The analysis of the structure of micro-blog users relation networks

Kai Yang and Ning Zhang

Business School University of Shanghai for Science and Technology, Shanghai Province, China
yang_kai_2008@163.com, zhangning@usst.edu.cn

Abstract - In this paper, the individual micro-blog is taken for example to research the structure characteristics of the micro-blog networks.Thedata of individual micro-blog is tracked, and complex networkisbuilt. Some static properties of one network are given - the degree distribution, clustering coefficient, betweenness, and average path length, it isfound to have short average path length and highaverage clustering coefficient. The distributions of out-degree, in-degree post present power-law characters. By analyzing the node centrality of micro-blog networks and describingthree network centrality indicators, including node degree, betweenness and k-Core,this paper discusses importance of networks' nodes and the impact on the propagation of information. The research shows that k-Core can be accurate representations of the core of the network location, which helps us to identify influential nodes in theinformation dissemination network. From the types of nodes, it is clear understanding of users' interests and behaviors.

Index Terms - Micro-blog Networks; Power-law Distribution; Small-word; Node Centrality

1. Introduction

With the development of Web 2.0 and the popularity of computer networks, online social networks have become an indispensable tool in human's life and work [1]. Micro-blog has been an important social network platform in recent years. Comparing with other traditional social network platforms, micro-blog is charactered by simple information release and timely and efficient information propagation. Because of the unique features of structure and information propagation, micro-blog plays a pivotal role in various aspects, like information propagation, opinions collection, exchange of information, emergency forecasts, marketing and so on.

Some papers researched the structure of micro-blog users relation networks at home and abroad. The empirical study of Twitter made by HaewoonKwak et al [2] found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. In order to identify influentials on Twitter, they have ranked users by the number of followers and by PageRank and found two rankings to be similar. Akshay Java et al. [3] present our observations of the micro-blog phenomena by studying the topological and geographical properties of Twitter's social network. They found that people used micro-blog to talk about their daily activities and to seek or share information. They analyzed the user intentions associated at a community level and showed how users with similar intentions connect with each other. Sina micro-blog network were researched by Fan Pengyi, etc. [4]. The network had apparent small-world effect and scale-free characteristic, specially, the out-degreedistribution appeared to have multiple separate power-law regimes with different exponents. They also observed the overlay graph of SinaMicro-blog represents assortative mixing pattern and weak correlation of in-degree and out-degree.Guozhengbiao et al [5] mainly studied the topology of Sinamicro-blog, and found that: the overlay of micro-blog is dynamic, most of the links between users are one-way, there is a core network in Sinamicro-blog, and the radius of Sine micro-blog is very short, which make it different from human social networks and other online social networks.

In this paper, after presenting empirical data thatwe extract the individual micro-bolgrelation users, construct complex network, analyzethe properties of the network, and study the structure of the personal micro-blogusers relationnetwork based on complex network theory.Finally , the node centrality of the network is given, we analyze the import nodes to identify the influentials of the information propagation.

2. RelatedTheory

Micro-blog users relation networks are the directed unweighted networks. Social network analysis is based on the importance of relationships or links between interactive units or nodes [6]. Some descriptivemeasures to be considered are:

Betweenness [7]: Represents how much a node is part of all the shortest paths between any given nodes. In other words, vertices that occur on many shortest paths between other two vertices have higher betweenness than those that do not.

$$BC_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}} \quad (1)$$

whereg_{st} is the total number of shortest paths from node s to node t and n_{st} is the number of shortest paths from s to t going through i.

The k-shell decomposition[8]: Nodes are assigned to k-shells according to their remaining degree, which is obtained by successive pruning of nodes with degree smaller than the $k_s$value of the current layer. We start by removing all nodes with degree k = 1. After removing all the nodes with k = 1, some nodes may be left with one link, so we continue pruning the system iteratively until there is no node left with k = 1 in the network. The removed nodes, along with the corresponding links, form a k-shell with index $k_s = 1$. In a similar fashion, we iteratively remove the next k-shell, $k_s = 2$, and continue removing higher k-shells until all nodes are removed. As a result, each node is associated with a unique k_s index, and the network can be viewed as the union of all k-shells.

© 2013. The authors - Published by Atlantis Press

The centrality of nodes, or the identification of which nodes are more “central” than others, has been a key issue in network analysis. Linton C.Freeman (1978) [9]argued that central nodes were those “in the thick of things” or focal points. Freeman summarizes three network centrality indicators, including node degree, closeness, betweenness.Degree is the number of nodes that a focal node is connected to, and measures the involvement of the node in the network. Its simplicity is an advantage: only the local structure arounda node must be known for it to be calculated. However, there are limitations: the measure does not take into considerationthe global structure of the network. A main limitation ofcloseness is the lack of applicability to networks with disconnected components: two nodes that belong to different components do not have a finite distance between them. Thus, closeness is generallyrestricted to nodes within the largest component of a network. The last of the three measures, betweenness, assess the degree to which a node lies on the shortest path between two other nodes, and are able to funnel the now in the network. In so doing, a node can assert control over the now.

3. Data Description AndNetwork Construct

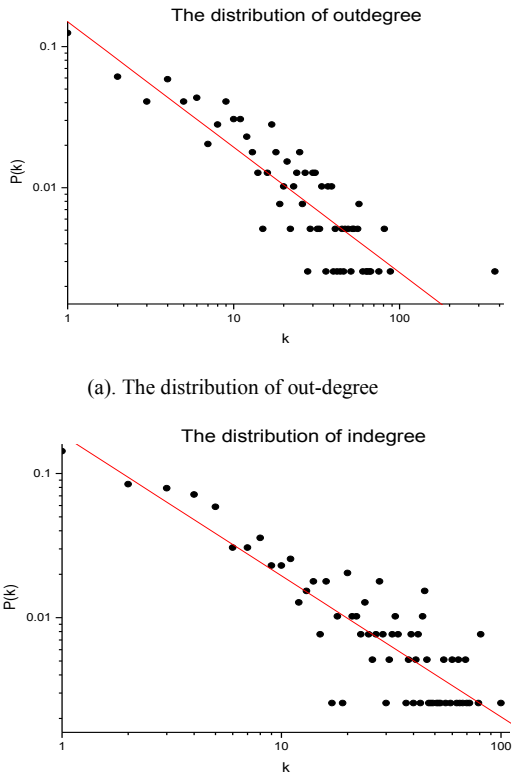
In this paper, we research Sina micro-blog, extract a common active user –the ID is"2010694415" (recorded as the user A), collect the user associated with the user A directly (his followings and followers). We take users as vertices, using users’ following and followed relationship construct micro-blog users relationnetwork.The users who are not directly associated with the user Aare not within the range of the network. We collect the data of the user A on a certain stabilization period,it showed the user's long-term variation in stable stage. The total number of nodes andedges of the networks is 392, 6049. We comparativelyanalysethe topological characteristics of the networks.

4. The Structure of Micro-blog Users Relation Networks

In this section, we analyze the structural properties of micro-blog, such as degree distribution, the average path length and clustering coefficient.

A. The scale-free characteristic

Scale free characteristic has frequently appeared in variousreal-world networks such as WWW, protein networks, e-mailnetworks, actor networks, scientific collaboration networksand some online social networks such as Twitter, Flickr andYoutube. It means that these networks are heterogeneous anddisplay power law shaped degree distribution $P(k) \sim k^{-\gamma}$ [10]. The micro-blogusers relationnetworksare the directed network, so the degree distribution includes in-degree and out-degreedistribution. Out-degree is the number of followings relationship of vertex;In-degree is the number of followed relationship.Based on the topological dataset of SinaMicro-blog, the scatter plot of N ’s double logarithm degree distribution is shown in figure 1.



(a). The distribution of out-degree

(b).The distribution of in-degree

Fig. 1Degree distribution of sina-microbolg

We observe the distribution of in-degree and out-degree has apparent power-law characteristics. After removing head, out-degree and in-degree respectively obey 0.97788 and 0.88854 power law distribution, which has obvious scale-free feature.

B. The small world characteristic

We employ two measurement properties to valid the smallworldcharacteristics of Sina-Micro-blognetwork, that is theaverage path length and clustering coefficient. Average pathlength is defined as the average number of steps along theshortest paths for all possible pairs of network nodes. It is ameasure of the efficiency of information or mass transport on anetwork. Clustering coefficient is a measure of degree to whichnodes in a graph tend to cluster together, including globalclustering coefficient and local clustering coefficient. Due tothe global definition can depict the clustering characteristic ofwhole network more precisely, we select it as our computingindicator, and it is defined as follow:

$$C = \frac{3 * \text{number of triangles}}{\text{number of connected triples of vertices}} \quad (2)$$

We calculate the average path length and clustering coefficient of the micro-blognetwork,that are 3.79876,0.3283. Compared with the same scale random network, we also calculate the average path length and average clustering coefficient of random network,that are 4.67257 and 0.08098.Micro-blog network has smaller average path length and larger clustering coefficient than the same scale random

network, shows that the smallworld characteristics. On the other hand it also shows that the micro-blog network has close connection, which is conducive to the spread of information in the network.

5. Node Centrality

In micro-blog system, users can send and read text based posts composed of up to 140 characters, called tweets, which are displayed on the user's profile page. Users can subscribe to other users' tweets – this is known as followings and subscribers are known as followers. Micro-blog users can browse information, share information, enjoy a different culture, retweet or comment on the information of interest. Centrality can describe vertex's position in the network, so as to find the center of the network. For social networks, we usually analyze centrality of degree, betweenness and k-shell.

We research node centrality based on complex network. According to our results, micro-blog users are classified[11], the first class in accordance with the user's relationship: friends, classmates, entertainers, strangers; the second class according to the interests of the user A: media, video, education, leisure, life. The top 20 users ranked by betweennessand degree are listed in table 1. That two indicators represent respectively the active nodes in the activities of information propagation (degrees), ability to control the flow of information (betweenness).

From Table 1, we can see some special nature of the network. Because the user A as the central node in these networks, the betweenness and out-degree of the user A is the largest in the networks, the table shows the interests and the focus of the user A. Ranked by betweenness, the tops of users are education and campus which is more important and control the information for the user A; User A is a student, these education-related users connect with user A tightly in network. Behind of the list, the types of users are video, emotion and entertainment stars, a lot of entertainers and the leisure user are followed by the user A, indicate that user A is interested in the entertainment. In-degree reflects the influence and activity of the users. Itis not difficult to find that the tops of users are entertainment stars and media ranked by in-degree. These users is relatively high activity, has a lot of fans, known as "power center" or "opinion leaders"[7] in the network, when they post a message, the information will spread quickly, these users are the main disseminator of information. These nodes are the motion of information propagation, the influence of the users is quite large, but rather those users are the interests of the user A, it is conducive to get the information for user A. Out-degree represents the user's level of followings and the ability to obtain information. Ranked by out-degree, the tops of the network generally collect information, we can see that these users are mostly some leisure or entertainment media, indicating that the ability of information propagation of these users is large. These users not only collect a lot of information, but also are able to spread information, make other users can obtain information faster in the network, play a key intermediary role in the information propagation. They include user A's classmates and friends, that User A also follow the reality people that know each other, to pay attention to their dynamics.

Of course, there are the users that are less close relationship with disseminators of information and rank behind in the network, they are the perimeter of the network; out-degree and in-degree (activity and influence) in the network is relatively small; the information is transmitted only in a small range. However, they echo information and thereby form the secondary propagation, which is important for information propagation, if a message can stir the response of these nodes, then the influence of information propagation will be improved. In a sense, the acceptance of these nodes play more important role in the scope of information propagation.

Some nodes show "community structure" in the network,i.e., groups of vertices that have a high density of edges within them in the core of the network, and are the key nodes of information propagation. From another perspective, how to promote sorting methods based on the value of the nodes is important, that is k-shell decomposition method. Using concept of k-shell, we separate the core network of the users, and analyze it. The largest k-shell of the network is 34 - shell, using Pajek software, we draw 34 - shell network diagram as shown in figure 2.

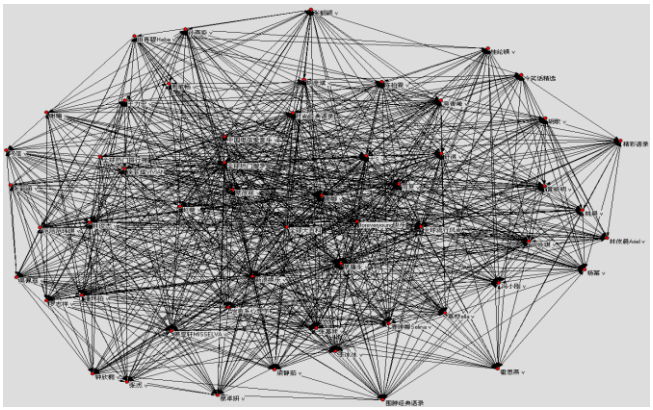


Fig.2The 34-shell of network on time point 3

From figure 2 can be found that the out-degree and in-degree of each node is very large, and these nodes connect special closely, so these users form a connected group, micro-blog users relation network show a good information communication network, this nature is more conducive to the spread of information. Fromthe figurewe can see that these users in addition to the user A are mostly entertain stars or news media, and related research shows that the nodes with the same characteristics are more easily connected. User A focus on the entertainment stars, his interestsare the film and television entertainment.Besides he focuses on the activity of students and spends most of the time in leisure entertainment and news in spare time. These users have strong relations in the core position of the network, play a central role in the propagation of information on the micro-blog. When one of the users send a message, other users receive the information, retweetit to the next user, so that it soon spreads on the entire users network. Thus the users become the key nodes in the information propagation. These nodes play a crucial role in the network.

6. Conclusions

This paper presented the structure of micro-blog users relation network as well as influentials related to the information.

Sina was described and classified as one of the most rapidly growing social networks; its simplicity allows applying Social Network methodologies along time, such as strongly connected components, in degree, clustering coefficient. The technical mechanisms were also mentioned. A series of measurements were performed over the social network.

Besides, we research the node centrality of the network, find that the interests and behaviors of users, and the effect of information propagation.

Finally related work was surveyed showing that potential research opportunities might be found in online social network dynamics, suggesting micro-blog and its rapid growth as a good exercise to perform measurements.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 70971089, Shanghai Leading Academic Discipline Project under Grant No. XTKX2012, and the Innovation Fund Project For Graduate Student of Shanghai (JWCXSL1202).

References

[1] Hu Haibo，Wang Ke，Xu Ling,Wang Xiaofan. Analysis of online social networks based on complex network theory[J]. Complex Systems and Complexity Science,2008,5(2):1-12.

[2] Kwak H, Lee C. "What is Twitter，a social network or a news media?" . Proceedings of the 19th international conference on World wide web ACM, pp.591-600,2010.

[3] Akshay J, Song X D., Tim F, Belle T. Why We Twitter: Understanding Micro-blogging Usage and Communities. In Proceedings of Joint 9th WEBKDD and 1st SNA-KDD Workshop 07,2007.

[4] Fan P Y, Li P, Jiang Z H, et al. Measurement and analysis of topology and information propagation on Sina micro-blog[C].Proceeding of IEEE International Conference on: Intelligence and Security Informatics. New York: IEEE Press 2011:396-401.

[5] Guo Z B, Li Z T, Tu H. Sina Micro-blog: An Information-driven Online Social Network. cw, pp.160-167,2011 International Conference on Cyber worlds,2011.

[6] Coulon, Fabrice, "The use of Social Network Analysis in Innovation Research: A literature review", retrieved online on March 2nd 2009 from: http://bit.ly/2Im7IM, 2005.

[7] Barthélemy, Marc. Betweenness centrality in large complex networks. Eur. Phys. J. B 38, 163–168 (2004).

[8] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11): 888-893.

[9] Freeman L C. Centrality in social networks conceptual clarification[J]. Social networks, 1979, 1(3): 215-239.

[10] S. Boccaletti, V. Latora, Y.Moreno, M. Chavez, and D.U. Hwang. Complex networks: structure and dynamics. Physics Reports, 424(2006): 175-308,2006.

[11] Wu Xinmiao, Wang Jianmin.Micro-blog in China:Indentify Influential Users And Automatically Classify Posts on Sina Micro-blog.Journal of Ambient Intelligence And Humanized Computing 2012, DOI: 10.1007/s12652-012-0121-3

TABLE 1Top 20 Users Ranked by Betweenness, In-degree, And Out-degree

	Ranked by betweenness			Ranked by in-degree			Ranked by out-degree		
	Betweeness	Users'Name	Type	In-degree	Users'Name	Type	7	Users'Name	Type
1	0.302907	foreveryoung	Sudent	100	蔡康永	Entertainment star	375	foreveryoung	Student
2	0.230545	新浪教育	Education	81	新周刊	News media	88	全球大百科	Life
3	0.212094	上海理工大学	Campus	81	头条新闻	News media	81	全球热门搜罗	Leisure
4	0.209869	sinjay	Friends	81	姚晨	Entertainment star	81	我彻底笑抽了	Leisure
5	0.208866	上理俱乐部	Campus	79	谢娜	Entertainment star	75	新浪娱乐	Leisure
6	0.051016	新浪视频	Video	72	李冰冰	Entertainment star	67	中国电信爱音乐	Music
7	0.047549	冷笑话精选	Leisure	70	杨幂	Entertainment star	66	新浪教育	Education
8	0.043109	头条新闻	News media	69	何炅	Entertainment star	64	治愈系心理学	Emotion
9	0.034557	潮人徐峰立	Education	69	李开复	Entrepreneur	63	时尚经典语录	Emotion
10	0.033488	精彩语录	Emotion	67	冷笑话精选	Leisure	60	优酷网	Video
11	0.030661	时尚经典语录	Emotion	65	赵薇	Entertainment star	57	土豆网	Video
12	0.030409	蔡康永	Entertainment star	64	冯小刚	Entertainment star	57	蔡康永	Entertainment star
13	0.030369	新浪娱乐	Leisure	64	新浪娱乐	Leisure	57	谢娜	Entertainment star
14	0.024902	黄晓明	Entertainment star	63	大 S	Entertainment star	56	sinjay	Friends
15	0.022402	谢娜	Entertainment star	60	舒淇	Entertainment star	56	新浪视频	Video
16	0.021806	杨幂	Entertainment star	60	范范范玮琪	Entertainment star	53	YouTube 精选	Video
17	0.01897	全球热门搜罗	Leisure	59	精彩语录	Emotion	53	全球流行风尚	Leisure
18	0.017227	土豆网	Video	56	黄晓明	Entertainment star	52	RenaChao 的世界	Leisure
19	0.015865	头条博客	News media	55	新浪视频	Video	52	黄晓明	Entertainment star
20	0.015371	治愈系心理学	Emotion	55	王力宏	Entertainment star	51	头条博客	News media