cascade effect than the edges nodes[9]. But a large number of nodes can have the same K-shell value and which is more critical in these nodes will not be able to differentiate [10].

According to the above problem, to depict the influence status of the node in the network more comprehensively and accurately, this paper puts forward a new node influence ranking algorithm DKIN (degree and k-shell and important neighbor), which combines with the node degree, node location influence, K-shell value and node neighbor influence. On this basis, the algorithm chooses the top K points as seed nodes, then using the Independent Cascade Model (Independent Cascade Model) to simulate the propagation process After obtaining the average propagation range, it uses the different propagation probabilities to complete the contrast experiment. Finally the proposed method is verified by experiments to find the most influential nodes having a better influence range.

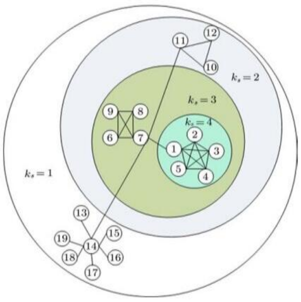

Figure 1. Network sample

II. RELATED DENIFITIONS AND THEORETICAL BASIS

The complex network graph is defined as a two tuple G= (V, E), where V represents the set of nodes in the network, E represents the set of edges between nodes and the nodes number is n=|V| while the edge number is m=|E|. The following we will introduce the related centricity indicator and K-shell decomposition as well as the general model of information dissemination.

A. *Degree centrality*

The degree centrality[11] is the most commonly and easily used method to depict the influence of nodes, which is specifically defined as follows.

Given a complex network G = (V, E) and any node $v_i \in V$. The degree of v_i is defined as the number of the nodes that were connected with the node directly, and it was denoted by $C_{d(vi)}$, as shown in (1), where $N(v_i)$ represents the set of the neighbors of v_i.

$$C_{d(vi)} = degree(v_i) = \sum_{j \in N(v_i)} a_{ij} \qquad (1)$$

For example, in the information dissemination, the large V users of the micro blog network have millions or even tens of millions of fans, and advertising businesses are likely to consider these people firstly. Therefore, the node degree can reflect the local nature of the nodes in the network, the computation complexity is low and the time is fast, which is suitable for large networks.

B. *Betweeness centrality*

Betweeness centrality was proposed by Freeman in 1977 [4]. The centricity degree thinks that every neighbor node of the node has the same important degree. Obviously the truth is not the case. The betweeness centrality can measure the status of each node and it usually can be defined as the shortest path number ratio through node v_i, which is denoted by $C_{b(vi)}$, as shown in (2).

$$C_{b(i)} = \sum_{s \in V, t \in V} \frac{n_{st}^i}{g_{st}} \qquad (2)$$

Where, g_{st} is the number of the shortest path between node s and t, n_{st}^i is the number of the shortest path between s and t through the node v_i. For example, the people who play a part in bridge and bond in the social network own more interpersonal relationships and their influence are higher. Betweeness centrality is a measurement that can reflect the global situation and its computational complexity is high, which is not suitable for large networks.

C. *K-shell decomposition*

Kitsak et al. [9] think that the influence of the node is determined by its location in the network, and the importance of nodes is given by K-shell algorithm. The basic idea of K-shell is a iterative method, which shells the nodes in the network. The higher shell number indicates that the node is more important and it has the greater influence. Steps are as follows: Firstly, calculate the degree of all the nodes in the network and remove all the nodes whose degree is 1 and the edges related to them. Scan again on this basis. If there is still nodes whose degree are less than 1, continue to remove these nodes and the edges related to them, and these node's K-shell value is 1. If there is no one can be removed, then K value plus 1. Repeat the above process until all the nodes have been assigned a K-shell value. This completes the K-shell decomposition of the network and each node has a corresponding K-shell value.

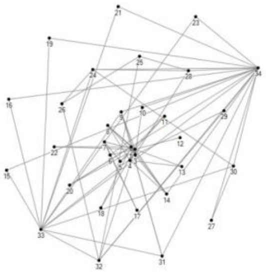

Figure 2. Zachary karate club graph

To take the Zachary Karate Club dataset [13] as an example, as shown in Fig. 2, the network contains 34 nodes and 78 edges. The K-shell decomposition to the Zachary karate club network is shown in table 1. The results show that the network can be divided into four shells. The higher the shell number is, the greater the node influence is. And the nodes whose K-shell value are 4 are the most influential nodes.

According to table 1, a number of nodes are in the core shell that is the fourth shell, but which one is the most influential node? Obviously, the K-shell decomposition can result that many nodes have the same KS value but it is unable to measure the influence of these nodes. If the network scale is larger, the number of the nodes having the same KS value is more. So it is obviously unscientific if we think the influence of these nodes are equal. K-shell just divides the network, but for the influence of the nodes, it also needs more accurate measurement.

TABLE I. K-SHELL DECOMPOSITION TO ZACHARY KARTE CLUB

	Nodes
K-shell=1	12
K-shell=2	10, 13, 15, 16, 17, 18, 19, 21, 22, 23, 27
K-shell=3	5, 6, 7, 11, 20, 24, 25, 26, 28, 29, 30, 32
K-shell=4	1,2, 3, 4, 33, 34, 9, 8, 14, 31

D. *Information dissemination model*

In the network G= (V, E), it is considered that each node has only two states, namely, active and inactive. The state of the node can only be transformed from inactive state to active state and it can't be transformed oppositely.

The progress that node is transformed from inactive to active is an activation event. The independent cascade model and the linear threshold model are the common communication models. This paper uses the independent cascade model to simulate the propagation process.

Independent cascade model[5] is a probability model based on independent particle system, which uses the probability to indicate the influence degree of a node to another node. If the probability is larger, the possibility of activation success is larger. Its communication process is as follows: when node v_i becomes active state in t time steps, it has one and only one chance to activate its inactive neighbor w, and the probability of success is PVW (a parameter independent to the system). If the node v_i has a number of neighbors, the order of v_i activates its inactive neighbor nodes is arbitrary. If the node v_i is successful, the w becomes active at t + 1 time. If the node v_i is fail, w won't be influenced in the back of the process. This process continues until there is no activation event occurs.

III. NODE INFLUENCE MODEL

The influence of nodes is not only related to the local properties (the degree of the node) of the network, but also the location of the nodes in the network. Therefore, a single indicator can't reflect the nature of the nodes in the network, and it does not apply to a variety of complex networks. In order to comprehensively and accurately describe the status of nodes in the network, considering the above three factors, we proposed a new node influence measurement standard DKIN. Specific definitions are as follows.

Given the complex network G= (V, E), any node $v_i \in V$, the influence of v_i is denoted by $DKIN(v_i)$, as shown in (3).

$$DKIN(v_i)=\alpha f_{degree}(v_i)+\beta f_{k\text{-}shell}(v_i)+\gamma f_{imptnb}(v_i),\quad \alpha+\beta+\gamma=1 \tag{3}$$

Where, $f_{degree}(v_i)$ represents the local influence of nodes, $f_{k\text{-}shell}(v_i)$ represents the location factor of the nodes, and $f_{imptnb}(v_i)$ is the important neighbor. A, β, γ represents the influence factor weight, which satisfies $\alpha+\beta+\gamma=1$.The three weight factors can be set according to the network structure, so that they can be applied to more networks.

$f_{degree}(v_i)$ is as shown in (4), where $Cd_{(vi)}$ represents the degree of node v_i and V is the set of all nodes in the network. $f_{degree}(v_i)$ represents the ratio that the local influence of node v_i in the entire network.

$$f_{\deg ree}(v_i)=\frac{cd_{v_i}}{\max\limits_{v_i \in V} cd_{(v_i)}} \tag{4}$$

$f_{imptnb}(v_i)$ is as shown in (5), where $KS_{(vi)}$ represents the k-shell value of the node v_i and V represents the set of all nodes in the network. $f_{k\text{-}shell}(v_i)$ represents the important degree of the location of the node v_i in the network.

$$f_{k-shell}(v_i)=\frac{Ks_{(v_i)}}{\max\limits_{v_i \in V} Ks_{(v_i)}} \tag{5}$$

$f_{imptnb}(v_i)$ is as shown in (6). The numerator represents the number of v_i's neighbors that the shells are bigger than v_i, which means the number of the important neighbors. The greater value means the number of the important neighbors is larger. The denominator represents the degree of the node v_i and $f_{imptnb}(v_i)$ means the important neighbors of v_i ..Max(KS) represents the maximum KS value of the network.

$$f_{imptnb}(v_i)=\begin{cases}\dfrac{\sum_{m\in N(vi)}\left| KSv(m)\geq KSv(v)\right|}{\sum_{m\in N(vi)} a_{vm}} & KSv(n)=Max(KS)\\[2ex] \dfrac{\sum_{m\in N(vi)}\left| KSv(m)> KSv(n)\right|}{\sum_{m\in N(vi)} a_{vm}} & KSv(n)<Max(KS)\end{cases} \tag{6}$$

The degree can only reflect the local influence of the node. And for many nodes, the k-shell value can give them the same shell value. So we comprehensively consider whether the neighbors of the node have influence. If the node has many important neighbors, it is easier to choose the node to spread information through these important neighbors.

Using the number of shells to measure whether the node has neighbors can continue to spread the information in the form of cascade diffusion.

The influence range is defined as follows:

In the network, select a node and use the propagation model to simulate the transmission. When reaching a certain times of the transmission, we can get the influencing nodes of the node and take the average value of the node as the influence range of the node. The greater the value is, the more important the node is.

To take Fig. 2 as an example, based on the four node influence measurement methods to calculate the node influence, as shown in table 2. Among them, the first column is the node identity, the second column is the degree of the corresponding node, the third column is the betweenness of the corresponding nodes, the fourth column is the K-Shell value of the corresponding node, the fifth column is the node influence value $DKIN(v_i)$ and the sixth column is the $Drank(v_i)$ that the $DKIN(v_i)$ value corresponding to.

TABLE II. Node Influence Ranking Of The Nodes In Fig.2

v	*Cd(v)*	*Cb(v)*	*Ks(v)*	*DKIN(v)*	*Drank(v)*
1	5	0.36601	4	0.36785	3
2	4	0	4	0.36785	3
3	4	0	4	0.36785	3
4	4	0	4	0.36785	3
5	4	0	4	0.36785	1
6	3	0	3	0.59642	1
7	6	0.75817	3	0.59642	1
8	3	0	3	0.59642	1
9	3	0	3	0.59642	1
10	2	0	2	0.43571	2
11	3	0.20915	2	0.43571	2
12	2	0	2	0.43571	2
13	1	0	1	0.43571	2
14	1	0.56863	1	0.36785	3
15	1	0	1	0.43571	2
16	1	0	1	0.36785	3
17	1	0	1	0.36785	3
18	1	0	1	0.36785	3
19	1	0	1	0.36785	3

IV. ALGORITHM DESCRIPTION

Algorithm: DKIN (degree and k-shell and important neighbor)

Input: A complex network G= (V, E)

Output: DKIN(vi), Drank(vi) and the influence range of the node

Begin

```
    1) Calculate the degree of each node and calculate fdegree(vi)
    2) Do K - shell decomposition to the network and calculate fk-shell(vi)
    3) Calculate the important neighbor indicator fimptnb(vi) for each node
    4) Set the weights of α, β, γ and calculate DKIN(vi)
    5) Select k nodes whose DKIN(vi) value are the largest
    6) Select one each time, use different probabilities and use the Independent Cascade Model to simulate the transmission process
    7) Calculate the average influence range to the whole network
    End
```

V. MODEL AND EXPERIMENTAL ANALYSIS

The experiment has been done in 2 real social networks datasets. The real dataset is the p2p-Gnutella08 network[13] and Wiki_Vote Wikipedia vote data[14]. All the data are from the Stanford University SNAP website. P2p-Gnutella08 is a directed P2P network in August 2002, in which the nodes represent the hosts and the edges represent connections between these hosts. Wiki_Vote is a directed network and points represent users. A points to B means A vote to B for whether B can become an administrator (by February 2008). The related information of the dataset are shown in table 3, where v is the number of nodes, e is the number of edges, max (k) is the maximum degree of nodes, d is average number of the shortest paths between nodes and max(ks) is the largest K-shell value in the network.

TABLE III. THE BASIC SITUATION OF EXPERIMENTAL DATASETS

Datasets	*Nodes*	*Edges*	*Max(k)*	*Max(ks)*
p2p-Gnutella08	6301	20777	98	10
Wiki-Vote	7115	100762	1065	53

The degree distributions of the two datasets are shown in Fig. 3 and Fig. 4, and the two networks are the uniform networks which are approximate to the power distribution. Each network has nearly half the number of nodes whose degree are 1, which indicates these nodes only have one neighbor. In the case of a very small probability of transmission, these nodes are difficult to become the key node in the network.

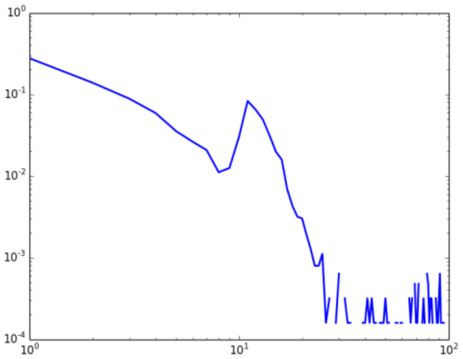

Figure 3. Degree distribution of the nodes in p2p-Gnutella08

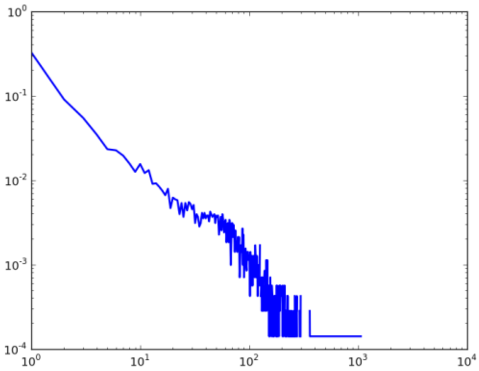

Figure 4. Degree distribution of the nodes in Wiki-Vote

Based on the independent communication model, the proposed DKIN index and K-shell method are analyzed and compared. In the process of experiment simulation, the top 10 nodes in the two methods are selected as the initial nodes, and we select the small probability of transmission. Because under the large probability, the transmission will be very large and can’t reflect the function of the key nodes. Repeat 1000 times to each seed node and calculate the average value as the result. Meanwhile, obtain the average propagation range of each method. In the equation of DKIN, if the value of the certain parameter is larger, it means it focuses much to reflect this index. According to the analysis and many experiments, we take α to 0.3, β to 0.4 and γ to 0.3.

A. P2p-Gnutella08 Dataset

Calculating the influence range of DKIN and K-shell algorithm under the three cases of p=0.05, p=0.075 and p=0.1.

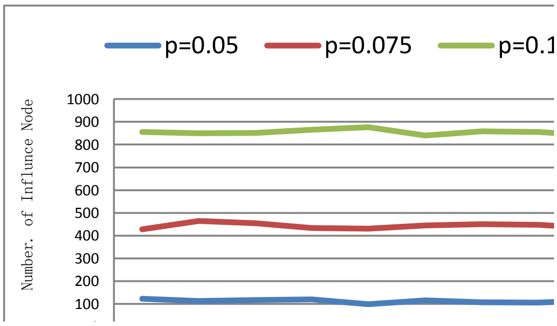

Figure 5. DKIN ranking propagation effect in p2p-Gnutella08 dataset

Fig. 5 shows ranking for the network using the DKIN algorithm and the top 10 nodes are 123, 367, 249, 127, 427, 149, 124, 352, 177. We can see, in the same probability, the propagation range of these nodes are relatively uniform. When the probability is smaller, the influence of nodes are smaller. With the increase of the probability of transmission, the influence range will be significantly increased.

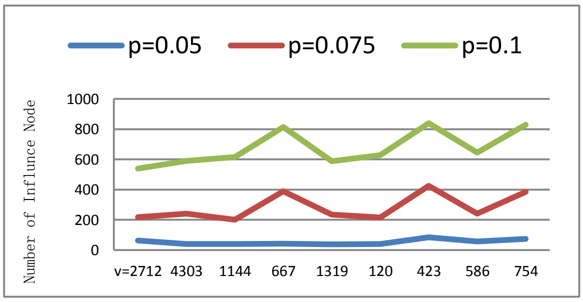

Figure 6. K-shell ranking propagation effect in p2p-Gnutella08 dataset

Fig. 6 is the K-shell ranking of the network. The K-shell value of the 10 nodes selected randomly in the tenth shell is the largest. We can see, in the same probability, the propagation range of these nodes are larger, and the fluctuation is not obvious when the probability is small. But they are not larger than the influence range of DKIN. With the increase of the probability of transmission, the influence range will be significantly increased, the curve will be volatile and the propagation effect of these nodes not good enough as that of the node using DKIN ranking to select. It means that the influence is not accurate if we only use K-shell value to measure and the effect is not the best.

B. Wiki_vote Dataset

Because the Wiki-Vote network has many edges, if the probability of transmission is larger, it is easy to spread to the whole network from the critical nodes. So we complete the simulate propagation experiments under the three cases of p=0.025, p=0.05, p=0.075, and calculate the influence range of the nodes which are selected using DKIN and K-shell algorithm.

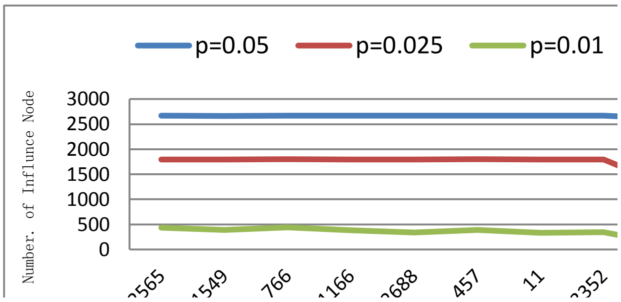

Figure 7. DKIN ranking propagation effect in Wiki-Vote dataset

Fig. 7 shows the ranking to the network using the algorithm DKIN. We can see that under the same probability, the propagation range of the nodes which are in the top of the ranking is relatively uniform and the curve is smooth. The higher the ranking is, the greater the influence range is. When the probability is smaller, the number of influenced nodes is smaller. With the increase of the propagation probability the influence range will be significantly increased.

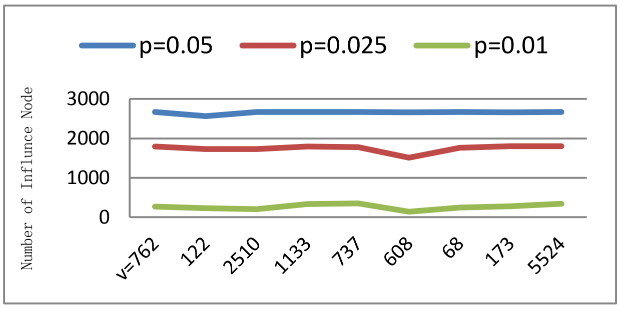

Figure 8. K-shell ranking propagation effect in Wiki-Vote data set

Fig. 8 is the ranking to the network using K-shell. Select 10 nodes randomly from the top 53 shell, and each time chooses one node to simulate the propagation, then you can see in the case of p=0.01, K-shell, p=0.05, the propagation range of the nodes which are selected using K-shell algorithm are different. There are obvious ups and downs, under the case of p=0.05, the curve of influence range is relatively flat, but the propagation effect of these nodes are not as good as DKIN rankings.

VI. Conclusions

Through the research on the node influence of complex network, we find that the method proposed by this paper can better identify influential nodes, and can adjust the parameters according to the different structure of the network so as to be applied to different complex networks. And the problem for the same KS value in the K-shell method has been solved. For the case of single propagation source, the method of this paper can well identify the most influential nodes. For the case of multiple propagation sources, because of the existence of cross propagation, the nodes with lager K-shell value are distributed in the core position, and the nodes with large degree may be distributed in the area of different networks. So the propagation effect using the nodes with large degree may be better. In addition, the community structures of some networks are obvious. For these networks, we can consider joining the community factor. How to improve the situation based on multiple propagation sources is the next step.

References

[1] X. F. W,X. L, Complex network theory and its application[M], Tsinghua University press,2006.

[2] N. H, D. Y. L, W. Y. G and X. Z., "Mining Vital Nodes in Complex Networks[J]," Computer Science, vol. 34, pp. 1–5, 2008.

[3] Freeman L C.Centrality in social networks conceptual clarification[J].Social networks,vol.1, 1979, pp. 215-239.

[4] Freeman L C.A set of measures of centrality based on betweenness[J].Sociometry, 1977, pp. 35-41.

[5] Kempe D, Kleinberg J, Tardo E. 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, pp.137-146.

[6] Page L, Brin S, Motwani R, et al. The PageRank citation ranking :Bringing order to the web[J]. 1999.

[7] Carmi S, Havlin S, Kirkpatrick S, er al. A model of Internet topology using k-shell decomposition[J]. Proceedings of the National Academy of Sciences, vol.104, 2007, pp.11150-11154.

[8] Q. C.H, Y. S. Y, P. F. M, Y. G,et al. A new approach to identify influential spreaders in complex networks[J]. Journal of Physics, vol.62, 2013, pp.140101-140101, **doi**:10.7498/aps.62.140101.

[9] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, vol.6, 2010, pp.888-893.

[10] J. G. L, Z. M. R, Q. G, B. H. W, Node importance ranking of complex networks, Journal of Physics, vol.62, 2013, pp.178901-178901, **doi**:10.7498/aps.62.178901.

[11] Sabidussi G. The centrality index of a graph[J]. Psychometrika, vol.31, 1966, pp.581-603.

[12] W.W.Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research, vol.33, 1977, pp.452-473.

[13] J.Leskovec, J.Kleinberg and C.Faloutsos. Graph Evolution:Densification and Shrinking Diameters. ACM Transaction on Knowledge Discovery from Data(ACM TKDD) ,vol.1, 2007.

[14] J.Leskovec, D.Huttenlocher, J.Kleinberg. Signed Networks in Social Media. CHI 2010.