

Finding Friend Groups in Blogosphere

Shih-Ta Kuan

Dept. of Electrical Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan, ROC
kuster.die9841@gmail.com

Bang-Ye Wu

Dept. of Computer Science and Information Engineering
Shu-Te University
Kaohsiung, Taiwan, ROC
bangye@mail.stu.edu.tw

Wan-Jui Lee

Information and Communication Theory Group
TU Delft, The Netherlands
w.j.lee@tudelft.nl

Abstract

In this work, we propose an algorithm based on transitive extension for finding friend groups in blogosphere to perform social network analysis. In today's blog service environment, the establishments of friend relationships are always unidirectional, and the recognition of both ends is not necessary. Therefore, general methods will find either too small or too few cliques from friend groups. This is because the bidirectional link is built incompletely in the social network under such circumstances. To solve this problem, we propose the 1.5-clique extension method to derive better social network structures for finding friend groups, and we use Wretch (www.wretch.cc) as a case for analysis. We further make a comparison among the results of finding groups in the original social network, and its 1-clique extension, 1.5-clique extension, and 2-clique extension. The experimental results suggest that our proposed method is effective and promising.

Keywords : friend groups, blogosphere, transitive extension, social network analysis.

1 Introduction

Due to the rapid growth of Internet in the last decade, Internet service providers (ISP) nowadays need to provide various social network services (SNS), e.g. E-mail, Internet hard disk, personal web space, and online photo album, to survive in the competitive environment. The increasing main stream of SNS at the present time is *blog* which is the portmanteau of web log. Blogs provide simple interfaces and therefore users can easily maintain or add some contents on a particular subject [1, 8]. Personal web space

has been gradually replaced by blogs, and the Internet ecology is also changing. People are no longer just consumers of the websites, but also the producers. In such an environment, it is not difficult to find people with same interests and make netizens through SNS. Moreover, there are usually some records in the process of people-to-people exchanges, such as friend lists, good website lists, etc. These enormous user-records have gathered into some serious social networks, and these networks are very important in the study of information technology [4] and social science. Unfortunately, because of the rapid expansion and late maturity of the social networks, this field of study hasn't drew much attention of the researchers, especially on clustering and classification issues. In this work, we develop a method to discover friend groups from the social networks. Our method is a pre-processing algorithm based on n -clique [5] extension for social networks. After the pre-processing of social network, we can directly finding the cliques and these cliques can form into some representative friend groups.

The rest of the paper is organized as follows. In Section 2, some background and related works about social network analysis are recapitulated. In section 3, our method for discovering friend groups in the social networks is presented. Simulation results are discussed in Section 4. Finally, conclusions are summarized in Section 5.

2 Background and Related Work

Social network analysis (SNA), is a new filed of research in recent years. It is related to graph analysis, and is also one branch of data mining [4]. Also, SNA attracts some social science researchers because it is essentially related to social science. In the following, we will introduce

some related research first, and then explain the case Wretch (www.wretch.cc) that we study.

2.1 Social Network Analysis

To discover knowledge in SNA, there is an assumption that if some people talk about the same things, they are interested in the same things, and therefore these people are considered to be latent friends. Some text mining methods are used to find latent friends from blogosphere [7, 10]. First, the articles which are published in blogs are collected, and then they are clustered with Latent Semantic Indexing (LSI) [7, 10] to form some features, and people who have the same features will be considered to be friends. Through keyword feature screening, the characteristics of blogs can be found. However, some of these characteristics are dangerous, for example, racial discrimination, and hatred towards some groups or countries. It is not easy to tell whether a person's risk level is high only depends on the features retrieved from his blog. But some researchers believe that a blogger is with a higher level of danger if his blog is highly linked to the other blogs with a lot of dangerous characteristics in one way or another. This person is very likely a member of some criminal organization and the blog is their contact center [7, 3].

From a statistical point of view, social networks can be described as directed graphes for there is an actor (node) who knows (directional edge) others. In a directed graph, transitions usually exist. Suppose there are $A \rightarrow B$ and $B \rightarrow C$, then $A \rightarrow C$ may exist because of the transition. Also, A and C are supposed to be not far away from each other and there is a certain degree of relationship between them [6]. Therefore, to find a friend group in SNA can be referred to finding cliques in the graph. The idea of a clique is relatively simple. At the most general level, a clique is a sub-set of network in which the actors are more closely and intensely tied to one another than they are tied to the other members of the network [9]. Clique is a complete graph, i.e., each node is connected with all the other nodes [5]. A n -clique tries to relax the definition of clique. Let $d(u, v)$ denote the distance from nodes u to v , i.e., the number of arcs of the shortest path from u to v . Note that $d(u, v)$ in general differs from $d(v, u)$ in a directed graph. For an integer $n \geq 1$, a n -clique is a graph (or a substructure in a network) such that $d(u, v) \leq n$ for each ordered pair of nodes u and v , i.e., the distance from each node to any other is at most n . In this work, we work on finding n -cliques to further discover friend groups.

2.2 Wretch: A Case Study

Wretch [2] is the largest web blog in Taiwan. It has more than 3.92 million distinct internet albums users and more

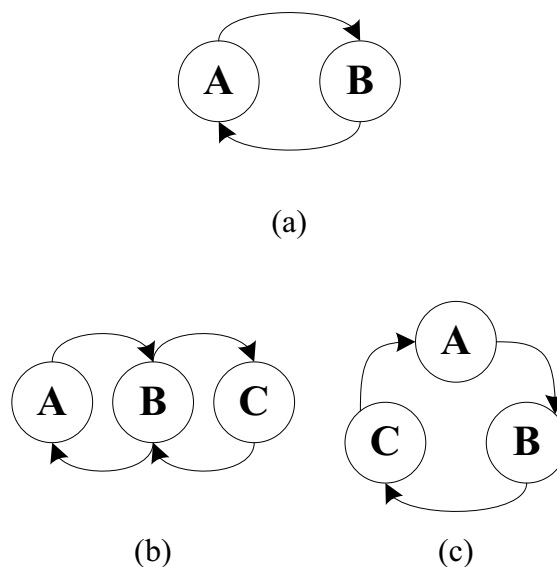


Figure 1. Unidirectional graph of (a) 1-clique, (b) 2-clique, and (c) 1.5-clique.

than 2.5 million distinct blog users. To test the number of blog users on Wretch, we chose a starting user, and went through this blogger's friend lists by breadth first search (BFS). Users can add their friends to their friend lists very easily because Wretch offers a simple user interface to operate friend lists. The linkage of adding friends in Wretch is unidirectional because it does not need the recognition of the other party. Therefore, the range of topology in this social network can reach 90% of users by linking friend lists.

3 Our Method

In the previous research, most researchers considered blog as text documents, and used the text mining technologies to cluster or classify blogs. Then, the relationship among bloggers are weighted by the text mining-based classification or clustering results. However, if the relationship among a certain group of people is highly weighted, we can only say that these people share the same characteristics or the same hobbies. It is very likely that some people in this group don't know each other and therefore it is not very appropriate to assort these people into the same friend group. On the other hand, in statistics, if $A \rightarrow B$ and $B \rightarrow C$ hold, it may be inferred that the A and C are with a considerable relationship. But such an inference on the friend lists of the Wretch blog system is not quite appropriate. After all, even if I know my friend, I may still be considered as a stranger to my friend's friend. In the unidirectional link of Wretch,

if two people know each other, there will be a relationship between them as Figure 1(a) where A and B are mutually connected to each other and therefore their relationship is a 1-clique. From the transitive extension perspective, suppose A knows B and B knows C, we say A already recognizes C and the relationship among them can be shown as Figure 1(b) which is a 2-clique graph. Suppose B is the friend of A and C, and there are bidirectional links between B and A, and B and C. Then, there will emerge a bidirectional link between A and C because they are in a 2-clique through B. As a result, all bidirectional link friends of B will also copy the friend lists of B, and the social network will be completely deformed and the relationships will be in chaos. In addition, many pop stars choose Wretch to publish their blogs and it might cause the star effect: a small number of users will be added as friends by a huge number of users. In such a case, it is very likely to find very large 2-cliques.

Intuitively, to find friend groups by n -clique, the more incomplete the data is, the more larger n we should use. But, the friend relationship data we obtained in Wretch is in fact not so incomplete so that even $n = 2$ is a too relaxed condition. One evidence of the above claim is the reciprocity of the relation ties. There are about up to 50% bidirectional ties in our data set.

Since the 1-clique is too restricted but the 2-clique is too sparse, we try to define a middle ground: the 1.5-clique. Formally, a 1.5-clique is a graph/subgraph such that $d(u, v) + d(v, u) \leq 3$ for any nodes u and v . In other words, two nodes in a 1.5-clique must be connected by either a bidirectional edge or a directed cycle of length 3. Usually, in real life, if A knows B, B knows C, and C knows A, we instinctively think that the three people know each other as shown in Figure 1(c). Especially in the cyberspace, the social network is usually incomplete. It is not very likely that each user would list everybody he knows, because even the user himself is not sure whether all his friends having blogs in Wretch. Another reason for the incompleteness is the limit of the number of friends in a blog, which will be described later in this paper. So we assume that if A, B, and C have the relationship in Figure 1(c), then we think A, B, and C should be in a group. We define this new kind of clique which is between 1-clique and 2-clique as the 1.5-clique.

It is very time-consuming if we want to search the n -clique groups in the social network which is built by the friend lists of Wretch. So we propose a pre-processing algorithm based on n -clique to extend the original network. We transform the unidirectional link graphs into bidirectional link graphs, and then find simple cliques in the bidirectional link graphs to simplify the problem. For each node, we proceed the following algorithm to extend the original network to (1) no extension bidirectional link network (BL1), (2) 1.5-clique extension bidirectional link network (BL1.5), and (3) 2-clique bidirectional link extension network (BL2):

Input: a set of nodes N_i , a set of unidirectional links UL
Output: BL1, BL1.5, BL2

Step1 I_1 = a set of nodes which are connected by N_i .

Step2 I_2 = a set of nodes which are connected by I_1 .

Step3 O_1 = a set of nodes which are connected to N_i .

Step4 O_2 = a set of nodes which are connected to O_1 .

Step5 $BL1 = I_1 \cap O_1$.

Step6 $BL1.5 = (I_1 \cap O_1) \cup (I_2 \cap O_1) \cup (I_1 \cap O_2)$.

Step7 $BL2 = (I_1 \cap O_1) \cup (I_2 \cap O_2) \cup (I_2 \cap O_1) \cup (I_1 \cap O_2)$.

The density of a 1.5-clique is the ratio of the number of edges in the original graph to the number of all possible edges.

4 Experimental Results

The concept of the original social network in Wretch and its transitive extensions derived with our algorithm was given in Table 1. The original network has 2,530,954 users, and 36,984,090 unidirectional links. A total of 9,444,848 bidirectional links were further composed from unidirectional links which means about 50% of the unidirectional links are relative. The maximal degree of bidirectional links is 614, and the mean degree of bidirectional links falls between 7 and 8. After the 1.5-clique extension, the number of bidirectional links increases about 150% (14,580,061 bidirectional links in total). However, the maximal degree only increases less than 10% which is rather low. It is very possible that there is a specific user who has a lot of friend lists and also takes his friends' friend lists very seriously, thereby he puts a lot of efforts in keeping the links bidirectional. But overall, the mean degree is increased by 150% and falls between 11 and 12. Therefore, the performance of extension is quite general.

After the 2-clique extension, it is very clear that the number of bidirectional links grows sharply to 1800%. The maximal degree is also increased by 1000%, and the mean degree is increased by about 1800%. The characteristics of the social network are obviously destroyed. We show the relationship between the degrees and the cumulative probability distribution in Figure 2 where x-axis indicates the degree in an exponential order, and y-axis is the probability of the nodes that are greater than the degree in the whole network. We can see the impact zone of 1.5-clique extension (middle line) is between 0.5 and 2.5. There is a small number of nodes have large degrees, and they also don't have much influence on the original network. However the

Table 1. The original social network, and its N -clique transitive extensions (TE).

	Original network	1.5-clique TE	2-clique TE
number of nodes N_i	2,530,954	2,530,954	2,530,954
number of BL	9,444,848	14,580,061	167,499,850
number of UL	36,984,090	-	-
maximal degree of BL	614	666	6122
mean degree of BL	7.46	11.52	132.36

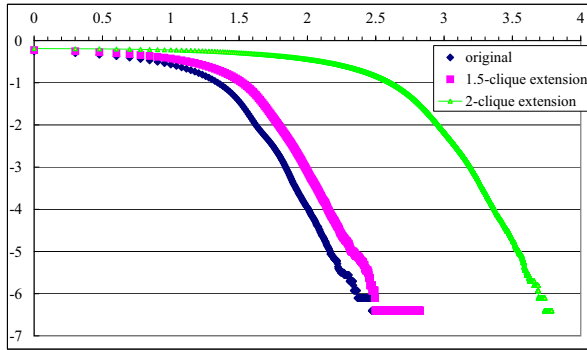


Figure 2. n -clique transitive extension.

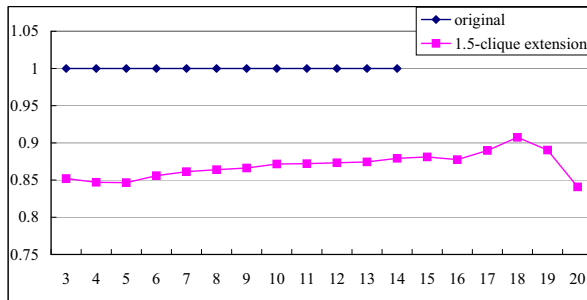


Figure 3. Sizes and the average densities of cliques in the original social network and the 1.5-clique extension of social network.

green line which represents 2-clique extension differs significantly with the original network all the way from beginning. To compute groups, we first sample 10,000 nodes which have at least one unidirectional link. Then, we find the maximal effective clique which connected to the sample nodes with the random algorithm, and filter out the group whose size is less than 3. If a group comprises only two nodes, the performance of such a group is usually insignificant. Therefore, we retain only the groups whose sizes are

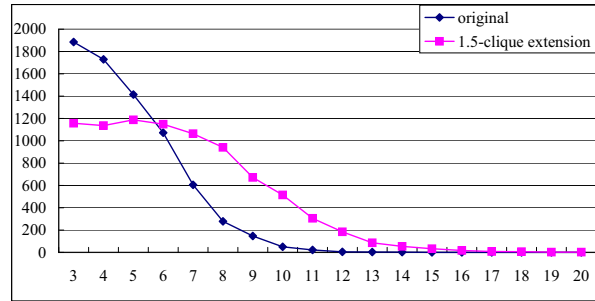


Figure 4. Sizes and numbers of groups in the original social network and the 1.5-clique extension of social network.

larger than 3. For the results of 2-clique extension have obviously destroyed the characteristics of the original social networks, we don't calculate the social network after 2-clique extension. With the same sampling nodes, 7,213 groups can be found in the original social network, while 8,519 groups are found in 1.5-clique extension of the social network. We show the sizes and the average densities of cliques in the original social network and the 1.5-clique extension of social network in Figure 3. From the figure, we can find that after the 1.5-clique extension of social network, even the sizes of groups are different, their average densities haven't differed much. With respect to the sizes, we can only identify groups with size 14 in the original social network, but in the 1.5-clique extension of social network, we can find groups of size 20. In Figure 4, we can see the differences between the original social network and its 1.5-clique extension. There are 1,885 groups of size 3 in the original social network. After 1.5-clique extension, there are 1,157 groups of size 3, 2 groups of size 20, and 68 groups with sizes larger than 14. The average size of group samples is 4.81 with the standard deviation of 1.67 in the original social network. For 1.5-clique extension, the average size of group samples is 6.52 with the standard deviation of 2.69, and there are 80% of groups with sizes smaller

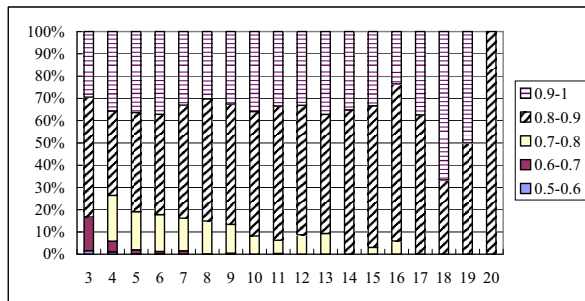


Figure 5. Density distributions after 1.5-clique extension.

than 9. Finally, we observe the density of distribution for each size of groups; the density of distribution after 1.5-clique extension is shown in Figure 5, but we don't show the result of the original network because it's density remains 1 in all cases. We can find that about 80% of the groups having densities greater than 0.8, and about 30% of the groups having densities more than 0.9 in the samples of 1.5-clique extension. In 1.5-clique extension, the worst case is with the density of 0.5. If the social network is randomly generated, the worst case may happen frequently. But we can see from the distribution of social network that it is with the characteristic of gathering groups, and only a few of groups have the densities of distribution fall between 0.5 and 0.6.

5 Conclusions

In this paper, we propose an algorithm based on transitive extension for finding friend groups in blogosphere to perform social network analysis. The simulation results suggest that our proposed method is effective and promising. We will keep working on finding clusters within groups base on this method, and hope to derive a hierarchical architecture of groups in the future. The hierarchical architecture shall be able to describe the phenomenon of groups in social network more clearly.

References

[1] Wikipedia. <http://en.wikipedia.org/wiki/Blog>.
 [2] Wretch. <http://www.wretch.cc/ad/>.
 [3] M. Chau and J. Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1):57–70, January 2007.
 [4] J. Han and M. Kamber. *Data Mining: Concepts and Techniques Second Edition*. Baker & Taylor Books, 2001.

[5] R. Hanneman and M. Riddle. Introduction to social network methods, 2005. <http://www.faculty.ucr.edu/hanneman/nettext/>.
 [6] P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, December 2002.
 [7] Q.-M. Li, M.-W. Xu, J. Hou, and F.-Y. Liu. Web classification based on latent semantic indexing. *Journal of Communication and Computer*, 3(1):24–27, 2006.
 [8] C.-C. Liu. The preliminary study of social network and self-presentation in blogosphere. Master's thesis, Department of Information Management, National Sun Yat-Sen University, 2004.
 [9] M. Morzy. Latent friend mining from blog data. In *The 1st Workshop on Internet and Network Economics*, pages 112–115, 2005.
 [10] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Latent friend mining from blog data. In *Proceedings of the Sixth International Conference on Data Mining*, pages 552–561, 2006.