International Journal of Database Theory and Application
Vol.8, No.5 (2015), pp.273-284
http://dx.doi.org/10.14257/ijdta.2015.8.5.24

Big Data Analytics of Multi-Relationship Online Social Network Based on Multi-Subnet Composited Complex Network

Gengxin Sun[1], Sheng Bin[2] and Yixin Zhou[3]

[1]*International College of Qingdao University, Qingdao, China*
[2]*Software Technical College of Qingdao University, Qingdao, China*
[3]*International College of Qingdao University, Qingdao, China*
[1]*sungengxin@qdu.edu.cn*, [2]*binsheng@qdu.edu.cn*, [3]*zyx@qdu.edu.cn*

Abstract

Online social networks such as Twitter and Facebook are becoming popular form of social information networks. There are frequently many kinds of relationships in an online social network. Complex Network acting as one kind of big data technologies is often used to analyze users' social activities. By studying the Douban network, which is a representative multi-relationship online social network in China, big data of friendship relationship and book comments similar relationship are crawled through network topology measurement software, from the perspective of topological characteristics of complex network, the basic topologies of the two relationship networks constructed individually by the two relationships are analyzed. Based on these, a multi-relationship online social network based on Multi-subnet Composited Complex Network Model is constructed through loading book comments similar relationship subnet to follower relationship subnet, and accurate understanding of topologies of Douban multi-relationship network is obtained. These findings provide a deep understanding on the evolution of multi-relationship online social network, and can provide guidelines on how to build an efficient multi-relationship online social network evolution model.

Keywords: *Complex Network, Big Data, Online Social Network, Multi-subnet Composited Complex Network*

1. Introduction

In recent years, with the development of Web 2.0, online social networks have become a major social exchange platform for people to interact with others. With the help of online social networks, people can share information, recommend favourite books or movies, make friends with like-minded individuals and so on. The popularity of online social networks has also attracted many researchers to study their topological characteristics [1-2], evolution models [3-5] and user behaviors [6-7].

Online social networks are typical examples of social network, there are many social relationships, say friendship, commerce, or others in social network. Traditional social network study can date back about half a century, the traditional researches the characteristics of structure and evolution law of social networks from the existing user relationship networks. Novel network structures of social network which integrates the theories of traditional social networks and modern complex networks have been revealed [8].

In this paper, we study the Douban, which is a popular Chinese online social network and media comments platform. Douban users can indicate their preference and comments on particular media items, such as books, movies, music, *etc*. Moreover, similar to Twitter, Douban also is a "friendships network" in which users can make online friends and broadcast short messages to their friends. We propose two user and implicit relationship. The explicit relationship is the tie of users of adding friends or following mutually. From

ISSN: 2005-4270 IJDTA

the behaviour of users, we can study the interest and preference of users, and then construct user profile. Based on user profile, we can calculate the similarity between users. If the similarity exceeds some given threshold, we can consider that there exists some implicit relationship between users. Douban is a multi-relationship online social network, friendship relationship is an explicit relationship, and book comments similar relationship is an implicit relationship. Through self-developed network topology measurement software, we have crawled measured data of friendship relationship and book comments similar relationship over 110,000 Douban users. We have analyzed and compared individually topological characteristics of the two relationship networks. Lastly, Based on Multi-subnet Composited Complex Network Model [9], Douban multi-relationship network are constructed by the two relationships, whose topological characteristics are also analyzed.

2. Topological Characteristics of Complex Network

Complex network is a graph (network) with non-trivial topological features that do not occur in lattices or random networks but often occur in real networks. Empirical analysis of social networks has been much studied by complex networks [10-12]. Online social networks are typical examples of complex network. Using complex network theory to analyze topologies of online social network is an effective way to understand accurately topology and evolution of online social network.

Complex networks display substantial non-trivial topological characteristics, with patterns of connection between their nodes that are neither purely regular nor purely random. Such characteristics include a heavy tail in the degree distribution, a high clustering coefficient, a small average path length, hierarchical structure and community structure [13].

The degree of a node in a network is the number of connections or edges the node has to other nodes. The degree distribution $P(k)$ of a network is defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a network and n_k of them have degree k, we have $P(k) = n_k / n$. A network is named scale-free [14] if its degree distribution follows a particular mathematical function called a power law. The power law implies that the degree distribution of these networks has no characteristic scale. In a network with a scale-free degree distribution, allowing for a few nodes of very large degree to exist, these nodes are often called “hubs”.

In complex network, clustering coefficient is a measure of the degree to which nodes in a complex network tend to cluster together. The clustering coefficient is based on triplets of nodes. A triplet consists of three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties. A triangle consists of three closed triplets, one centered on each of the nodes. The clustering coefficient is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed). Watts and Strogatz defined the clustering coefficient as follows, “Suppose that a vertex v has k_v neighbors; then at most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when every neighbor of v is connected to every other neighbor of v). Let C_v denote the fraction of these allowable edges that actually exist. Define C as the average of C_v overall [15]”.

Average path length is one of the three most robust measures of network topology, along with its clustering coefficient and its degree distribution. In a network, the distance $d(v_i, v_j)$ between two nodes v_i and v_j is defined as the

number of edges along the shortest path connecting them. Assume that $d(v_i, v_j) = 0$ if v_i cannot be reached from v_j.

A network is called a small-world network [15] by analogy with the small-world phenomenon. The small world phenomenon means co-occurrence of a small diameter and a high clustering coefficient in a network. It is known that a wide variety of abstract networks exhibit the small-world phenomenon, such as, random networks and scale-free networks. Further, real-world networks such as the World Wide Web and the social network also exhibit this phenomenon.

3. Empirical Analysis of Douban Users Friendship Relationship Network

Douban is a "follower network". A distinct feature of a following relationship in network is that the user being followed can provide useful information to all his followers. In Douban, user A can establish following relationship to user B without waiting for permission from user B. So the following relationship is directed.

Reciprocity is a quantity to specifically characterize directed networks. Link reciprocity measures the tendency of node pairs to form mutual connections between each other. A traditional way to define the reciprocity r is using the ratio of the number of links pointing in both directions to the total number of links. With this definition, $r = 1$ is for a purely bidirectional network while $r = 0$ for a purely unidirectional one. Real networks have an intermediate value between 0 and 1. Through statistical analysis over the crawled original data, we found that the reciprocity of Douban users friendship relationship network is higher than 0.86. So we regard Douban users friendship relationship network as a undirected network in the empirical analysis.

The big data was obtained from Douban through network topology measurement software. After removing duplicate bidirectional edges, we perform empirical analysis of the structure of Douban users friendship relationship network which is composed of 115, 460 nodes and 2, 235, 741 edges (we viewed this network as undirected one).

The maximum vertex degree of network is 10405. There are 12190 nodes with degree equal to 1. There are 41943 nodes with degree less than 6, which accounts for 36.3% of the total network nodes. The average degree of network is 38.728, average path length of network is 6.12 and clustering coefficient of network is 0.095. Modularity of network is 0.325.

In Figure 1, we can see clearly that $P(k)$ follows two different scaling with k, depending on the specified threshold value of degree k_c. $P(k)$ obeys a power law form $\sim k^{-\gamma_1}$ with $\gamma_1 = 0.72$ when $k < k_c = 30$. Otherwise, $P(k) \sim k^{-\gamma_2}$, where $\gamma_2 = 2.11$ for $k > k_c$. The degree distribution above the threshold degree k_c is consistent with findings of social networks with the degree $2 < \gamma < 3$. Whereas, for small degree k below k_c, the scaling exponent of $P(k)$ is far less than 2.

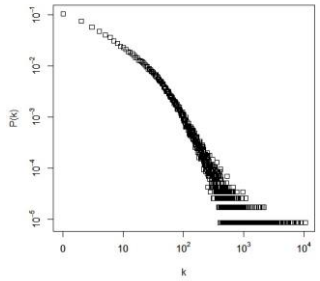

Figure 1. Degree Distribution for the Empirical Network

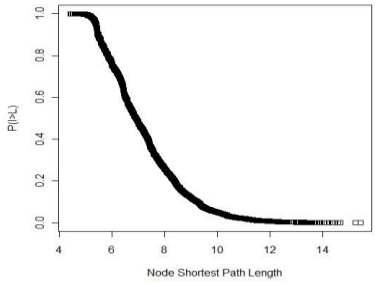

Figure 2. Cumulative Average Shortest Path Length Distribution for the Empirical Network

The average shortest path length is calculated, which is the mean of geodesic distance between any node pairs that have at least a path connecting them. The cumulative average shortest path length distribution of the empirical network is shown as Figure 2. From Figure 2, we can see that percentage of nodes whose average shortest path length exceeds 10 is very small, average shortest path length of most nodes is between 6 and 9. So the empirical network has small-world property, that is, high clustering coefficient and short average shortest path length.

Figure 3 plots the distribution of clustering coefficient $C(k)$ vs. degree k. The clustering coefficient distribution shown in Figure 3 suggests that the dependency of $C(k)$ on k is nontrivial, and points that empirical network has fundamental characteristic of hierarchy, low-degree nodes generally belong to well-interconnected clusters, while high-degree nodes are linked to many nodes that may belong to different groups.

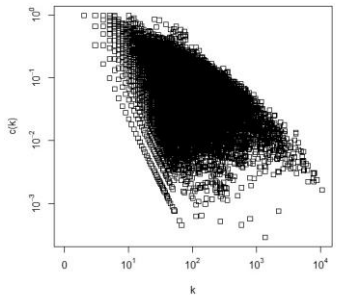

Figure 3. The Plot of Degree-Dependent Clustering Coefficient vs. Degree

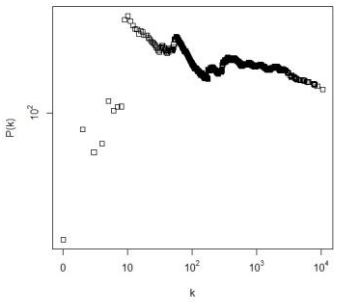

Figure 4. The Plot of Degree Correlation Distribution

Another important elemental characterizing of the complex networks is the degree correlation of node i and its neighbour. The degree correlation distribution is shown as Figure 4.

4. Empirical Analysis of Douban Book Comments Similar Relationship Network

The traditional researches of the topological characteristics and evolution law of social networks is always from the existing explicit relationship networks. But, in fact, implicit relationships often imply deeper user relationships.

Douban is a very typical social interest graph, which provides bibliographic recommendations, book comments services and so on. Interest graphs are used to create people's interest networks. Where online social networks are organized around an individual's friends, interest networks are organized around an individual's interests. Interest graphs can in some cases be derived from social networks and may maintain their context within that social network. These are specifically social interest graphs. For an interest graph to be accurate and expressive, it needs to take into account explicitly declared interests, for example "Likes" on Facebook, as well as implicit interest inferred from user activities such as books comments on Douban.

We obtained comments for about 80000 books of 115, 460 users of Douban users friendship relationship network. Comments contained user's score for book (score is from 1 to 5). If two users commentated on more than six books, and score for these books are basically the same, we can define that the two users have book comments similar relationship. After removing outliers, we perform empirical analysis of the structure of Douban book comments similar relationship network which is composed of 114, 122 nodes and 1, 651, 883 edges.

The maximum vertex degree of network is 4143. There are 2050 nodes with degree equal to 1. There are 42590 nodes with degree less than 10, which accounts for 37.3% of the total network nodes. The average degree of network is 28.949, average path length of network is 7.27 and clustering coefficient of network is 0.042. Modularity of network is 0.383.

In Figure 5, we can see clearly that the degree distribution is power law. Different from the largest number of leaf nodes in normal power law distribution, the largest number of nodes with degree is equal to 6 in the degree distribution of Douban book comments similar relationship network. It suggests that interest for books of most users is consistent.

The average shortest path length of Douban book comments similar relationship network is 7.27, clustering coefficient of network is 0.042, which is consistent with the small-world networks with higher clustering coefficient and shorter average path

length. It can explain that the empirical network is a small-world network. The cumulative average shortest path length distribution of the empirical network is shown as Figure 6.

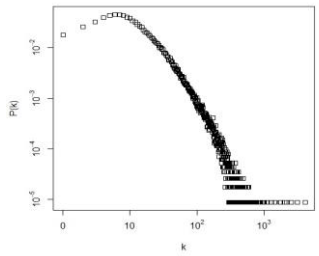

Figure 5. Degree Distribution for the Empirical Network

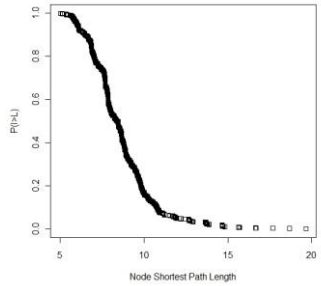

Figure 6. Cumulative Average Shortest Path Length Distribution for the Empirical Network

Like with Douban users friendship relationship network, the distribution of clustering coefficient $C(k)$ vs. degree k of Douban book comments similar relationship network shows approximate power-law, whose index is -2.3. It points that Douban book comments similar relationship network has apparent characteristic of hierarchy.

The degree correlation distribution of Douban book comments similar relationship network is shown as Figure 8.

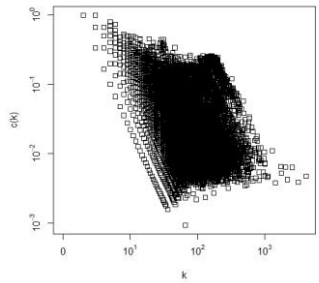

Figure 7. The Plot of Degree-Dependent Clustering Coefficient vs. Degree

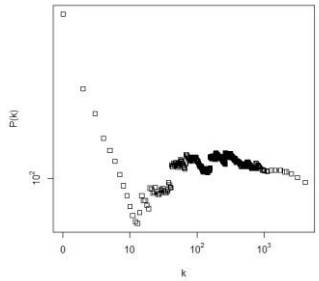

Figure 8. The Plot of Degree Correlation Distribution

In Figure 8, we can see clearly that when node degree is less than 10, with the increase of node degree, the average degree of neighboring nodes connected to the node decreases rapidly, when node degree is more than 60, the average degree of neighboring nodes connected to the node is no significant change. It means that the empirical network is neither assortative nor disassortative, It shows that the comments and interests for the book are made entirely determined by their own preferences.

5. Empirical Analysis of Douban User Multi-Relationship Network

Online social networks mainly study relationships between network users, and there are often multiple relationships between people. Classical complex networks mainly describe the same kind entities and their relationships. However, multi-subnet composited network is one model describing different kinds of entities and their relationships. Multiple complex networks could be composited and one complex network could be decomposed into multiple complex network by network operations presented by this model [16].

Multi-subnet composited network is defined as (V, E, R, F), where

(1) $V = \{v_1, v_2, ..., v_m\}$ is set of nodes, $m = |V|$ is the order of V;

(2) $E = \{< v_h, v_l > | v_h, v_l \in V, 1 \le h, l \le m\} \subseteq V \times V$ is set of edges;

(3) $R = R_1 \times ... \times R_i \times ... \times R_n = \{(r_1, ..., r_i, ..., r_n) | r_i \in R_i, 1 \le i \le n\}$ is relationships set, R_i denotes one kind of relationships set, n is the total number of relationships, R can be empty set;

(4) $F : E \to R$ is a mapping from E to R.

Communication method diversity of online social network users determines the diversity of relationships between network users. Some of these relationships are direct, explicit, sometime there are no direct obvious relationships between users, but user behaviors have obvious similarities, because user's behavior is performance of user's characteristics, implicit relationships existing between users may be analyzed based on such characteristics.

In this paper, relationships in online social network are divided into explicit relationship and implicit relationship. Explicit relationship refers to the direct relationships formed by adding friends or concerning users. Implicit relationship is from the perspective of user behavior, analyzing their interests and hobbies, the similarity between users is obtained through similarity calculation. If the similarity exceeds setting threshold, then there are implicit relationships between users. So Douban users friendship relationship network is an explicit relationship network, Douban book comments similar relationship network is an implicit relationship network. Using loading operation of Multi-subnet Composited Network Model, A

new Douban user multi-relationship network can be composed of Douban users friendship relationship network and Douban book comments similar relationship network.

A maximal connected subgraph of Douban user multi-relationship network is composed of 30, 453 nodes and 55, 856 edges. Every nodes denotes a Douban user, every edges denotes that there are friendship relationship and book comments similar relationship between the two users at the same time.

The maximum vertex degree of network is 3190. There are 13910 nodes with degree equal to 1, which accounts for 45.6% of the total network nodes. The average degree of network is 3.688, average path length of network is 5.18 and clustering coefficient of network is 0.162. Modularity of network is 0.418.

According to statistics, the maximum vertex degree of network is 3190, and nodes with degree less than 4 account for 85.2% of the total network nodes. It is consistent with characteristics of power-law degree distribution. The degree distribution of Douban user multi-relationship network is shown in Figure 9.

According to statistics, average path length of network is 5.18 and clustering coefficient of network is 0.162, which is consistent with characteristics of small-world network. The average shortest path length of all nodes in network are sequenced in descending, each node would correspond to a rank in the sequence. The correlation distribution between average shortest path length and its rank of each node is shown in Figure 10.

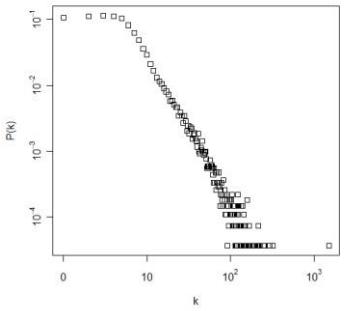

Figure 9. Degree Distribution for the Empirical Multi-Relationship Network

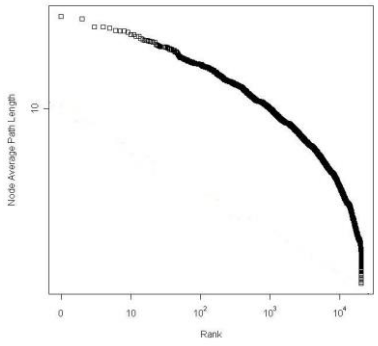

Figure 10. Correlation Distribution between Node's Average Shortest Path Length and its Rank

In Figure 10, we can see clearly that the correlation distribution is approximate power law distribution, and the value of average path length would decline sharply when the value exceeds a certain threshold.

According to analysis of node degree and node clustering coefficient, correlation between node degree and clustering coefficient is power law shown as Figure 11. From Figure 11 we can see clearly that except leaf nodes, correlation between node degree and clustering coefficient of other nodes show approximate power-law, whose index is -2, node clustering coefficient decrease continuously with increasing of node degree, the average clustering coefficient of lower degree node is higher than the higher degree node. So Douban user multi-relationship network has apparent characteristic of hierarchy.

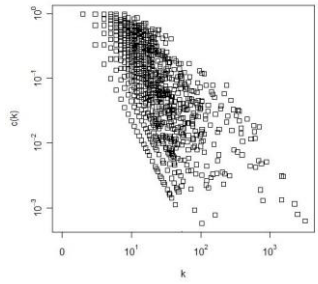

Figure 11. Correlation Distribution between Node's Degree and its Clustering Coefficient

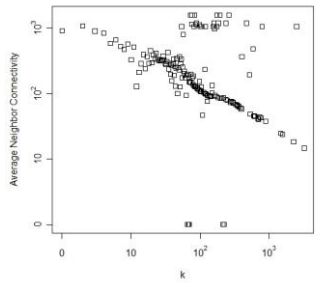

Figure 12. The Plot of Degree Correlation Distribution

The degree correlation distribution of Douban user multi-relationship network is shown as Figure 12. Figure 12 shows that with the increasing of node degree, the average degree of neighboring nodes connected to the node is gradually decreased, which indicates that Douban user multi-relationship network shows "disassortative mixing".

6. Conclusion

According to empirical analysis, we found that for the purposes of the node degree distribution, explicit relationship network, implicit relationship network and multi-relationship network all show power law characteristics. But node values of multi-relationship network are more dispersed, "Long tail" features are more obvious. It indicates that breadth and depth of multi-relationship network user network behavior are greater.

Explicit relationship network, implicit relationship network and multi-relationship network all show small-world characteristics of high clustering coefficient and low average shortest path, small-world characteristics of multi-relationship network is most obviously among them. Explicit relationship network and multi-relationship network shows "disassortative mixing". Whereas, implicit

relationship network do not shows obviously "assortative mixing" or "disassortative mixing". At this point, explicit relationship network and multi-relationship are different from traditional social networks, implicit relationship network is similar to technical networks. Modularity of explicit relationship network, implicit relationship network and multi-relationship network are all high, especially for multi-relationship network. It indicates that user community structure of explicit relationship network, implicit relationship network and multi-relationship network are very apparent, communication between users is easy to form community structure.

Acknowledgement

This paper is granted by the key research in Statistics Foundation of Shandong Provincial Bureau of Statistics (No. KT140215).

References

[1] Y. Ahn, S. Han and H. Kwak, "Analysis of topological characteristics of huge online social networking services", Proceedings of the 16th international conference on World Wide Web. ACM, **(2007)**.
[2] S. Wu, J. M. Hofman and W. A. Mason, "Who says what to whom on twitter", Proceedings of the 20th international conference on World Wide Web. ACM, **(2011)**.
[3] F. Fu, X. Chen and L. Liu, "Social dilemmas in an online social network: the structure and evolution of cooperation", Physics Letter A, vol. 1, no. 371, **(2007)**.
[4] J. Leskovec and L. Backstrom, "Microscopic evolution of social networks", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, **(2008)**.
[5] R. Kumar, J. Novak and A. Tomkins, "Structure and evolution of online social networks", Link mining: models, algorithms, and applications. Springer New York, **(2010)**.
[6] J. Jiang, C. Wilson and X. Wang, "Understanding latent interactions in online social networks", ACM Transactions on the Web, vol. 4, no. 7, **(2013)**.
[7] F. Benevenuto, T. Rodrigues and M. Cha, "Characterizing user behavior in online social networks", Proceedings of the 9th ACM SIGCOMM conference on Internet measurement. ACM, **(2009)**.
[8] B. Wellman, "Computer networks as social networks", Science, vol. 5537, no. 293, **(2001)**.
[9] F. J. Shao, R. C. Sun and S. J. Li, "Research of Multi-Subnet Composited Complex Network and Its Operation", Complex Systems and Complexity Science, vol. 4, no. 9, **(2012)**.
[10] M. Kurant and P. Thiran, "Layered Complex Networks", Physical Review Letters, vol. 13, no. 96, **(2006)**.
[11] A. Barabâsi and R. Albert, "Emergence of scaling in random networks", Science, vol. 5439, no. 286, **(1999)**.
[12] G. Kossinets, "Empirical Analysis of an Evolving Social Network", Science, vol. 5757, no. 311, **(2006)**.
[13] R. Albert, "Statistical mechanics of complex networks", Reviews of modern physics, vol. 1, no. 74, **(2002)**.
[14] A. L. Barabási, "Scale-free networks: a decade and beyond", Science, vol. 5939, no. 325, **(2009)**.
[15] D. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks", Nature, vol. 6684, no. 393, **(1998)**.
[16] F. J. Shao and Y. Sui, "Reorganizations of complex networks: Compounding and reducing", International Journal of Modern Physics C, vol. 5, no. 25, **(2014)**.

Authors

Gengxin Sun, received his Ph.D. degree in Computer Science from Qingdao University, China in 2013. He is currently an Associate Professor in the School of Computer Science and Engineering at Qingdao University. His main research interests include embedded system, operating system, complex networks, web information retrieval and data mining.

Sheng Bin, received her Ph.D. degree in Computer Science from Shandong University of Science and Technology, China in 2009. She is currently a lecturer in the School of Software Technology at Qingdao University, China. Her main research interests include embedded system, operating system, complex networks, cloud computing and data mining.

Yixin Zhou, received her Ph.D. degree in Computer Science from Qingdao University, China. She is currently an Associate Professor in the School of Computer Science and Engineering at Qingdao University. Her main research interests include complex networks, web information retrieval and data mining.

