Journal of Computers Vol. 28, No. 5, 2017, pp. 105-115
doi:10.3966/199115992017102805010

Study of Knowledge Discovery in Social Network Data Mining

Chang Zhang[12*], Han Yu[2], Qi Wang[2],
Yuan-yuan Ren[2] and Yan-ling Su[2]

[1] Beijing University of Posts and Telecommunications
100876, Beijing, China
airmagicz@163.com

[2] Shi Jiazhuang Posts & Telecommunications Technical College
050021, Shijiazhuang, China
{yuhan, wangqi, renrx, suyl}@cptc.edu.cn

Received 25 October 2016; Revised 15 February 2017; Accepted 16 February 2017

Abstract. Data mining as a high-tech subject is increasingly popular than before. Some data mining methods、data mining algorithms and application in social networks is introduced here. Especially, some specific social networks datasets are used to analyze the characters of social networks, the datasets are analyzed from aspects of link proportion、adjacency、out-degree、in-degree、similarity、clustering、distance、density、transitivity、groups and reachability. Some innovative data mining knowledge of social networks can be drawn by analyzing a crowd of data sets.

Keywords: clustering, data mining, distance, similarity, social networks, transitivity

1 Introduction

Modern sociology is major to study modern social development and systematical action in the society. Sociologists find there are mutual dependence and contact between social entities, and the related contact has a significant impact for every social entities. Based on this observation, they use network model to depict the relationship between social entities, furthermore, to analyze the implied rules in the social relations. Different from usual sociology methods, it provides a formal and conceptual way to treat the nature and development processes of society. So the social network is a very powerful means of sociology.

After the sociologists establish an accurate network model, logic reasoning can be used to study the social nature. However, because of the limitation of data collection, the early social network is limited to small groups; often contain only a few nodes. Some simple analysis of the nature and pattern can be draw out by using graph theory and statistics. But as the development of modern communication technology, more and more data are collected together, which make it possible to establish a large social network. For example, email log can be used to establish a contact network between the internet users. Therefore, the social network now has a larger size than before; it often contains thousands of or nodes, and even millions of node in a network. Face such a huge and complicated network, a simple mathematical knowledge and the manual processing have been impossible to carry out effective analysis.

Society network analysis pays attention to the link which is a very important character. From the aspect of data mining, the society network analysis is also called link mining. Through the mining of links, some richer and more accurate information can be got. At the same time, the link itself is often of concerned as information. For example, in certain circumstances, not all links are observed, thus people possibly interested in predicting the existence of link between instances. In other domains, links transform along with the time, then, our goal is based on the current observation to predict some links

* Corresponding Author

whether exist in some time in future. Furthermore, because of considering the links between data, the structure attribute of social networks also can provide important information, just like node degree, connectivity, and so on. At the same time, more complex patterns appear, like subgroup, so how to obtain more complex information about these patterns is becoming a big challenge [1].

Traditional data mining [2] uses a "property-value" table to represent data. Data is shown as vector; each dimension of vector corresponds to a value of conditional property. Social network data is a structured relational data, except the properties of each node, the more important thing is the link between nodes. The links contain a lot of information. However, the vector form shows the independence of nodes, ignore the link must not be good to knowledge discovery. Therefore, to analyze the social network data people should use the relational data mining methods which are rapid developing.

Relational data mining [3-4] discovers knowledge from the data which contain multiple objects that have links between themselves. It does not change the relational data into the single table mode, but mining the relationship mode in such data directly. The data type of relational data mining is more consistent with the application data in a relational database. So it can satisfy the current data and analysis requirements, Relational data mining become more and more important and many efficient algorithms are proposed.

Through the above it can be found that many research approaches and algorithms of social network data mining are proposed in recent years. But there are not many scholars to study the specific social network data and draw the conclusions based on the actual social networks, such as some small but real social networks, because it is quite difficult to summary conclusions from the dynamic and small datasets. However, specific social network datasets contain many useful and practical characteristics, a lot of valuable information can be got by a comprehensive analysis of these characteristics.

Compare to previous social network studies, the scientific contribution of this research work is to analyze some specific and real social network datasets by using data mining approaches innovatively, some traditional data mining methods are no longer applicable because of the particularity of social networks datasets, the new methods such as adjacency, similarity, density, transitivity, groups and reachability are used here to analyze social notworks, constructive conclusions of social network characteristics are summarized by the research of specific social network datasets, which has important practical significance.

2 A Typical Description of Link Mining

2.1 Node Ordering and Classification Based on Link

The nodes ordering based on the link in social network analysis is a core analysis task, its purpose is to analyze the link structure of the graph, according to the measurement of the importance of some measuring nodes, to sort nodes of the graph, the importance which can be measured called centrality. According to the different complexity, it can be divided into local and global measurement metrics. The former includes "degree centrality" [5], that the degree of each vertex, the latter includes "eigenvector/ power centrality", which is based on the connection of a node to other important nodes to characterize the importance of the node.

The classification problem of traditional machine learning is based on the data instance [6] that is the assumption of independent and identically distributed. However, many practical problems do not satisfy this assumption. In a classification problem of links, a data graph G = (O; L) means the node set O and the links set L between set O, our mission is to give a class label to the members of O. Link node classification problem with the traditional classification of the most significant difference is the type of nodes is related to each other. How to design a reasonable classification algorithm can effectively use information is a challenge for researchers.

Chakrabarti [7] consider this issue in the Reuters news data set, Lafferty proposed a concept of conditional random fields, which extends the limit of traditional maximum entropy model for the structure of the data must be chain. Lu and Getoor for each data instance by adding new property to expand a simple machine learning classifier, purpose is to enable them to deal with classification problems based on the node link. New property added measure of class labels in the node distribution of the composition of the Markov blanket.

2.2 Node Clustering

Node cluster is known as group detection, its purpose is to cluster the nodes which have a common characteristic. First, assume that graph nodes and links are the same species, in this assumption; the group detection technology can be divided into polymerization clustering and abruption clustering [8-9]. The mission of block modeling in social network analysis is divided social networks into individual set; this set is called position, which shows the set of similar links in the network. A similarity measure [10] which is defined between the collection and aggregation cluster are used to find location. Spectral graph partitioning methods for determining the order for a specified number of groups and graph can be removed in the near minimum set of links to solve the group detection. Another group of recent detection method used to connect the edge of a characterization to determine the measure of the link connecting groups.

Different from above method, many groups, the method of testing is based on stochastic block modeling in social network analysis. In the randomized block model, the observed social networks are assumed to be an implementation of the random block model. We assume the position of individual in the network is to meet the independent and identically distributed random variables; a given type of link only depends on the position of two individuals which the link connected. Nowicki and Snijders [11] proposed a generalized method of random block modeling, which can deal with the relationship between weight and number of any number of locations.

2.3 Link Prediction and Subgroup

Link prediction is based on the node attribute it links and links have been observed to predict the existence of a link. Link prediction is widely used, including the prediction of social friendships between people, e-mail, telephone contact and cooperation relations. Some links which have not been observed often need us to speculate. Popescul [12] introduced a structure of logistic regression model which can use relational features to predict the existence of the link.

Since most of data sets people are interested in is sparse, so the link is more difficult to predict, one difficult thing for the link prediction to structure statistical model is: links often have low prior. Such difficult thing is not just exist in the model evaluation, but also in the quantity of prediction feasibility. A good way to improve the quality of prediction is to make a comprehensive prediction. So Markov random fields are very useful.

Subgroup is also very important. Researchers need to find subgroup which they are interested in or some frequent group in a set of graph, an effective approximation algorithm is necessary for finding subgroup in a graph. Inokuchi [13] described the AGM, which can find all subgroups meeting the minimum support. Kuramochi [14] used an adjacency graph of data to describe new optimization and also used candidate subgroups' generation process to improve the AGM. Yan [15] described gSpan; this approach can avoid the cost of the candidate subgroup generation. It was first mapped graph to depth-first search code, and a dictionary is used to order them, then doing the depth-first search on the search tree of dictionary sequence definition. Other methods came from inductive logic programming. Dehaspe [16] first used the technique of inductive logic programming in the field of toxicology to find frequent patterns.

2.4 The Models of Graph Generation

Social networks analysis study graph generation models of different kinds of graph. There are many random distribution graphs, such as: Bernoulli graph distribution, conditional uniform graph distribution, dyadic dependence graph distribution and P* models. Bernoulli graph is the simplest generation model at present. It assume that the edges between instruction nodes O_i, O_j whether exist random variables I_{ij} are independent and identically distributed. When the probability of connection existence is 0.5, this distribution is often called the uniform random graph distribution. The definition of conditions uniform graph distribution is the uniform distribution on the graphs which have special structural features. These special features are fixed links, in-degree, out-degree, and so on. In the dyadic dependence distribution, just the dyad (I_{ij}, I_{ji}) are interdependent and the dyad meet the multinomial distribution. P* model assumes that the links which have at least a common vertex are interdependence. The dependence structure of graph generation model is wider than Markov graph model, including multi-node, link type and link structure of the dynamic network model.

3 Methodologies of Social Networks Mining

3.1 Method Based on Similarity Measurement

Many data mining methods based on similarity measure. For example, k-nearest neighbor algorithm and some clustering algorithm. And given an evaluation criterion in some sort tasks, the definition of similarity is a key step in this type of algorithm. The definition of similarity associated with tasks, the best definition of similarity is likely to be different when the same data set under different tasks. Sometimes it is difficult to choose an appropriate similarity measurement, especially when there are lot of attributes whose relationships are not clear with aim and task. However, if given the appropriate similarity measure, such algorithm has a good intuitive explanation.

Similarity measure [17] is very useful in the link prediction. Link prediction is to determine whether there is link between the two actors. In the social network G, the similarity measure function for each pair of nodes: <x, y>, is given a possibility of a link: score(x, y). In some applications, the function can be seen as the topology structure of network G, for each node x and y calculated the degree of similarity between them. However, in some social network analysis tasks, the weight is not calculated the degree of similarity between nodes, but in order to do appropriate changes for specific target. Some of these weights are based on node neighborhood of node; others are based on the ensemble of all paths.

If two authors' colleagues have a large intersection, the possibility of their future cooperation will be greater than the same two who do not have same colleagues. Two people who have overlapping social circles have more probability to become friends. Starting from this intuitive observation, the probability that node x and node y contact each other in the future is related to their neighborhood nodes. $\Gamma(x)$ can be used to denote neighborhood of x in graph G. Some methods measure the intersection level of $\Gamma(x)$ and $\Gamma(y)$ as the probability of two node intersection.

Common neighbors use the number of neighbor's intersection as a measurement of intersection degree, which is a very straight idea. It defines:

$$score\ (x, y) := |\Gamma(x) \cap \Gamma(y)| . \quad \textbf{(1)}$$

Jaccard coefficient refers a similarity measure which is often used in information retrieval:

$$score\ := \left| \frac{\Gamma(x) \cap \Gamma(y)}{\Gamma(x) \cup \Gamma(y)} \right| . \quad \textbf{(2)}$$

These two methods are simple counts, which treats all the neighbours equally, but Adamic/Adar method takes neighbours property into consideration, which weighted the neighbours:

$$score\ (x, y) := \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} . \quad \textbf{(3)}$$

Preferential attachment is very suitable for the growth model of network, which focuses on the probability of increase an edge from node x is proportional to current number of neighbours of node x. It means the node which has more neighbours currently may form more links in the future, which is defined like:

$$score\ (x, y) := |\Gamma(x)| \cdot |\Gamma(y)| . \quad \textbf{(4)}$$

In the sort of object, according to the node characters itself, as well as local or global structure of the networks, the importance of nodes can be judged. For example, node degree can simply be used as the importance standards of local standards. Overall standards can use approach of eigenvectors to describe node importance which is related to the important nodes they linked.

3.2 Method Based on Statistics

Statistical Relational Learning [18] is a class of algorithms which combine statistical methods and data relationship. It focuses on the joint probability distribution of the data. Statistical relational learning model is established for relational data, which can be well described by the social network, thus completing the required analysis tasks.

MLN is often used in social networks. Richardson and Domingos [19] hope that using MLN can make the sum of probability and logic integrated into an expression. Because the probability model can handle uncertainty and knowledge can be succinctly expressed by logic. MLN makes Markov network and the first-order logic together.

Markov network describes a set of random variables: $X=(X_1, X_2, \ldots, X_n) \in x$. Typically, the use of log-linear model is expressed as:

$$P(X = x) = \frac{1}{Z}\exp\left(\sum_j w_j f_j(x)\right). \tag{5}$$

Z is the standardized item, $f_j(x)$ is the characteristic property of the state x.

4 Data Analysis

Key research problem of this paper is how to study specific social network datasets by data mining approaches. If this problem is solved, some useful information can be drawn here. But it is not like the traditional data mining approaches, specific social network datasets not only focus on link information but also have many variability and unpredictable nature, which makes the work more difficult than before. Now some specific datasets will be analyzed to find common characters which show universal and important features of whole social networks by using innovative data mining methods based on connectivity and statistics.

4.1 Specific Examples of Social Networks

Dataset Zachary karate club [20] will be analyzed here, the datasets are collected from the members of a university's karate club by Wayne Zachary. The ZACHE matrix represents the presence or absence of ties among the members of the club; the ZACHC matrix indicates the relative strength of the associations, which is the number of situations in and outside the club in which interactions occurred.

Zachary used these data and an information flow model of network conflict resolution to explain the split-up of this group following disputes among the members. If the data uploaded in software Ucinet [21], two matrices of nodes can be shown in Fig.1 and Fig. 2.

Matrix 1: ZACHE

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	1
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	1
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	1	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0

Fig. 1. ZACHE

Matrix 2: ZACHC

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	4	5	3	3	3	3	2	2	0	2	3	1	3	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	2	0	0
2	4	0	6	3	0	0	0	4	0	0	0	0	0	5	0	0	0	1	0	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0
3	5	6	0	3	0	0	0	4	5	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	2	0
4	3	3	3	0	0	0	0	3	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	3	0	0	0	0	0	2	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	0	0	0	5	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	3	0	0	0	2	5	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	2	4	4	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	2	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	3	4
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
11	2	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	3	5	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4
17	0	0	0	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
20	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1
22	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	4	0	3	0	0	5	4
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	3	0	0	0	2	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	0	0	0	0	0	0	7	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	2
28	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	0	0	0	0	0	0	0	0	4
29	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	4	0	0	0	0	0	4	2
31	0	2	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3
32	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	7	0	0	2	0	0	0	4	4
33	0	0	2	0	0	0	0	0	3	0	0	0	0	0	3	3	0	0	1	0	3	0	2	5	0	0	0	0	0	4	3	4	0	5
34	0	0	0	0	0	0	0	0	4	2	0	0	0	3	2	4	0	0	2	1	1	0	3	4	0	0	2	4	2	2	3	4	5	0

Fig. 2. ZACHC

There are two 34×34 matrices; the ZACHE is a symmetric and binary matrix which shows links among members in the club, the ZACHC is a symmetric and valued matrix which indicates the relative strength of the associations. The two datasets can be analyzed from aspects of link proportion, adjacency, out-degree, in-degree, similarity, clustering, distance, density, transitivity, groups and reachability.

Link Proportion can be used in Ucinet to do a basic analysis about the ZACHE matrix. Theoretically, for each actor, other actor is related to it more frequently or which other actor has more probability to link it. But the link matrix just uses 1 and 0 to represent whether the two actors are linked or not, for one column, the proportions are all same if two actors have relation. So ZACHC matrix which shows the relative strength can be used here.

Adjacent shows connection having actors in common, for example, the relation A to B and relation B to C have the actor B in common. The two relations are being “adjacent” when they have one actor to share. Use the “Linegraph” in Ucinet the actor can be transformed by actor matrix into a relation by relation matrix, and then the research are focused on relations rather than actors. Some useful information about social structure can be got by using this transformation.

An actor’s out-degree is the sum of connection from this actor to other actors, which means the actor sends information to others. The degree is very important, because it not only tells how many connections an actor has, but also shows the influence of an actor. An actor’s in-degree is the sum of connection from other actors to this actor, which shows how many actors send information to the specific one. The in-degree is very meaningful, because actors that receive much information from sources seem to be powerful and prestigious. But sometimes it maybe means information overload or noise interference.

Reachability means there are connections which can be traced from the source to the specific actor. Some actors cannot reach others in the network, which indicate that the network has more than one subgroup. Ucinet can be used to analyze the reachability of ZACHE data. For each pair of nodes, the algorithm finds whether there exists a path of any length that connects them. In the Zachary karate club data, one actor can reach any other actors. The reachability of this network is perfect. So the reachability can discover whether there is an indirect or direct connection from one actor to another.

Now think about another question. If there are two person A and B, each of them have 10 friends, each of A’s friends have no friend except A, but each of B’s friends all have other 10 friends. So whose influence is greater? Obviously the answer is B, and also B’s potential of influence is much greater than A. To explain it, distance can be to measure the adjacency in a social graph. If two actors are adjacent, the distance between them is 1. If A connects to B, B connects to C, then distance between A and C is 2. The distance is important to understand actors’ different in the constraint and the probability that they have their position. Furthermore, people are also interested in how many pathways exist between two actors. Sometimes there is more than one way between two actors. If people need to send messages from actor A to actor B, they should choose the shortest distance between A and B. Because it is faster and less dependence trouble.

Ucinet can be used to analyse the ZACHE data, algorithm will find the shortest path between nodes, which is called quantity geodesic distance. The average distance of all the pair of nodes is 2.408, the distance-based cohesion which can show the compactness of the network is 0.492, and the range of the cohesion is from 0 to 1, the greater value it is, the greater cohesiveness social network is. So geodesic distances are generally small, which suggests information can transmit quite quickly in the network. The network is fully connected, because there is a pathway from each actor to each other actor. So information that starts anywhere will eventually reach every node.

Sometimes, there are many efficient pathways between two actors; the connection redundancy is a feature of networks. So the number of geodesic paths should be calculated. There are usually some multiply geodesic paths exist in the graph.

If compare two populations, actors in one population have fewer ties, but actors in another population have many ties, which means that the social life is quite different in these two populations. So density measure gives us an index of degree of dyadic connection in one population. Attributes (different clubs) have been used to divide the network into two subgroups before. Ucinet is a useful tool to calculate the density of Zachary karate club social network. The network is divided into two subgroups; the connection between subgroup actors is shown clearly in the table. The density value is the proportion of connection between subgroups. The density of 1 0 block is 0.1765, which shows there are 54 connection actually exist in the possible 18×18-18 connections, here we should ignore the diagonal. The 1 0 subgroup have same in-ties and out-ties, and it is not quite dense. The 2 1 subgroup have more out-ties than in-ties, and it is not dense either. The 1 0 subgroup’s ties to 2 1 and 2 1 subgroup’s ties to 1 0 are same.

Some researchers suggest that social relations can be observed in triads. So people should be interested in the proportion of transitivity in triads. The transitivity is a kind of balance, where, if actor A has a tie to actor B, B has a tie to C, and then A also has a tie to C. Some researchers think the transitivity can show the equilibrium of a network. Ucinet can be used to find the transitivity of ZACHE data as below (Fig. 3).

```
Relation: ZACHE
---------------

Number of non-vacuous transitive ordered triples: 270
Number of triples of all kinds: 35904
Number of triples in which i-->j and j-->k: 1056
Number of triangles with at least 2 legs: 2628
Number of triangles with at least 3 legs: 270

Percentage of all ordered triples: 0.75%
Transitivity: % of ordered triples in which i-->j and j-->k that are transitive: 25.57%
Transitivity: % of triangles with at least 2 legs that have 3 legs: 10.27%
```

Fig. 3. Transitivity

There are 270 cases where if AB and BC are exist, AC is exist too. And it has 35904 triples in all, which shows that 0.75% of the triples are transitive. So the network is not equilibrium. Moreover, there are 1056 triples like: AB, BC, anything. The proportion of transitivity in these triple is 25.57%, which means there are 25.57% triple could be transitive.

Division of actors into groups is a very important aspect of social structure. It can make us understand how the network as a whole is likely to behave. Some groups from a social network overlap each other; the information may spread rapidly in these groups. But if the groups don't overlap, one group's traits will not diffuse to others. When two actors have a connection, a group is formed. The group structure begins with such a basic group, and tries to find how far this kind of relationship can be extended. So clique is introduced here. A clique means to extend the dyad by adding its members who are connected to all of the members in the group. A clique which is a kind of group is a subgroup of a network; the actors in the clique are more closely and intensely connected to one another than they are to other cliques. Formally, a clique is the maximum numbers of actors which have all possible ties exist among themselves. People are interested in where the subgroups overlap, and which actors are most central and isolated from cliques. The answer is shown in Fig. 4.

```
HIERARCHICAL CLUSTERING OF OVERLAP MATRIX

           1 1   1     1 1           1 1 2 2 2 2 2 3 2 2 2 2 1 1 1   3 3 2 3 3
Level      0 2 5 1 6 7 7 3 1 2 3 4 8 4 8 0 2 9 5 6 2 8 7 3 1 9 6 5 9 1 0 4 3 4
-----      - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
8.000      . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . XXX
5.000      . . . . . . . . XXX . . . . . . . . . . . . . . . . . . . . . . XXX
2.333      . . . . . . . . XXXXX . . . . . . . . . . . . . . . . . . . . . XXX
2.133      . . . . . . . . XXXXXXX . . . . . . . . . . . . . . . . . . . . XXX
2.000      . . . . XXX . . XXXXXXX . . . . . . . . . . . . . . . . . . . . XXX
1.667      . . . . XXX . . XXXXXXX . . . . . . . . . . . . . . . . . . . XXXXX
1.500      . . . . XXX . . XXXXXXX . . . . . . . . . . . . . . . . . . XXXXXXX
1.000      . . XXX XXXXX . XXXXXXXXX . . . . . XXXXX . . . . . . . XXX XXXXXXX
0.667      . . XXX XXXXX . XXXXXXXXX . . . . . XXXXX . . . . . . . XXXXXXXXXXX
0.600      . . XXX XXXXX . XXXXXXXXX . . . . XXXXXXX . . . . . . . XXXXXXXXXXX
0.450      . . XXX XXXXX . XXXXXXXXX . . . . XXXXXXX . . . . . . XXXXXXXXXXXXX
0.444      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . . . . . . XXXXXXXXXXXXX
0.394      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . . . . . XXXXXXXXXXXXXXX
0.350      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . . . . XXXXXXXXXXXXXXXXX
0.315      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . . . XXXXXXXXXXXXXXXXXXX
0.286      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . . XXXXXXXXXXXXXXXXXXXXX
0.263      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX . XXXXXXXXXXXXXXXXXXXXXXX
0.242      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXX
0.210      . . XXX XXXXX . XXXXXXXXXXX . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.178      . . XXXXXXXXX . XXXXXXXXXXX . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.127      . . XXXXXXXXX XXXXXXXXXXXXX . . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.030      . . XXXXXXXXX XXXXXXXXXXXXXXX . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.014      . . XXXXXXXXX XXXXXXXXXXXXXXXXX . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.007      . . XXXXXXXXX XXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.001      . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.000      . . XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
0.000      XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Fig. 4. Actor-by-actor

It shows hierarchical clustering of clique overlap. It is shows clearly that actor 33 connect to actor 34 in 8 different cliques. They are closest. Then it turns to actor 1 and actor 2, actor 6 and actor 7 and so on. The ucinet also generate a tree diagram below, which shows the clustering more intuitive (Fig. 5).

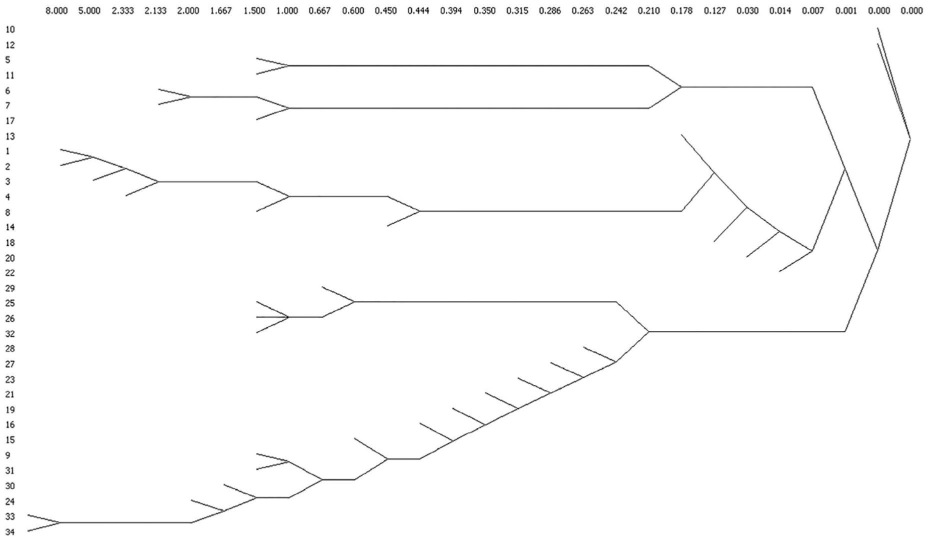

Fig. 5. Tree diagram

Clustering analysis can be used to identify patterns, which can be used in actor-by-actor similarity or distance matrix. The similarity of actors is an important concept, which can be a criterion of clustering. For the binary data, a common way of finding similar actors is to look at the vectors of actors' tie, and make the entries in one match the entries in another, see how closely they are. If ZAHCE data are measured, the figure is got as below (part) (Fig. 6).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.000	0.438	0.263	0.333	0.133	0.125	0.125	0.200	0.056	0.059	0.133	0.000	0.067	0.188	0.000	0.000	0.125	0.067	0.000	0.063
2	0.438	1.000	0.308	0.444	0.091	0.083	0.083	0.375	0.273	0.100	0.091	0.111	0.222	0.333	0.000	0.000	0.000	0.125	0.000	0.111
3	0.263	0.308	1.000	0.400	0.083	0.077	0.077	0.333	0.182	0.000	0.083	0.100	0.200	0.300	0.091	0.091	0.000	0.200	0.091	0.182
4	0.333	0.444	0.400	1.000	0.125	0.111	0.111	0.600	0.222	0.143	0.125	0.167	0.200	0.500	0.000	0.000	0.000	0.333	0.000	0.286
5	0.133	0.091	0.083	0.125	1.000	0.750	0.250	0.167	0.143	0.000	0.333	0.333	0.250	0.143	0.000	0.000	0.250	0.250	0.000	0.200
6	0.125	0.083	0.077	0.111	0.750	1.000	0.500	0.143	0.125	0.000	0.250	0.250	0.200	0.125	0.000	0.000	0.333	0.200	0.000	0.167
7	0.125	0.083	0.077	0.111	0.250	0.500	1.000	0.143	0.125	0.000	0.750	0.250	0.200	0.125	0.000	0.000	0.333	0.200	0.000	0.167
8	0.200	0.375	0.333	0.600	0.167	0.143	0.143	1.000	0.286	0.200	0.167	0.250	0.500	0.800	0.000	0.000	0.000	0.500	0.000	0.400
9	0.056	0.273	0.182	0.222	0.143	0.125	0.125	0.286	1.000	0.400	0.143	0.200	0.167	0.429	0.400	0.400	0.000	0.167	0.400	0.333
10	0.059	0.100	0.000	0.143	0.000	0.000	0.000	0.200	0.400	1.000	0.000	0.000	0.000	0.400	0.333	0.333	0.000	0.000	0.333	0.250
11	0.133	0.091	0.083	0.125	0.333	0.250	0.750	0.167	0.143	0.000	1.000	0.333	0.250	0.143	0.000	0.000	0.250	0.250	0.000	0.200
12	0.000	0.111	0.100	0.167	0.333	0.250	0.250	0.250	0.200	0.000	0.333	1.000	0.500	0.200	0.000	0.000	0.000	0.500	0.000	0.333
13	0.067	0.222	0.200	0.200	0.250	0.200	0.200	0.500	0.167	0.000	0.250	0.500	1.000	0.400	0.000	0.000	0.000	0.333	0.000	0.250
14	0.188	0.333	0.300	0.500	0.143	0.125	0.125	0.800	0.429	0.400	0.143	0.200	0.400	1.000	0.167	0.167	0.000	0.400	0.167	0.600
15	0.000	0.000	0.091	0.000	0.000	0.000	0.000	0.000	0.400	0.333	0.000	0.000	0.000	0.167	1.000	1.000	0.000	0.000	1.000	0.250
16	0.000	0.000	0.091	0.000	0.000	0.000	0.000	0.000	0.400	0.333	0.000	0.000	0.000	0.167	1.000	1.000	0.000	0.000	1.000	0.250
17	0.125	0.000	0.000	0.000	0.250	0.333	0.333	0.000	0.000	0.000	0.250	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
18	0.067	0.125	0.200	0.333	0.250	0.200	0.200	0.500	0.167	0.000	0.250	0.500	0.333	0.400	0.000	0.000	0.000	1.000	0.000	0.667
19	0.000	0.000	0.091	0.000	0.000	0.000	0.000	0.000	0.400	0.333	0.000	0.000	0.000	0.167	1.000	1.000	0.000	0.000	1.000	0.250
20	0.063	0.111	0.182	0.286	0.200	0.167	0.167	0.400	0.333	0.250	0.200	0.333	0.250	0.600	0.250	0.250	0.000	0.667	0.250	1.000
21	0.000	0.000	0.091	0.000	0.000	0.000	0.000	0.000	0.400	0.333	0.000	0.000	0.000	0.167	1.000	1.000	0.000	0.000	1.000	0.250
22	0.067	0.125	0.200	0.333	0.250	0.200	0.200	0.500	0.167	0.000	0.250	0.500	0.333	0.400	0.000	0.000	0.000	1.000	0.000	0.667
23	0.000	0.000	0.091	0.000	0.000	0.000	0.000	0.000	0.400	0.333	0.000	0.000	0.000	0.167	1.000	1.000	0.000	0.000	1.000	0.250
24	0.000	0.000	0.154	0.000	0.000	0.000	0.000	0.000	0.250	0.167	0.000	0.000	0.000	0.111	0.400	0.400	0.000	0.000	0.400	0.143
25	0.056	0.000	0.083	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
26	0.056	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
27	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.167	0.333	0.000	0.000	0.000	0.167	0.333	0.333	0.000	0.000	0.333	0.250
28	0.053	0.083	0.000	0.111	0.000	0.000	0.000	0.143	0.286	0.500	0.000	0.000	0.000	0.286	0.200	0.200	0.000	0.000	0.200	0.167
29	0.118	0.091	0.000	0.125	0.000	0.000	0.000	0.167	0.333	0.667	0.000	0.000	0.000	0.333	0.250	0.250	0.000	0.000	0.250	0.200
30	0.000	0.000	0.077	0.000	0.000	0.000	0.000	0.000	0.286	0.200	0.000	0.000	0.000	0.125	0.500	0.500	0.000	0.000	0.500	0.167
31	0.111	0.000	0.273	0.111	0.000	0.000	0.000	0.143	0.400	0.200	0.000	0.000	0.000	0.286	0.500	0.500	0.000	0.200	0.500	0.400
32	0.000	0.071	0.231	0.091	0.125	0.111	0.111	0.111	0.375	0.143	0.125	0.167	0.143	0.222	0.333	0.333	0.000	0.143	0.333	0.286
33	0.120	0.105	0.053	0.059	0.000	0.000	0.000	0.067	0.250	0.167	0.000	0.000	0.000	0.133	0.091	0.091	0.000	0.000	0.091	0.071
34	0.138	0.130	0.286	0.045	0.000	0.000	0.000	0.000	0.111	0.000	0.000	0.000	0.000	0.000	0.063	0.063	0.000	0.000	0.063	0.000

Fig. 6. Similarity

It means, for example, actor 1 to actor 2 is 0.438, which shows that compare actor 1 and actor 2, they have same tie to other actors 43.8% of the time. So the 43.8% can represent the similarity between actor 1 and actor 2. Hierarchical clustering of nodes is based on the similarity of actors' ties. The two most similarity cases are combined into a class. Then the similarity of new class to others is then computed. The join process can be repeated until all the cases are agglomerated into a single clustering. Now with the ucinet the clustering result can be got like Fig. 7.

```
HIERARCHICAL CLUSTERING

          1 2 1 2   1 2     3 2   1 2 3 1       1   2 2 3 1 1 2 2 2 1 1     3 3
Level   1 0 8 6 3 5 5 1 9 2 0 6 3 2 4 1 9 6 7 1 4 2 7 2 8 7 5 0 9 3 4 8 3 4
------  - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
0.0000  XXXXXXXXX XXXXXXXXXXXXX XXXXXXXXXXX XXXXXXXXXXXXX XXXXXXXXXXXXXXX .
0.0714  XXXXXXXXX XXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX .
0.0857  XXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX .
0.1442  XXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX .
0.1587  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX .
0.5152  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

Fig. 7. Clustering

4.2 Some Features of Zachary Karate Club Dataset

From the analysis above, some features can be found. First, the actor 34 which has most important ties to 8 other actors is the most important actor. Then are actor 3 and actor 33. Second, each entry's maximum degree is 1, which means no relation is the source of multiple relations. Third, actor 34 sends ties to 17 other actors. So the actor 34 is the most influential actor in this network, just like the first conclusion. Actor 1, actor 3, actor 33 are all influential. Actor 12 just sends one tie, so it is the least important one as a source of information. And the actors which have very low out-degree or very high out-degree have less variability than those which is in medium levels of out-degree. The fact shows that the actors which send information to almost everyone or no-one are more predictable and constrained of their behavior. Fourth, actor 34 which receives 17 ties is the most powerful one. And the actor 12 is almost isolated. Fifth, one actor can reach any other actors in this network; the information can be spread fully and completely. Sixth, the average distance of all the pair of nodes is 2.408; the distance-based cohesion is 0.492, which means the compactness of the network is in a middle class. And the largest geodesic distance is 5; others are 1, 2, 3, which suggest that information can transmit quite quickly in the network. Moreover, all the actors have flows from themselves to others, which suggest that there is a great chance that communication happens between two actors; so the connection of this network is quite strong in general. However, actor 12 only has one way to gain information, so 12 is the most vulnerable actor in the network. Seventh, network can be divided into two subgroups. The 1 0 group contains actor 1, 19, 28, 21, 30, 23, 24, 25, 26, 10, 27, 29, 17, 31, 15, 16, 33, 34, which have same in-ties and out-ties and is not quite dense. The 2 1 subgroup contains actor 3, 20, 5, 22, 6, 7, 4, 18, 2, 11, 12, 13, 14, 32, 8, 9, which have more out-ties than in-ties and is not dense either. The 1 0 subgroup's ties to 2 1 and 2 1 subgroup's ties to 1 0 are same. Eighth, 0.75% of the triples are transitive. So the network is not quite equilibrium. And there are 25.57% triple could be transitive. Ninth, 25 cliques exist in the network. There are also overlaps in these 25 cliques. The actor 17 is special here, because it is even not adjacent to any clique. Actor 33 and actor 34 are both members in 8 cliques of 25 cliques, so they are closest under the concept of cliques. Cliques 12 and 13 share 4 actors here, which have the most actors in common. Tenth, actor 1, 10, 16, 28, 23; actor 5, 15, 21, 9, 2, 30, 26; actor 3, 12, 24, 31, 19, 6; actor 7, 11, 4, 22, 27, 32, 18; actor 17, 25, 20, 29, 13, 14, 8, 33 are most similar, they get five new cluster A, B, C, D, E. Then C and D form a new cluster F, then A and B form a new cluster G, then F and E form a new cluster H, then G and H form a new cluster I, last, I and actor 34 form a new cluster J which contain all 34 actors.

5 Conclusions

Zachary karate club dataset has been analyzed from some different aspects above, like link proportion, adjacency, degree, reachability and so on. Actually, the result can tell some important features of social network. However, social network often refers to a very large dataset; the Zachary karate club dataset is small. So in order to make the analysis more effective and accurate, some other datasets which are different to Zachary club can be used to do the similar analysis, and then compare the results with the karate club one; some credible conclusions of whole social networks can be drawn.

First, basically, the social network data often focus on the link, but not the attributes of actors. The relation ties between actors are very important. Relation ties are often interactive and asymmetric; it is different in content and intensity. Relations link directly or indirectly connected to the network members.

Second, most of relations are often not sources of other relations; each entry's maximum degree is often 1, which means no relation is the source of multiple relations.

Third, the actors which have very low out-degree or very high out-degree have less variability than those which is in medium levels of out-degree. The fact shows that the actors which send information to almost everyone or no-one are more predictable and constrained of their behavior.

Fourth, one actor often can reach almost any other actors in the network; so the information may be spread fully and completely.

Fifth, information often can be transmitted quite quickly in the network and the compactness of the network is often in a middle class.

Sixth, there are often some multiply geodesic paths exist in this network, which suggest that the information flow in the network will not be broken down and each actor is difficult to be a broker, because most of actors have alternative efficient pathways to link other actors.

Seventh, actors often have flows from themselves to others, which suggest that there is often a big chance that communication happens between two actors. However, there are often several actors that just have a few ways to gain information exist, but the whole networks' communication will not be influenced.

Eighth, the network is often not quite equilibrium. And there are triple which could be transitive exist.

Ninth, there are clique and subgroup exist in the social network. And the cliques are often overlaps. Furthermore, all the networks can be clustered.

Tenth, social ties' structure can produce non-random networks, which can generate network clusters, network boundaries. Asymmetrical ties and complex networks make the distribution of scarce resources unequal.

Eleventh, there is a prediction that network can produce the action of cooperation and competition for the purpose of scarce resources.

Twelfth, several subgroups can form a big social network. From the network-subgroup-actor concept, a guess of social structure can be drawn: as individuals people may different and changing, the ties between actors may also change locally, which can make the subgroups minor changes. However, the whole network will be stable. Because the social network relations as a whole are the result of long accumulation. When it local changes, the social network will return to a previous state after a period of time.

In brief, Application of social network analysis is a very strong sociological research approach; it can be used to describe the nature of the community, and can solve many practical problems. Social network data collection, both to the field with an unprecedented opportunity for data analysis also presented enormous challenges. Data mining as a tool help people to discover useful knowledge from a lot of data. Through constant development, it has been able to deal with the structure of social networks as network data. Different from the traditional tasks of data mining, which assume the instances are independent; instances in social network are dependent. Such dependence can be described as links. Mining from links can provide us more accurate and richer information about the social network.

However, there are still limitations of the work. Although some datasets which are different to Zachary club dataset are used here to do analysis and some conclusions are drawn, the total data quantity for research is still not big enough. It may lead to the analysis results are not very accurate. But the massive social network data is really difficult to be uploaded by existing analysis software or computer compiling system because of the huge data volume, especially the huge link information. So the important aspects of mining social network in future are to improve the software and enhance the performance of the data warehouse system, which can make them have a better effect on deal with massive data, in addition, some new social network data mining algorithms and more reasonable distributed processing methods should be created to analyze the data effectively. The approaches and technology of social networks study will be increasingly advanced in future.

Acknowledgements

This work is supported by the scientific research foundation supporting project of postal development research Center, the project name is data analysis and mining research based on of postal rural e-commerce, the project No. is PY2016033.

In addition, the work is also supported by the financial project coming from federation of social sciences of Hebei Province, the project name is study on the internet communication strategy of guiding the social public opinion under media convergence, the project No. is 201601733.

The contents and conclusions of this work are based on deep study of sub-topic of the two research projects above.

References

[1] C. Zhang, Y.F. Jin, W. Jin, Study of data mining algorithm in social network analysis, in: Proc. the 2015 3rd International Conference on Mechatronics, Robotics and Automation, Shenzhen, China, 2015.

[2] Y.F. Jin, G. Li, C. Zhang, Study of data mining in social network analysis, Journal of Information & Computational Science 16(11)(2015) 5947-5955.

[3] H. Yu and C. Zhang, Discussion on the development model of rural e-commerce in Hebei, Journal of Business Economic Research 12 (06) 2016, 172-174.

[4] B. Gu, Data Warehouse and Data Mining, vols.1, Beijing University of Posts and Telecommunications Press, Beijing, 2014.

[5] C. Zhang, Mobile E-Commerce, vols.1, Beijing University of Posts and Telecommunications Press, Beijing, 2016.

[6] B. Taskar, P. Abbeel, D. Koller, Discriminative probabilistic models for relational data, in: Proc. 18th Conf. on Uncertainty in Artificial Intelligence, Edmonton, Canada, 2003.

[7] S. Chakrabarti, B. Dom, P. Indyk, Enhanced hypertext categorization using hyperlinks, in: Proc. Int. Conf. on Management of Data, San Francisco, 1998.

[8] J. Neville, D. Jensen, Collective classification with relational dependency networks, in: Proc. 2nd Int. Workshop on Multi-Relational Data Mining, 2003.

[9] M.E.J. Newman, Detecting community structure in networks, European Physical Journal 38(5)(2004) 321-330.

[10] P.J. Carrington, J. Scott, S. Wasserman, Models and Methods in Social Network Analysis, vols.1, Cambridge University Press, Cambridge, 2005.

[11] K. Nowicki, T.A.B. Snijders, Estimation and prediction for stochastic block structures, Journal of the American Statistical Association 96(455)(2001) 1077-1087.

[12] A. Popescul, L.H. Ungar, Structural logistic regression for link analysis, in: Proc. 2nd Int. Workshop on Multi-Relational Data Mining, 2003.

[13] A. Inokuchi, T. Washi, H. Motoda, An Aprioribased algorithm for mining frequent substructures from graph data, in: Proc. Conf. on Principles and Practice of Knowledge Discovery and Data Mining, 2000.

[14] M. Kuramochi, G. Karypis, Frequent subgraph discovery, in: Proc. IEEE Int. Conf. on Data Mining, 2001.

[15] X. Yan. J. Han, GSpan: Graph-based substructure pattern mining, in: Proc. Int. Conf. on Data Mining, 2014.

[16] L. Dehaspe, H. Toivonen, R. King, Finding frequent substructures in chemical compounds, in: Proc. Int. Conf. on Knowledge Discovery and Data Mining, 1997.

[17] P. Bonacich, Power and centrality: a family of measures, American Journal of Sociology 92(5)(1987) 1170-1182.

[18] N. Lavrac, S. Dzeroski, Inductive Logic Programming, vols.1, Harwood Press, New York, 1994.

[19] M. Richardson, P. Domingos, Markov logic networks, Machine Learning 62(1)(2006) 107-136.

[20] W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33(6) (1977) 452-473.

[21] UCINET IV Datasets. <http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet/UciData.htm#zachary>, 2010 (accessed 05.03.16).