A Strategy to Build Connection with a Stranger Based on Social Network Topology

Shaojun Zhang, *Department of Statistics, University of Florida*

***Abstract*—Nowadays people are connected in various social networks such as phone calls, emails and Facebook accounts. Then an interesting question arises whether we can improve our chance to make friends with some stranger in the network? Here the stranger can be our potential spouse, colleague, advisor or boss. Obviously, we can send an invitation through Facebook, send an email or make a phone call. But we are likely to be ignored since we are complete strangers. Moreover, email address and phone number are unavailable most of time. A solution is to get connected one by one along the shortest path to the target. However, the problem becomes more complicated when the shortest path is still long or when we have more than one shortest path to choose, especially in a dense network. In this work, I will develop a model mainly based on network topology including shortest paths, centrality, and community structure to address these problems. The model and the accompanying analysis will be illustrated on a dataset covering email communication of about 150 users of Enron, an American energy company.**

Project Plan

A. *Descriptive Analysis*

I will start by computing the network graph characteristics including degrees, three types of centrality as well as clustering coefficient to capture the basic topology of the Enron email communication network. Most of these results will play an important role in the following analysis. This part should be done before Nov. 9th, 2015.

B. *Community Detection*

There are many algorithms for community detection. But what I need is one that can generate overlapping communities, which usually contains more information and is very common in social networks. Label propagation will be my first choice since it is easier to implement and runs faster on large datasets. This part should be done before Nov. 16th, 2015.

C. *Model Development*

My current idea is to combine the shortest paths, community structure as well as centrality to find the optimum path based on some criterions. As we can see, the shorter the path length, the faster we can reach the stranger. The more communities covering the path that are shared by the target, the more information of the stranger we can get. Furthermore, the model should set a maximum number of people that are allowed to connect along the shortest path. This constraint is reasonable since every connection requires some form of cost such as time or money. If the maximum number is reached before connecting to the stranger, we should stop and try other methods with all the information obtained before. This part should be done before Nov. 30th, 2015.

D. *Statistical Analysis*

I will illustrate the method on the Enron email communication network dataset. A quantity of interest is the average path length between any two people in the network. It is a well-known fact that the people in the United States are separated by about six people on average, which is also known as the six degrees of separation. Therefore, I would like to see if the average distance between any two Enron email accounts agrees with that result. This is also similar to ask whether it is a small-world network. Another interesting problem is to identify whether it is more difficult to make friends with a user of more importance in the social network, namely with higher centrality score. This part should be done before the project deadline, Dec. 16th, 2015.

References

[1] B. Klimt and Y. Yang, “Introducing the Enron corpus”, in *Proc. of the First Conference on Email and Anti-Spam*, 2004.

[2] Eric D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 2009.

[3] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, “Analysis of Topological Characteristics of Huge Online Social Networking Services,” in *Proc. of WWW*, 2007.

[4] S.P. Borgatti, “Centrality and Network Flow,” *Social Networks*, vol. 27, no. 1, pp. 55-71, Jan. 2005.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, *Introduction to Algorithms*. The MIT Press, 1990.

[6] M.E.J. Newman, “Detecting Community Structure in Networks”, *The European Physical J. B*, vol. 38, pp. 321-330, 2004.

[7] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, “Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters,” *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[8] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, “Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society,” *Nature*, vol. 435, no. 7043, pp. 814-818, 2005.

[9] U. N. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.

[10] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks”, *Nature*, vol. 393, pp. 440–442, Jun. 1998.