The Network of Network Scientists

CHEN Xiaolong

The Hong Kong University of Science and Technology
xchenfg@connect.ust.hk

Abstract. This report mainly introduces some properties of the co-authorship network of some network scientists, including the degree distribution, clustering coefficient, average shortest path and the similarity. In addition, the Girvan-Newman algorithm is applied to this network to do community detection and site percolation is conducted to test the resilience of this network.

Keywords: Co-authorship network · Degree distribution · Community detection · Site percolation.

1 Introduction

1.1 Data Selection

The data set is downloaded from `https://networkrepository.com`, describing the co-authorship of scientists in network theory and experiments[4].

1.2 Network Description

In this undirected network, which consists of 379 nodes and 914 edges, the nodes represent scientists and the edges represent co-authorship. Every scientist were encoded as a number and every edge represents the existence of collaboration between corresponding the two scientists. The picture of this network is presented in Fig. 1. Using color to denote the value of degree, Fig. 2 can be generated. From this picture, one can conclude that most of the network scientists do not have over 10 collaborators and there may be several communities, but it is not easy to classify just from the picture.

In the remainder of this report, a number of properties of this network will be introduced, such as the degree distribution, small world properties and network resilience and so on.

2 Degree Distribution

The definition states that the degree of a vertex in a network is the number of edges connected to it[3]. Normally, we use k_i to represent the degree of node i.

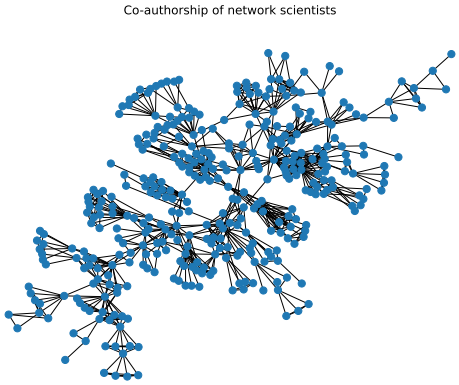


Fig. 1. This is the picture of the network. The nodes represent scientists and the edges represent the existence of co-authorship.

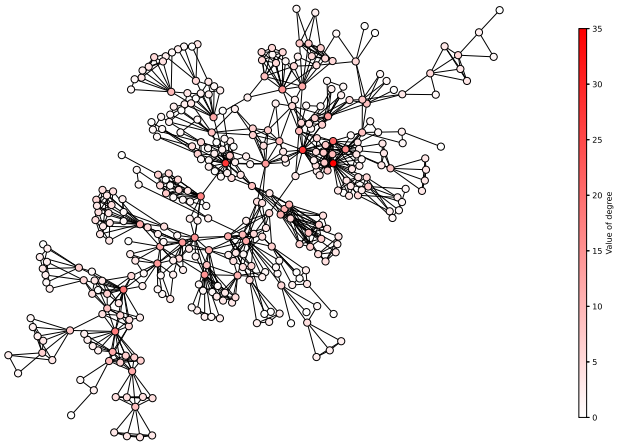


Fig. 2. This is the picture of the network with the color of a node representing the value of the degree.

The degree distribution is the probability distribution of the degrees over the whole network[5]. Using n_k to denote the number of nodes with degree k and n

to denote the total number of nodes, the degree distribution can be written as

$$P(k) = \frac{n_k}{n}. \tag{1}$$

The degree distribution of this network is given in Fig. 3. One can draw the conclusion that in this network, most of network scientists have $1 \sim 15$ collaborators.

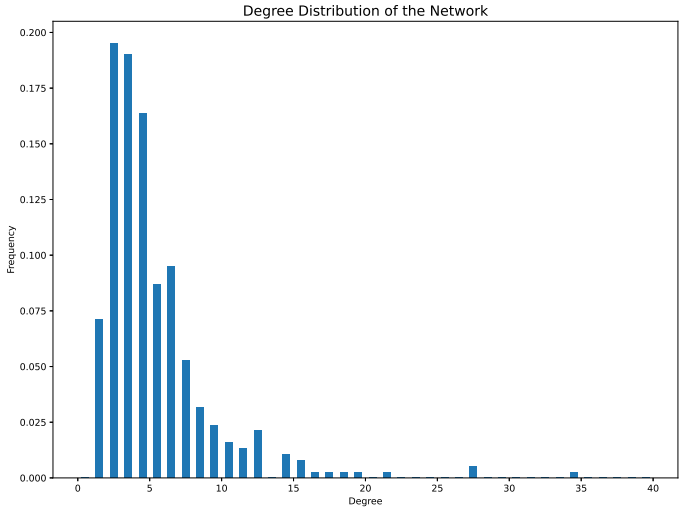


Fig. 3. This histogram illustrates the degree distribution of the co-authorship network. The horizontal axis represents the degree values and the vertical axis represents the proportion of nodes with the corresponding degree.

3 Small World Properties

A small-world network is a network with high clustering and low average shortest path. In small-world networks, most nodes are not neighbors of one another, but it is likely that the neighbors of any given node are neighbors of each other and most nodes can be reached from every other node by a small number of steps[6]. If a network satisfies conditions (2) and (3), it can be regarded as a small-world network.

$$C \gg C_{random} \tag{2}$$

$$l \approx ln(N) \tag{3}$$

The network is connected, meaning that there is only one component in this network, which allows us to compute the shortest average path directly. In this co-authorship network, the average clustering coefficient is 0.74123 and the average shortest path is 6.04187 (both rounded to keep five digits after the decimal point), which satisfies the conditions mentioned above, indicating that this network is indeed a small-world network.

$$C = 0.74123 \tag{4}$$

$$l = 6.04187 \tag{5}$$

An interesting phenomenon of small-world networks is six degrees of separation, which was first proposed by Frigyes Karinthy[2], referring to the idea that everyone is on average approximately six steps away from any other person. This phenomenon can also be discovered in the co-authorship network. The average shortest path of this network happens to be around 6, meaning that on average, a network scientist in this network can reach any other scientist in about six steps.

4 Scale-free Properties

Distributions of the form of equation (6) are called power law and networks with power-law degree distributions are called scale-free networks[3].

$$p_k = Ck^{-\alpha} \tag{6}$$

$$P(X > k) \sim \int_k^{\infty} x^{-a} dx \sim \left(x^{-a+1}\right)_k^{\infty} \sim k^{-a+1} \tag{7}$$

The cumulative degree distribution is given by Fig. 4. The tail part of this can be fitted to the function (all of the parameters are rounded to keep five digits after the decimal point)

$$P(X > k) = 112.022k^{-3.059}. \tag{8}$$

The fitting result suggests that the tail part of the degree distribution conforms the power law, indicating that this network is scale-free.

5 Similarity

Another concept in social network analysis is the similarity between nodes. Basically, there are two fundamental ways to construct measures of network similarity, which are structural equivalence and regular equivalence[3]. Two nodes are structurally equivalent if they have many of the same neighbors and are regularly equivalent if they have neighbors who are themselves similar. Since measures for structural equivalence are better developed than those for regular equivalence, this report will mainly introduce two measures for structural equivalence, which are cosine similarity and Pearson correlation coefficient.

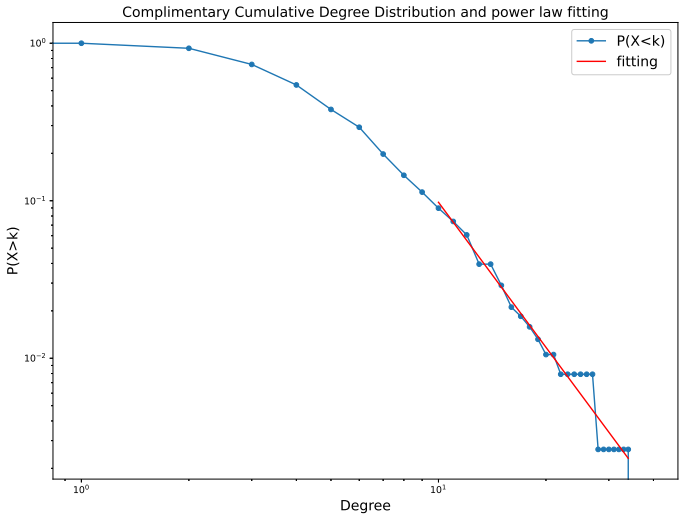


Fig. 4. This figure describes the complimentary cumulative distribution of the co-authorship network of network scientists.

5.1 Cosine Similarity

The number of common neighbors of two nodes is given by

$$n_{ij} = \sum_k A_{ik} Akj, \tag{9}$$

where A is tha adjacency matrix. Equation (9) can be normalized as

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{k_i k_j}}. \tag{10}$$

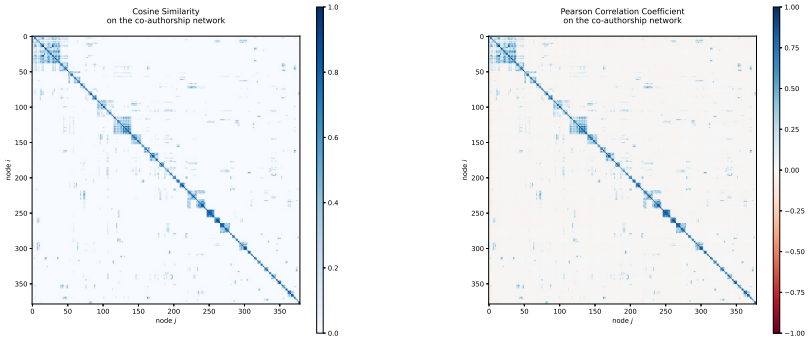
This is called cosine similarity because it is similar to the calculation of cosine value of an angle.

5.2 Pearson correlation coefficient

Another way to normalize the count of of common neighbors is to compare it with the expected value that count wold take on a network in which vertices choose their neighbors in a random way[3], which leads us to Pearson correlation coefficient. The normalized calculation is given by

$$r_{ij} = \frac{\sum_h \left(A_{ih} - \langle A_i \rangle\right)\left(A_{jh} - \langle A_j \rangle\right)}{\sqrt{\sum_h \left(A_{ih} - \langle A_i \rangle\right)^2}\sqrt{\sum_h \left(A_{jh} - \langle A_j \rangle\right)^2}} = \frac{\mathrm{covar}(A_i, A_j)}{\sqrt{\mathrm{var}(A_i)\mathrm{var}(A_j)}}. \tag{11}$$

The similarity of the co-authorship network is given in Fig. 5, from which one can see that there are only small groups of similar nodes, corresponding to the fact that it is not easy to tell the number of communities in this network.



(a) Cosine similarity (b) Pearson correlation coefficient

Fig. 5. This figure describes the similarity of the co-authorship network.

6 Community Detection

The goal of community detection is to separate the network into groups of nodes so that every group has few connections with any other group.

6.1 Hierarchical Clustering

With different criteria, the number of the communities that will be detected may change. For example, we can group the network into 2 communities while it can be separated into 5 communities in other conditions. In order to describe all of the situations, hierarchical clustering is proposed, which is about building a hierarchical structure of communities based on network topology. One of the algorithms that can handle this job is the Girvan-Newman algorithm.

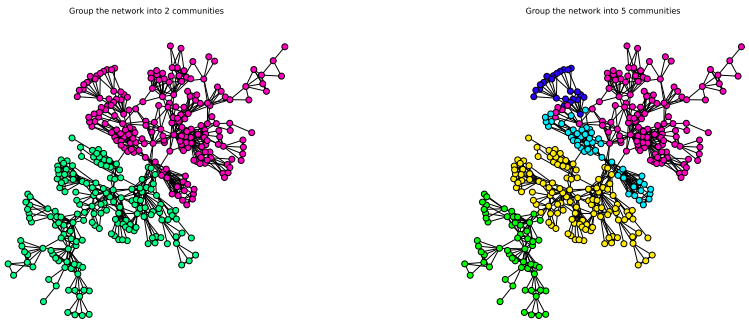
6.2 Girvan-Newman Algorithm

The Girvan-Newman Algorithm is an algorithm often used to detect communities of a network. Before introducing this algorithm, the definition of the edge betweenness should be given first. The edge betweenness of an edge is defined

as the number of shortest paths between pairs of nodes that run along it[1]. The process of the Girvan-Newman algorithm is given below.

1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweennesses for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.

By applying this algorithm, a network can be divided into a certain number of components, as is shown in Fig. 6, and the hierarchical structure is shown in Fig. 7, in the form of a dendrogram.



(a) 2 Communities (b) 5 Communities

Fig. 6. By using the Girvan-Newmen algorithm, nodes in this network can be grouped into $2 \sim n$ communities. (a) shows the situation where the network is separated into 2 communities and (b) describes the situation where it is grouped into 5 communities.

6.3 Modularity Maximization

A good measurement of the integrity of communities is modularity. To find the best way to distinguish communities, we just need to figure out the number of communities which results in the maximum modularity. The relation between the modularity value and the number of communities is given in Fig. 8. From the calculation, the maximum modularity value is around 0.844, which is reached when the network is grouped into 18 communities. With this result, Fig. 9 can be generated.

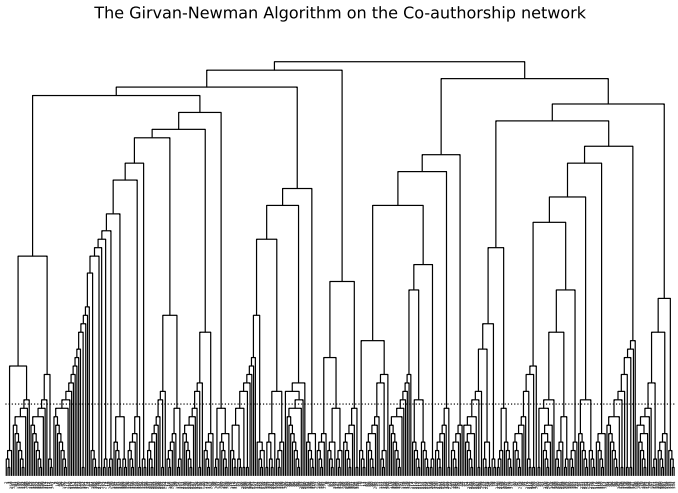


Fig. 7. This figure illustrates the hierarchical clustering of the co-authorship network, using the Girvan-Newmen algorithm.

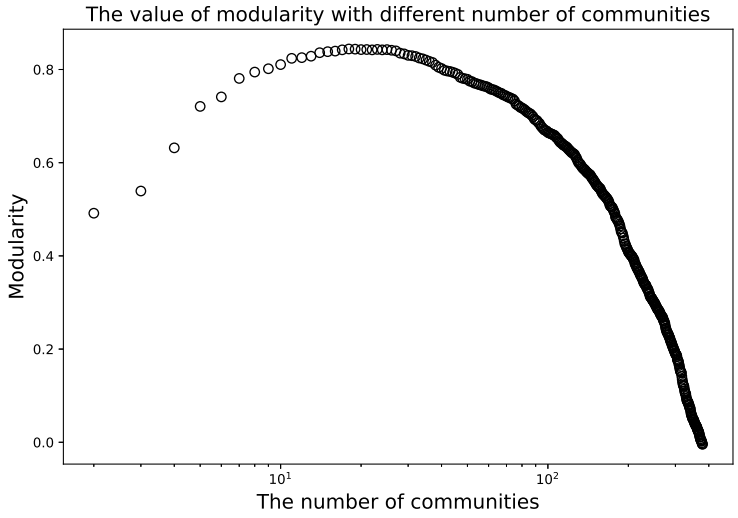


Fig. 8. This figure describes the change of modularity with different number of communities.

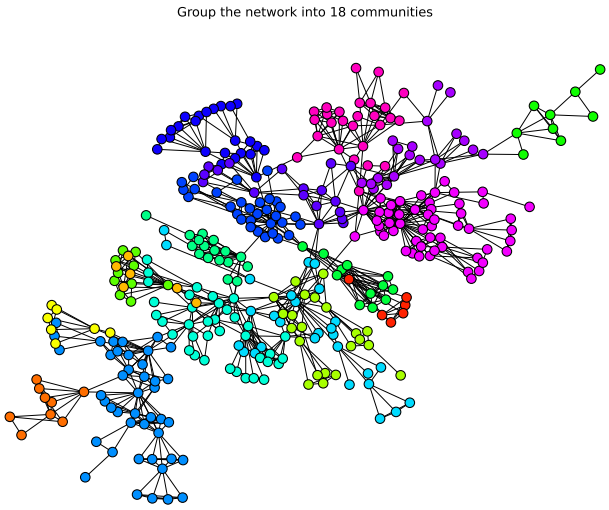


Fig. 9. In the case where the modularity is maximized, the network scientists is divided into 18 communities. Nodes with the same color represent network scientists belonging to the same community.

7 Network Resilience

Normally, site percolation is often used to test a network's resilience, which is about the removal of nodes. This process can be used as a model of a variety of real-world phenomena. For example, the failure of routers on the Internet, the vaccination or immunization of individuals against the spread of disease and so on[3].

Random attack means that in the process of percolation, the removal of nodes is random. Targeted attack means that the removal of nodes is in a certain order. In this report, random attack is repeated for 50 times to generate a smooth curve and targeted attack is based on the order of the degree of nodes, which means that the node with the highest degree will be removed in every turn. Performing random attack and targeted attack on this network, the relation between the giant component's size and the fraction of nodes removed is shown in Fig. 10. The giant component's size decreases much more slowly in the situation of random removal than targeted removal.

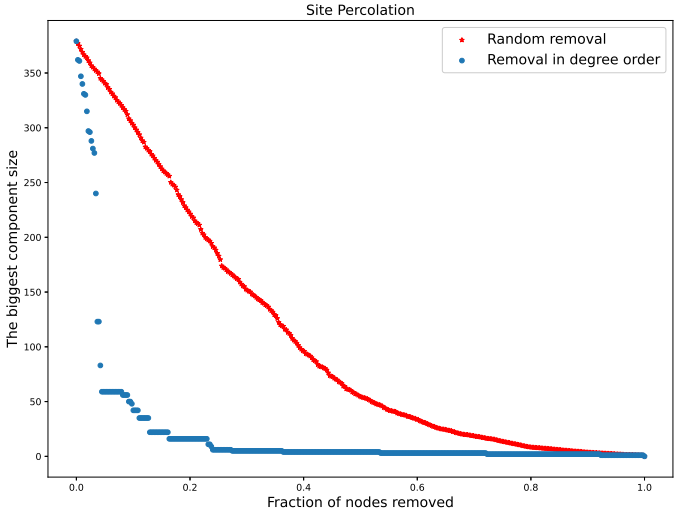


Fig. 10. This figure illustrates the change of the size of the giant component with different fraction of removed nodes.

References

1. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the national academy of sciences **99**(12), 7821–7826 (2002)
2. Karinthy, F.: Chains. Everything is different, Budapest (1929)
3. Newman, M.: Networks: An Introduction. Oxford University Press, Inc., USA (2010)
4. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI (2015), `https://networkrepository.com`
5. Wikipedia contributors: Degree distribution — Wikipedia, the free encyclopedia (2021), `https://en.wikipedia.org/w/index.php?title=Degree_distribution&oldid=1025200244`, [Online; accessed 4-December-2021]
6. Wikipedia contributors: Small-world network — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Small-world_network&oldid=1055631284` (2021), [Online; accessed 5-December-2021]