International Journal of Hybrid Information Technology
Vol. 9, No.10 (2016), pp. 211-218
http://dx.doi.org/10.14257/ijhit.2016.9.10.21

Network Characteristics Analysis of Microblog Public Opinion

Lirong Qiu[1], Jie Li[1], Zhen Hai Zhang[2] and Yi Lei Wang[2]

[1] *School of Information Engineering, Minzu University of China*
[2] *School of Information and Electrical Engineering, LuDong University*
qiu_lirong@126.com

Abstract

This paper uses public opinion propagation structure of Sina microblog as research object, it sets up simulation model of public opinion propagation structures based on complex network theory. Through the analysis of the degree distribution, clustering coefficient and community structure of the model, the validity of the model has been well confirmed. The paper also conducts research and analysis towards propagation quantity of the hot topic within 800 well-known spread microblog from 2013 to 2015, the paper establishes a community structure, through the simulation and demonstration of the community structure, it declares the community structure characteristic under the complex network environment.

Keywords*: Community structure; Network model; big data; Simulation analysis*

1. Introduction

A few years ago people named large volume of data as large-scale data, but in fact, this concept of Big Data had been proposed as early as 2008. In 2008, the 10th anniversary of Google establishment. Nature, the prestigious journal, published a special issue, it devoted to a series of technical problems and challenges of big data processing in future, which put forward the concept of Big Data [1].

In recent years, with the rapid development of computer and information technology, the popularization of the industry application system has expanded rapidly, and the data generated by the industry application has shown explosive growing trend. Hundreds of TeraByte (TB) or even tens to hundreds of PetaByte (PB) size of the data has been far beyond the existing traditional processing power of computing technology and information system. Therefore, searching for effective big data processing techniques, methods and tools has become the urgent needs around the world. Due to the importance and the urgent needs of big data processing, big Data technology has been highly concerned and valued in the global academic, industry and governments. Furthermore, the world has set off a research boom that can be compared with the 1990s information superhighway. The United States and some European countries have proposed a series of big data technology research and development plan at the national science and technology strategy level in order to promote the research and application of big data technology in government agencies, major industries, academic circles and industry.

With the development of semantic-net-featured infrastructures and data resources, the reform of organization becomes more and more inevitable. Big data will drive the network structure to generate organizational strength in unorganized form. The structural characteristics are firstly reflected in various decentralize Web2.0 applications, such as Really Simple Syndication (RSS), Wikipedia, microblog, etc [2].

Due to the extensive use of these applications, the complex network characteristics are studied. Using network science, we can explore a variety of complex systems in nature and human society. With the in-depth study of the physical and mathematical characteristics of the network, it is found that many real networks have a common

ISSN: 1738-9968 IJHIT

property, that is, community structure. That is to say, the whole network is composed of a number of "areas" or "group". Revealing the community structure in the network is very important for understanding the network structure and analyzing the characteristics of the network. [3]

With the rise of the study of complex networks, researchers began to focus on how to use complex network theory to study the online social network. Haewoon Kwak, et al through the analysis of the Twitter sample network to explore the structure and characteristics of the micro blog social network, and compared with the traditional online community. [4] Lu hang, et al analyzed the formation of public opinion in the microblogging process, to explore the microblogging become a powerful media of public opinion. [5] Shravan Gaonkar analyzed of the influence of Twitter from the mobile phone application platform Perspective. [6]

This research want to get the characteristic of sina weibo public opinion network by constructing a simulation network and analyzing the clustering coefficient of it, so that we can make full use of this network. For example, the mainstream view of the public on a hot topic is very import for the policymakers, they can make better decisions that conform to interests of the majority of society. And we can know the abilities of sina weibo public opinion network, so it can be better used to do more things.

2. Theory of Complex Network

2.1. Basic Concept

In short, Complex Network is the network that presents high complexity. Its complexity is mainly manifested in the following respects:

(1) Structure complexity: it is reflected in huge number of nodes and the network structure presents a variety of different features.

(2) Network evolution: it is performed in the produce and disappear of the node or link, such as the world-wide web network, in which web page or link may appear or disconnect at any time, leading to changes in network structure continually.

(3) Connection diversity: there exists difference in connecting weights between nodes, which may be directional.

(4) Dynamics complexity: the node sets may belong to nonlinear dynamical systems, such as the status of node can change intricately over time.

(5) Node diversity: nodes in complex networks can represent anything. For example, nodes of complex network constituted by interpersonal relationships represent separate individuals, it can represent different web pages when the complex network is composed of the World-Wide Web.

(6)The fusion of multiple complexity: all the complexities mentioned above can influence each other, leading to more unpredictable results. For example, when you design a power supply network, you need to consider the evolution of the network, and the evolutionary process would determine the topology of the network. When there exists energy transmission frequently between two nodes, the connection weights between them will increase, the network performance will be gradually improved through continuous learning and memory [7].

2.2. Average Degree and Degree Distribution

When describing the attribute of a single node, there is a simple and important concept: degree. In a network, the degree of a node V_i refers to the adjacent edge number K_i. However, the degree of the node is not necessarily equal to the number of nodes connected to it. Because there may be a lot of edges that connect any two nodes in the network. Moreover, the larger the degree of a node in the network is, the more important it is for this network. Add the sum of all nodes' degree of the network and calculate the

average value, then we can get the average degree of the network, as shown in formula (1).

$$< k >= \frac{1}{N}\sum_{i-1}^{n} i \quad (1)$$

Obviously, not all the nodes in the network have same number of edges. In other words, not every node has the same degree value.

We can use $P(k)$ to indicate the degree distribution of nodes in the network. That is to say, $P(k)$ is the possibility of randomly get a node in the network whose value of degree is k, as shown in formula (2). We can use histogram to illustrate the degree distribution of the network.

$$P(k) = N(k)/N \quad (2)$$

The $P(k)$ in this formula indicates the number of nodes whose value of degree is k.

2.3. Average Path Length

The shortest path is the path through the least edges between two nodes in the network. The paper uses d_{ij} to indicate the shortest length between any pair of nodes i and j. And the maximum length between any pair of nodes becomes the diameter of the network, we use D to indicate it. As shown in formula (3).

$$D = max_{i,j}\, d_{ij} \quad (3)$$

In the network, the average distance between any pair of nodes is called the average path length L. It has important significance for the study of the dispersion degree of network nodes. As shown in formula (4).

$$L = \frac{1}{\frac{1}{2}N(N+1)}\sum_{i>j} d_{ij} \quad (4)$$

2.4. Clustering Coefficient

There are circumstances that the friends you follow in Sina microblog may have followed each other. The clustering feature in complex network is called clustering coefficient, this feature can be used to describe the aggregation degree, the connection strength of each node. In the network, if there are K_i edges connect arbitrary node i and other nodes, then the number of neighbor nodes of node i is K_i. Therefore, it can be calculated that there are a maximum of $K_i(K_i - 1)/2$ edges between those K_i nodes. The calculation formula for the clustering coefficient of node i is shown in formula (5).

$$C_i = \frac{2E}{K_i(K_i-1)} \quad (5)$$

The clustering coefficient of the whole network is the average value of clustering coefficient of all nodes in the network. As shown in the formula (6).

$$C_i = \frac{1}{N}\sum_{i-1}^{n} C_i \quad (6)$$

2.5. Betweenness

If an individual is in the shortest path in the network, it can be said that this point has a very important position in the network. In other words, it has the ability to control the connection between two points, just like a bridge, or can be regarded as an intermediary, this is the betweenness we are going to talk about. The more the number of shortest path which passes this point, the bigger the point's betweenness is. According to this parameter, we can know the importance of nodes in the network.

2.6. Community Structure

With the further study of the physical meaning and mathematical characteristics of network, it is found that many real networks have a common character, which is the community structure. That is to say, the whole network is composed of a plurality of groups or clusters. The connections between nodes in each group are relatively close, but the connections between the groups are relatively sparse.

The research on the structure of network community has a long history. It has a close connection with graph partition in computer science and hierarchical clustering in sociology. A practical example of graph partition problem is parallel computing. If there are n mutual communicating computer programs distributed in g processors. But not every program is directly contact with other procedures. The communication patterns between all programs can be represented by a graph or network, a node in the graph represents a procedure, and each edge connects the two procedures which need to communicate directly.

Now the problem is how to put n programs into g processors, and make the number of programs running on each processor is approximately equal, in the meantime, the edge number between processors is the least, thus the communication between the processors is minimum(because communication between processors is relatively slow). In general, the exact answer to this segmentation problem is a Non-deterministic Polynomial problem. Therefore, there is no effective algorithms to obtain the accurate solution in the case of large graph. However, there are a lot of heuristic algorithms can get satisfactory solution in most cases. Among them, there are two famous algorithms: Kernighan-Lin Algorithm and Spectral Segmentation Method based on the Laplace Eigenvalue of Graph [8].

In the early 70s of 20th century, Zachary used two years of time to observe the social relationships between an American University karate club members. Based on the social relationships of the members in the club and the outside, he constructed a network of relationships between them. Using our method to analyze Zachary net, the results obtained are in agreement with the Zachary of the original network. When using the Kernighan-Lin Algorithm to analysis the complex network, the number of network societies and the size of them must be clear. But it's difficult to get this data, especially when the network is big and complex.

3. Construction of the Demonstration Network

3.1. Data Collection

The data collection process means, there are 800 bloggers from different industries who have a lot of fans and are relatively active, they are elected from the microblog fans list. And then select 70 events through the hottest search ranking list in recent years. By browsing microblog messages of the 800 bloggers in recent years, we tried to figure out the situation of each blogger's reaction to the 70 hot events and recorded it in a form. This article selects 800 bloggers as nodes of the network, it can reflect the network structure of the virtual community very well.

3.2. Data Reduction

The data collected above is in the form of text, and cannot be directly analyzed. It must be processed and converted into digital matrix form which can be directly processed by the analysis tool (this paper uses Matlab). When the popular microblogging bloggers and popular microblogging messages have been collected, arranging them in the form of Arabic numerals, and then convert them into tabular form. Next, transform the table data into the adjacency matrix by Matlab. So that the data can be analyzed by the software.

4. Analysis of the Demonstration Network

In order to better understand the degree of each point in the microblog public opinion, we have established a simulation network in which microblog serves as the main point, popular microblog serves as the edge, and then initialize a matrix.

After the initialization of the microblogging network, there are 800 points and 76 popular microblogging, each popular microblog is a side, if there is a blogger published on the hot spots, the number of the edge would plus one.

4.1. Degree Distribution

According to Table 1, one of the hottest microblog was forwarded by 166 bloggers, the blog was about Malaysia Airlines flight MH370 crash; the second hottest blog was forwarded by 153 bloggers, it was about the earthquake disaster. It is not difficult to find that people are highly concerned about the catastrophic disaster news. Although microblog news are frequently reported, such as the derailment of the famous celebrity, the forwarding amount is not very high. We can see that bloggers try to keep distance with the bad news, and concerned about the national affairs.

Table 1. Degree Distribution

5	10	13	22	27	30	31	34	43
166	56	51	61	51	58	52	153	67

According to the previous investigation, the degree distribution accords with the power-law distribution in regular situation, and it will accord with Poisson distribution if the degree distribution is random. And the decline of power-law distribution curve is much slower than exponential distribution curve. The simulation model is proved cannot meet the power distribution.

In double logarithmic coordinates, the degree distribution of the nodes from demonstration network is also very close to the power-law distribution. Due to the large number of nodes in the network, the degree distribution of nodes are denser, and the straight-line feature of negative slope is stronger. The power-law distribution characteristics of the virtual social network are preliminarily proved.

In the real network society, there are always some people often forwarded blogs and have lots of friends, while some others are isolated from the hustle and bustle, rarely voice in microblog. But most people remain at a normal level to satisfy the daily need, the actual situation is exactly in accord with the power-law distribution. It can be seen that the virtual social network and other network in the information field, such as AltaVista and Roche dictionary network, also have power-law distribution characteristics [9].

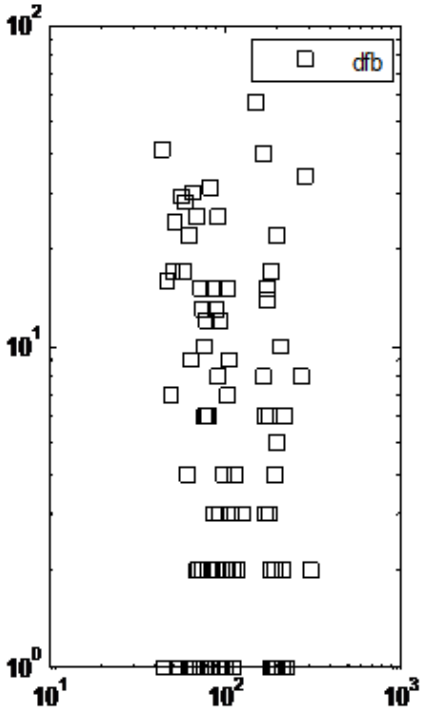

Figure 1. Degree Distribution

4.2. Clustering Coefficient

The clustering coefficient reflects the feature of social network, and the Sina Microblog which shares similar friend circle highlights this attribute. The results show that the clustering coefficient of the simulation network is 0.814, which means that the nodes in the simulation network are closely related. According to the experience, this is because social networks are more open and free than real life, the clustering coefficient of virtual network can be stronger, and the connection between people will be more closely. This changes the world into a global village.

According to the graph theory, clustering coefficient represents the aggregation degree of nodes in a graph. Evidence shows that in real network, especially in a particular network, due to the connection points have higher density, nodes always tend to establish a set of tight organizational relationships. In the real network, this possibility is often greater than the average probability of setting up a connection between two nodes randomly.

In many networks, if the node V_1 is connected to node V_2, and the node V_2 connected to node V_3, the node V_3 may connected to the node V_1.This phenomenon shows that dense connection between part of the nodes. It can be presented by the clustering coefficient (CC). In the undirected network, the clustering coefficient is defined as n, it represents the number of edges between the node V and its K neighbor nodes. However, the weighted value of microblog simulation model has no meaning, so the complete graph is an undirected graph.

4.3. Community Structure

The simulation model is processed in order to understand the community structure of the simulation network, and get effective data and complete graph. As shown in Figure 2, the simulation network has a strong community structure

characteristics, this shows that a variety of industries have different points of concern, and there also has a close link between the various associations, which states that there are always many topics cause widespread concern in the whole society. According to the different size of the community, we can know that bloggers from different industries have different forwarding amount towards popular microblog, this also shows that the entertainment circle is a very active part in the micro blog.

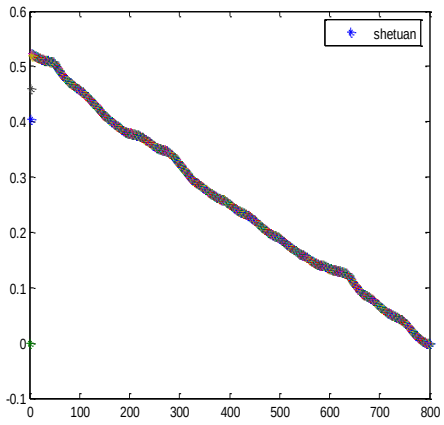

Figure 2. Community Structure (1)

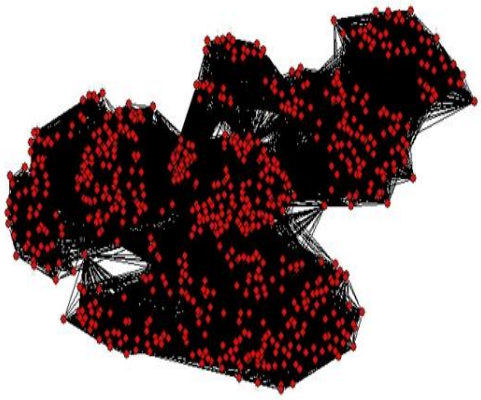

Figure 2. Community Structure (2)

Through the figure above, it can be seen that the simulation model has little to do with the community structure characteristics. This means that there is no obvious community structure in the complex network of microblog, it is also related to the high update speed of social information, and the rapid generation, dissemination and disappearance of a variety of different types of information in the network. The simulation model does not exist the basic characteristics of community structure, this may have the following reasons. Everyone concerned about different kinds of information, so one same popular blog will be forwarded by all walks of life. In addition, due to the timeliness of the microblog and the rapid update speed of information, people are more concerned about the new news, which also led to the balance of each point. Although the community structure is not obvious, parts of the characteristics of the community in the simulation network are still found in the empirical process.

5. Conclusion

This paper studies the relationship between the community structures of the Sina microblog network. Based on the complex network theory and the experimental data analysis, we established a simulation model about the community structure of Sina microblog network. At the same time, through analyzing the popular microblog forwarding quantity, we established the simulation model according to the characteristics. The comparison shows that the empirical data are in agreement with the simulation model, although there are some differences, it still can show the strong transmission capacity of Sina microblog. It can be used as a new way of communication in the future because it's faster than the traditional media, for example, the Sina microblog can be used to launch new products.

In the real world, there are many complex networks with small world or non-scale features. From the brain structure of a living body to a variety of metabolic networks, from Internet to World Wide Web, from large power networks to the global transportation network, from scientific research cooperation network to a variety of political, economic, social relations network and so on. At present, the research of various networks has been paid high attention in the world, and has become a very important and challenging frontier.

Acknowledgement

This research has been supported by the Nature Science Foundation of Beijing (No. 4153062), the National Technology Support Program (2014BAK10B03) and the Program for New Century Excellent Talents in University (NCET-12-0579).

References

[1] X.-F. Wang, L. Xiang and G. R. Chen, "Complex network theory and its application", Beijing: Tsinghua University Publication, **(2006)**.
[2] F. Wang, and H. Yang, "The author's research cooperation network model and empirical research, library and information work", vol. 51, no. 10, **(2007)**, pp. 68-71.
[3] C. H. Guo and L. Zhang, "An Analysis Method Based on PCA for the Community Structure in Complex Networks", Operations Research and Management Science, **(2008)**.
[4] H. Kwak, "What is Twitter, a social network or a news media", Proceedings of the 19th international conference on World Wide Web. ACM, **(2010)**.
[5] H. Lu and T. Zhang, "A preliminary study on the formation of public opinion in the micro-blog communication environment", no. 6, **(2010)**.
[6] S. Gaonkar and R. R. Choudhury. "Micro-Blog: map-casting from mobile phones to virtual sensor maps", Proceedings of the 5th international conference on Embedded networked sensor systems, ACM, **(2007)**.
[7] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the National Academy of Sciences of the United States of America, vol. 99, no. 2, **(2002)**, pp. 7821-7826.
[8] U. Brandes, "A faster algorithm for betweenness centrality*", Journal of mathematical sociology , vo. 25, no. 2, **(2001)**, pp. 163-177.
[9] M. EJ. Newman, "The structure of scientific collaboration networks", Proceedings of the National Academy of Sciences, vol. 98, no. 2, **(2001)**, pp. 404-409.

