Coauthorship among Statistical networks

Hui PAN
Department of ECE
HKUST
hpanab@connet.ust.hk

Abstract

In this report we aim to use the techniques in social network to investigate the coauthorship relationships in statistic community. To model and analyze coauthorship interactions among statisticians, we first mapped the relationship into a social graph. With this model, we could *a*) identify the most influential authors in this area; *b*) cluster the whole community into different clusters based on their coauthorship interactions; *c*) investigate the intrinsic features of statistician community. Moreover, inspired by the famous six degrees of separation concept, we found that degree of separation in statistician community is about 6.98, which means every two authors are connected by interactions in a maximum of 7 steps on average.

1 Introduction

Social network is a social structure determined by the interactions between individuals or groups. It provides a set of useful and convenient methods for analyzing the structure of whole social entities by explaining the patterns observed in these structures.

Thanks to the great effort of Ji and Jin (2), the coauthorship and citation dataset for statisticians was made available to us. The data sets are based on all published papers from 2003 to the first half of 2012 in four of the top statistical journals: Annals of Statistics (AoS), Journal of American Statistical Association (JASA), Journal of Royal Statistical Society (Series B) (JRSS-B), and Biometrika, providing a fertile ground for researches on statistics.

It is found that the interactions of coauthorship and citations in scientific community could be modeled as a social graph. The relationship of coauthorship can be mapped into an undirected graph where each author is represented with a node, and the links between any two nodes means that they cooperated to publish a paper or more.

2 Data Setup and Problem Formulation

2.1 Data Statement

The dataset provided is quite clean and well organized for the network setup. The coauthorship dataset is given in three versions: *a*) coauthorship network among the whole community, where each edge denotes at least one paper coauthored, containing 3607 authors in total; *b*) the giant component of coauthorship network, where each edge denotes at least one paper coauthored, containing 2263 authors in total; *c*) the giant component of coauthorship network, where each edge denotes at least one paper coauthored, containing 236 authors in total. Different versions are given because the complete citation network is scattered, with many isolated clusters. It is calculated that there are 369 separated components in the whole network. And the distribution of components follows the scale-free law(also called long tail distribution), i.e. fewer components have a lot of authors and most components have a few. The whole network has 3,607 nodes, while the giant component (the largest

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

connected cluster) of the network has 2264 nodes, accounting for up to 62.77% of the whole. All the other components has less than 30 nodes each. Though the largest connected cluster of the network only contain partial information of the network, it is also useful to get the main stream concept of the community. So the other two versions of network are also given.

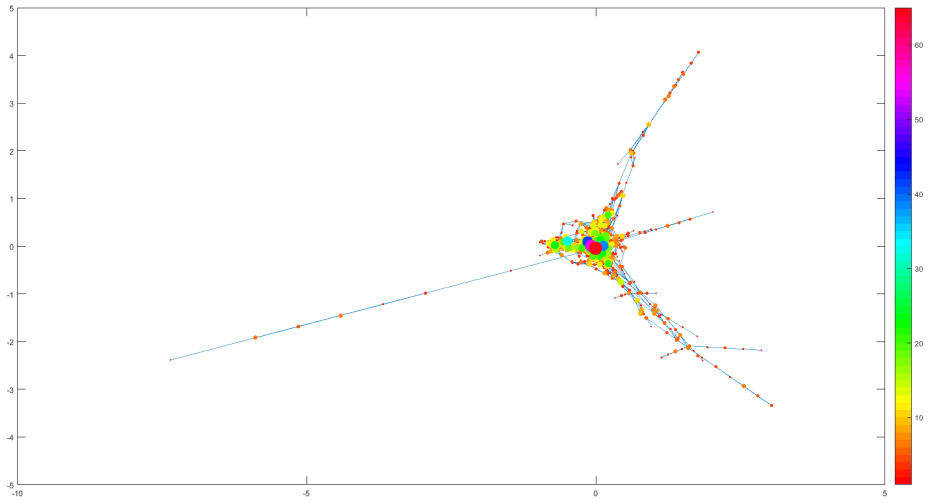

Figure 1: Giant component of Coauthorship network. The size and the color of nodes indicate the degrees of them.

2.2 Problem Formulation

We defined a graph G as an ordered pair $G = (V, E)$, where V is a set of nodes, and E is a set of links.

2.2.1 Centrality analysis

In network analysis, centrality is used to find the most important vertices within a graph. We to identify the most "influential" authors in statisticians community, There are different measurement of centrality, including degree, closeness, and betweenness.

Degree: the node with most direct links to other nodes;;

Betweenness: the node with the most shortest paths passing it;

Closeness: The node which has shortest path to all other nodes.

Degree Centrality:

$$C_D(v_i) = \sum_j A_{ij} \tag{1}$$

Betweenness Centrality:

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V} \frac{\sigma_{st}(v_i)}{\sigma_{st}} \tag{2}$$

where v_s and v_t denote the source and target nodes respectively. σ_{st} is the number of shortest paths between v_s and v_t.

Degree Centrality:

$$C_C(v_i) = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)} \tag{3}$$

where $g(v_i, v_j)$ denotes the geodesic(shortest path) distance between v_i and v_j.

Network	Degree of separation
Complete Coauthorship Network	6.9830
Giant Component of Coauthorship Network	6.9874
Giant Component w/ threshold	6.8580

Table 1: Degrees of separation of different networks

2.2.2 Degree of separation

Six degrees of separation means any two people are distanced by an average of six friendship links. It is a well-known idea that was originally set out by Frigyes Karinthy in 1929. Since then plenty of experiments have been conducted to explore this problem. In this project, we also want to calculate the degree of separation in statisticians community. They are calculated by Equation4:

$$DoS(G) = \frac{2\sum_{v_i \neq v_j \in V} distance(v_i, v_j)}{n(n-1)} \tag{4}$$

2.2.3 Clustering

To detect small communities in the large network, we investigate the network using various community detection methods: including Louvain algorithm by Blondel et al.(1), Newman-Girvan algorithm(4), , and Newman fast algorithm(3).

3 Experimental Results

3.1 Centrality

Based on the calculation of centrality for every nodes in the graph, we identified the most "influential" authors in this area using different method. The distribution of degrees for each node is show in Figure2 and 3. If follows the scale-free law. Based on degree centrality, we identified ten most cooperative authors: Ciprian M Crainiceanu, David Dunson, Holger Dette, Hongtu Zhu, James O Berger, Jianqing Fan, Joseph G Ibrahim, Lixing Zhu, Peter Hall, and Raymond J Carroll. Five of them are the winners of COPSS award up to 2015. COPSS is considered one of the two most prestigious awards in Statistics (The "Nobel Prizes" of Statistics). That means the estimation based on coauthorship network can give us meaningful result. Four among the 10 authors with the largest betweenness centrality won the COPSS award before 2015. Three of 10 authors with the largest closeness centrality won the COPSS award.

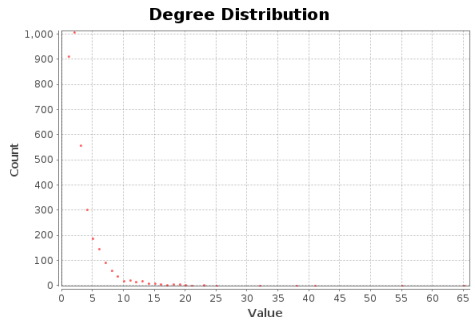

Figure 2: Degree of coauthorship network

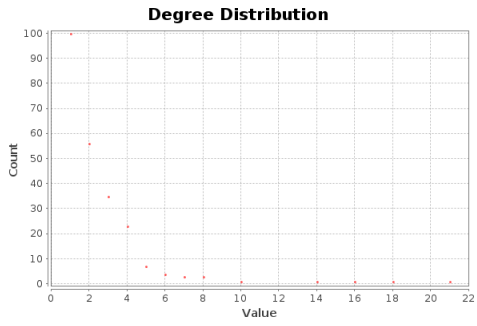

Figure 3: Degree of threshold giant coauthorship network

3.2 Degree of separation

As calculated, the degrees of separation of different networks are give in table1. The degree of separation of coauthorship network is about 6.9. That means each two members in the community can be connected by less than 7 people via coauthorship on average. The distance is a bit larger than

the real world separation. But considering the difficulty of coauthoring the same paper, the result is still very impressive. It also shows the "small world phenomenon" in statisticians' community.

3.3 Clustering

Visualization of several communities detected s shown in Figure 4 and 5. As is shown in Figure 4, which is the largest community detected from the threshold giant component of coauthorship network, all nodes centered at the nodes "Raymond J Carroll" and "Peter Hall". Both of them are the winners of COPSS award. They are also one of the nodes with largest centrality.

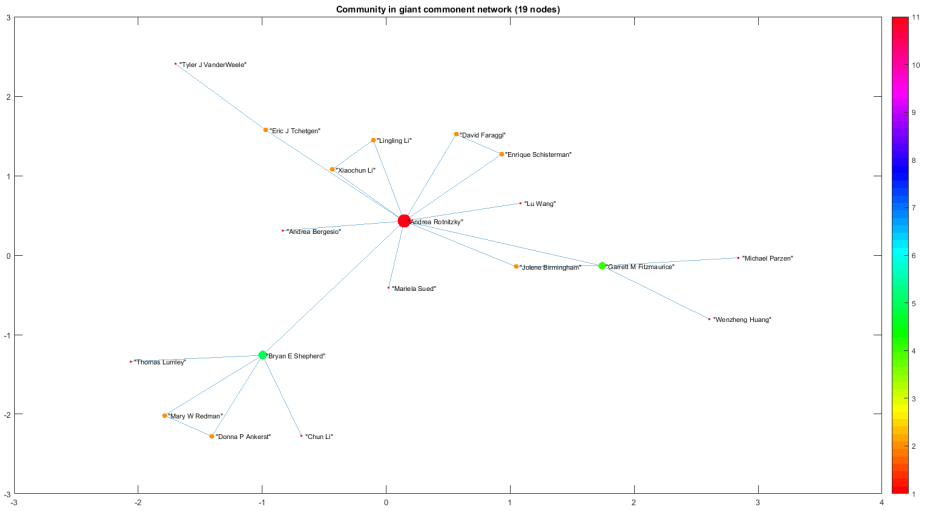

Figure 4: A community detected from threshold giant component of Coauthorship network (19 nodes).

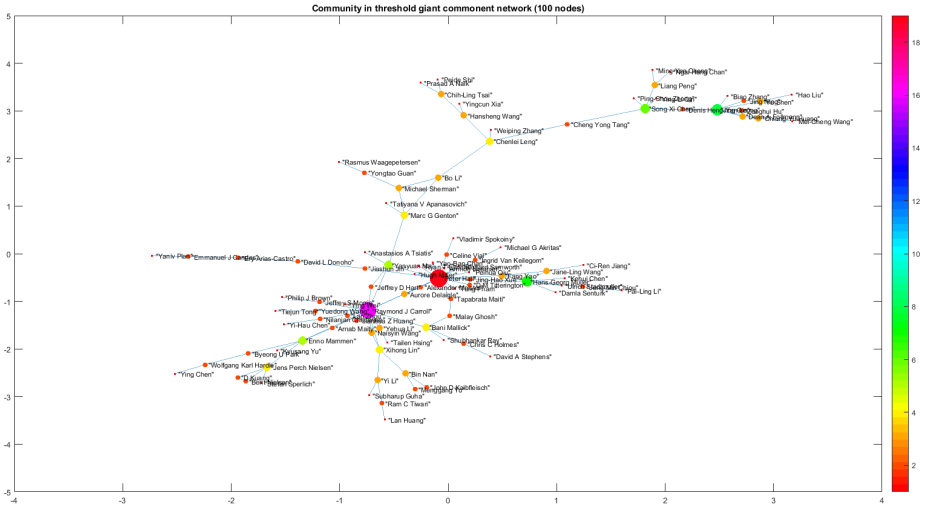

Figure 5: Largest community detected from giant component of Coauthorship network (100 nodes).

4 Discussion

More work can be done by combining the ground-truth knowledge about the community with network analysis in future. For example, even though we clustered the whole network into several small clusters, we couldn't analyze and explain the result effectively. Because the knowledge about various schools are often based on experience. Different schools usually have different focuses on research topic. Further study can be made combining the paper information and identify the focus of various schools, also verifying the reliability of clustering based on coauthorship relationships.

References

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[2] P. Ji, J. Jin, et al. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

[3] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[4] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.