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 and citation relationships for statistic community. To model and analyze coauthorship or citation 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/citation 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(1), 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. Similarly citation network is directed where the citation from one author to another is represented as a directed link.

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*) citation network among the whole community, where each edge denotes at least one paper coauthored, containing 3607 authors in total; *b*) the giant component of citation network, where each edge denotes at least one paper coauthored, containing 2263 authors in total; *c*) the giant component of citation network, where each edge denotes at least one paper coauthored, containing 236 authors in total. Different versions are given because the complete citaiton network is scattered, with many isolated clusters. It is calculated that there are 369 separated components in the whole network. And the components follows the scale-free distribution(also called long tail

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

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 gaint component (the largest 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 conceptof the community. So the other two versions of network are also given.

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.

In network analysis, to identify the most "influential" authors in statisticians community, centrality is used to find the most important vertices within a graph . 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.

3 Experimental Results

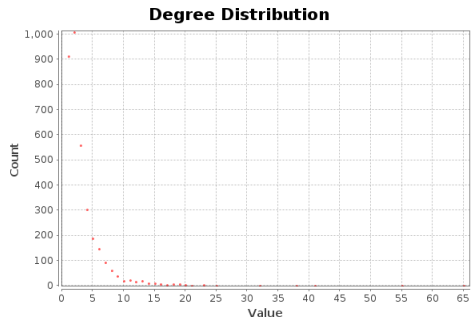

Figure 1: Degree distribution of coauthorship network

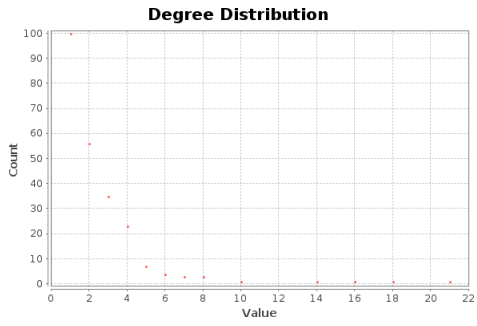

Figure 2: Degree distribution of treshold giant coauthorship network

References

[1] Ji, P., & Jin, J. (2016). Coauthorship and citation networks for statisticians. The Annals of Applied Statistics, 10(4), 1779-1812.

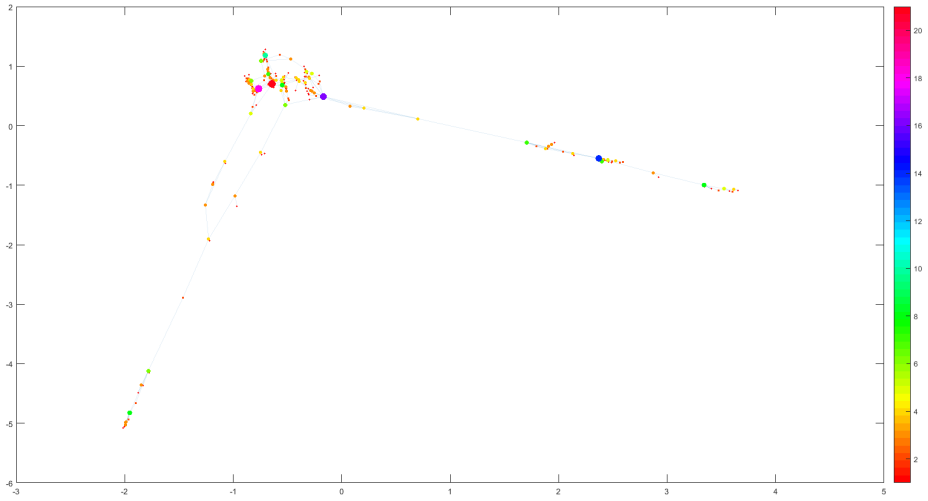

Figure 3: Threshold gaint component of Coauthorship network

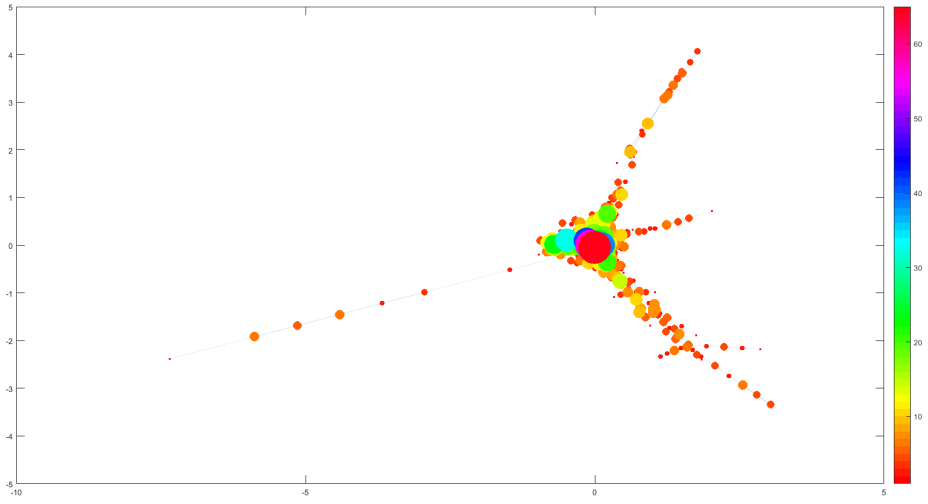

Figure 4: Gaint component of Coauthorship network