哈尔滨工业大学深圳研究生院

社交网络分析

Social Network Analysis

姓　　名：__胡文馨____　学　　号：__17S151563_

报告日期__2017/4/29__

Content

1 Introduction

1.1 The purpose of project

In this project, I will utilize an Astro Physics collaboration network. Based on this network, an experiment which compares several algorithms of centrality analysis will be conducted. It is interesting to do centrality analysis. In details, I compare centrality analysis of different algorithms according to some evaluation metrics and try to reveal some relative reference among these methodologies. Each node represents an author in the network, and central nodes mean that these authors has relatively great impact on this field as a consequence of many collaborations with others.

1.2 Experiment Content

The purpose of this project is to analyze a social network by using techniques I have learnt in the class. In the project, analyzing statistics of network is a necessary part, and I are supposed to choose 1 or 2 topics as following:

- community detection
- human evaluation and signed social networks analysis
- cascading behavior
- influence maximization
- outbreak detection
- network evaluation
- link prediction etc.

1.3 Experiment Submission

Each student submits a project report, and each group selects one students to give the presentation. Project reports from different students in a group should be different.

2 Dataset Information

2.1 Dataset Statistics

At beginning, the most necessary part is to learn properties of the network. Arxiv ASTRO-PH (Astro Physics) collaboration network is from the e-print arXiv and covers scientific collaborations between authors papers submitted to Astro Physics category. If an author i co-authored a paper with author j, the graph contains a undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on k nodes.

The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its ASTRO-PH section.

To get the properties and a better understanding of this network, I analyze this network and visualize this graph. There are some dataset statistics listed as following:

Property	**Value**
Nodes	18772
Edges	198110
Nodes in largest WCC	17903 (0.954)
Edges in largest WCC	197031 (0.995)
Nodes in largest SCC	17903 (0.954)
Edges in largest SCC	197031 (0.995)
Average clustering coefficient	0.6306
Number of triangles	1351441
Fraction of closed triangles	0.1345

Table 2.1 Basic Property of the Graph

2.2 Image Analysis

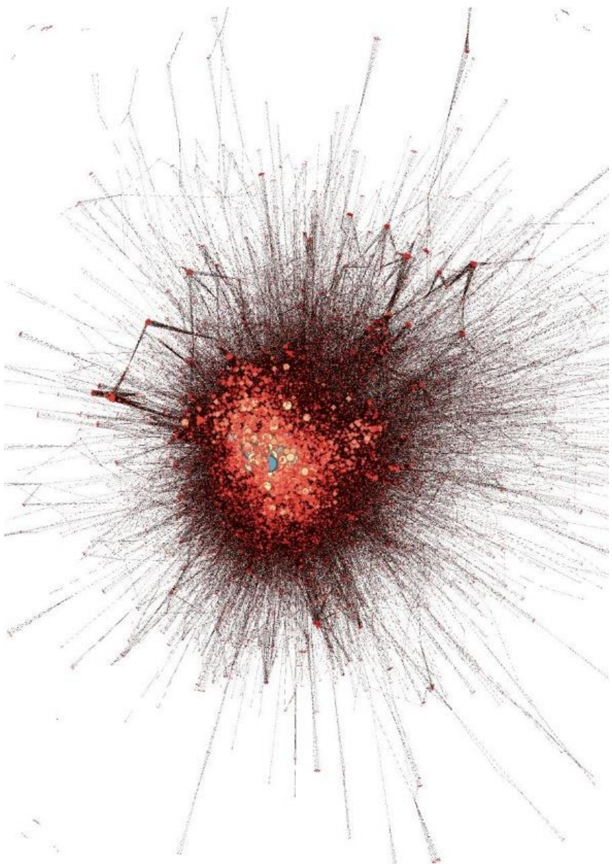

Fig1 visualization of network

In Gephi software, we can look at the visualization of network image. Through the image, I find that there is a biggest node that is the core. And there are too many nodes also close to the core. In the outer layer, there are a little nodes that have small degree.

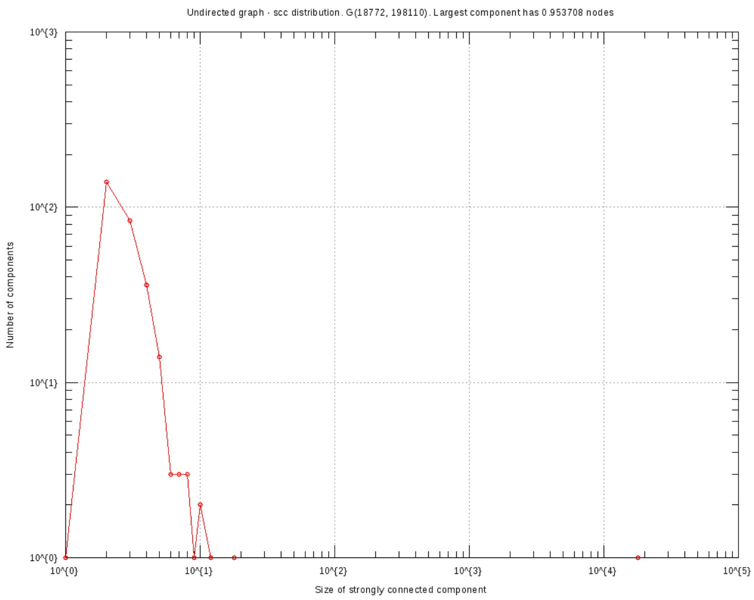

Fig2 the distribution of sizes of strongly connected components of Graph

Look at the Fig2, we can find that the connected components of each point are very different. The biggest connected component is just one node, the smallest connected components are four nodes.

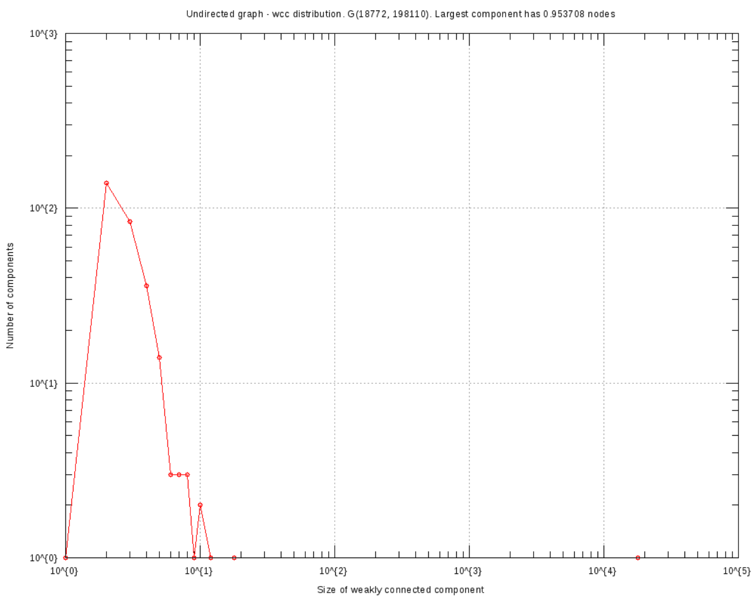

Fig3 the distribution of sizes of weakly connected components of Graph

Look at the Fig3, we can find that Fig3 and Fig4 are same. So we can conclude that it is very popular to create an article commonly in the history of Astro Physics.

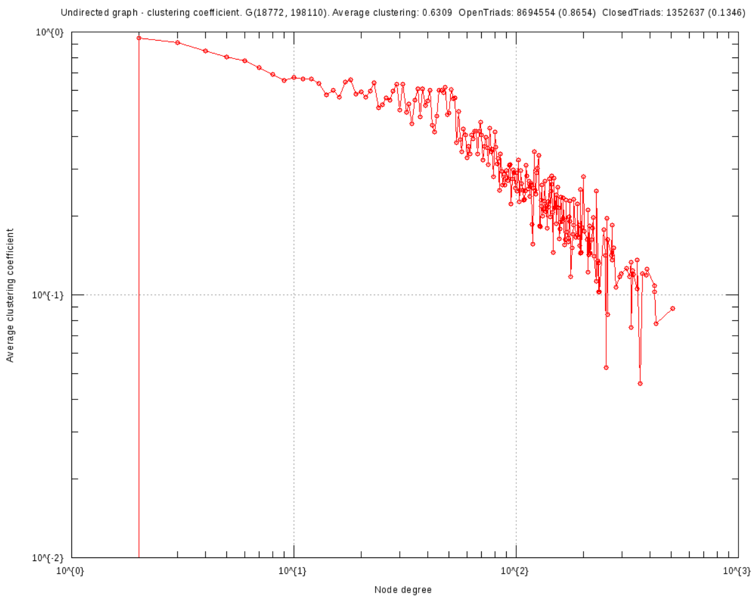

Fig4 the distribution of clustering coefficient of Graph

Look at the Fig4, we can find that high density connection nodes account for one fourth of the total. The clustering coefficient of the remaining nodes drops sharply.

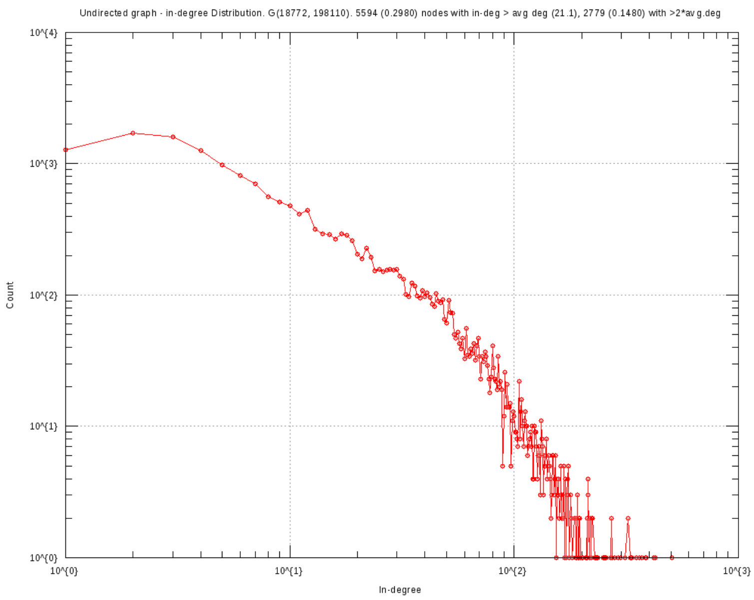

Fig5 the in-degree distribution of Graph

Look at the Fig5, we can find that degree distribution is basically linear. There are five nodes have evident hub effect. More than twenty nodes have one degree. So they are lonely doing research. Most nodes are concentrate on degree of ten.

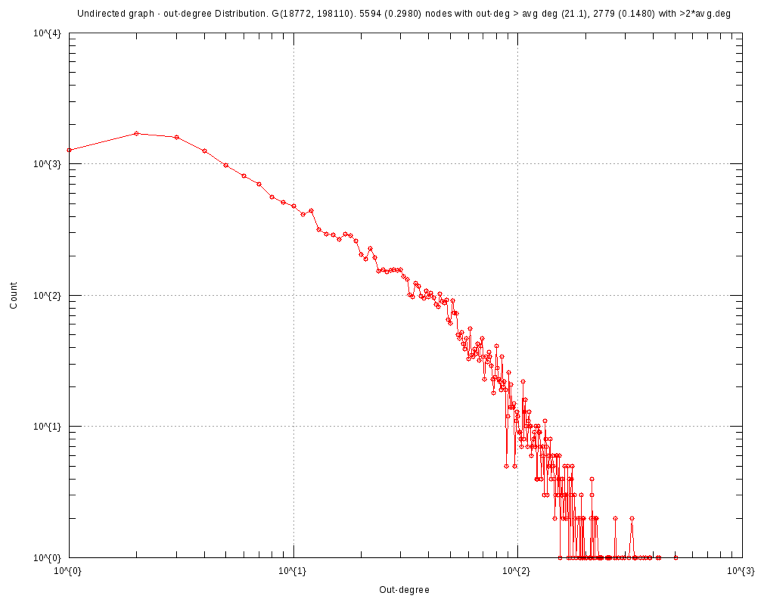

Fig6 the out-degree distribution of Graph

Look at the Fig6, we can find that Fig5 and Fig6 are same. So we can conclude that it is very popular to create an article commonly in the history of Astro Physics. But there are also independent nodes.

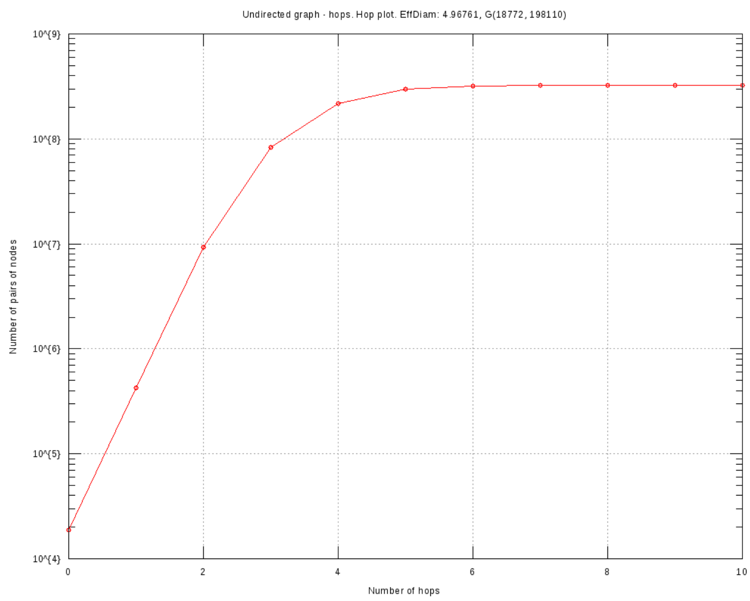

Fig7 the cumulative distribution of the shortest path lengths of Graph

Look at the Fig7, we can find that the cumulative distribution of the shortest path lengths grows fast sharply in the first half. It grows slowly in the second half.

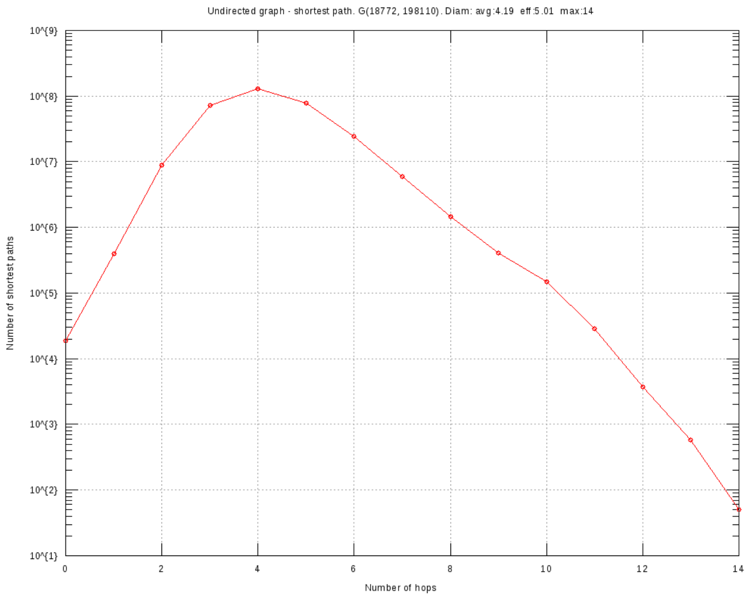

Fig8 the distribution of the shortest path lengths in Graph

Look at the Fig8, we can find that the distribution of the shortest path lengths is almost gaussian distribution. And the value that are distributed at most nodes shows a decreasing trend.

3 Centrality Analysis

3.1 Degree Centrality

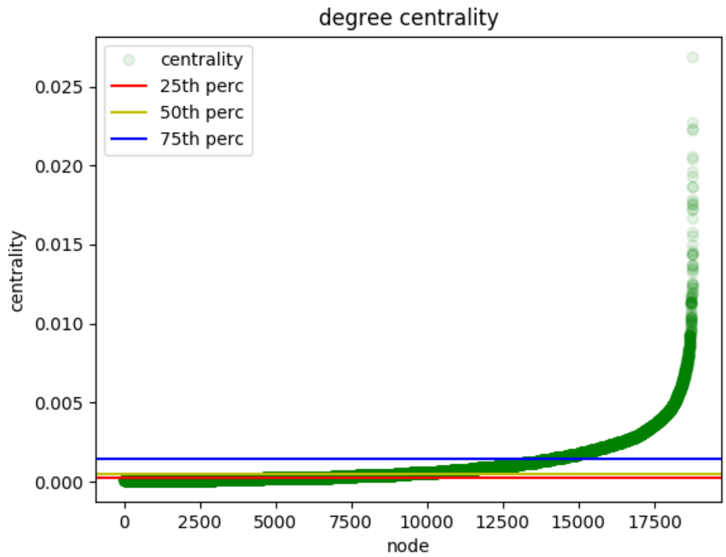

Fig9 plot of degree centrality

Look at the Fig9, we can see 17500 node have small degree, however, another nodes have huge degree. So in the network, it must be hub effect. A few nodes have more relation.

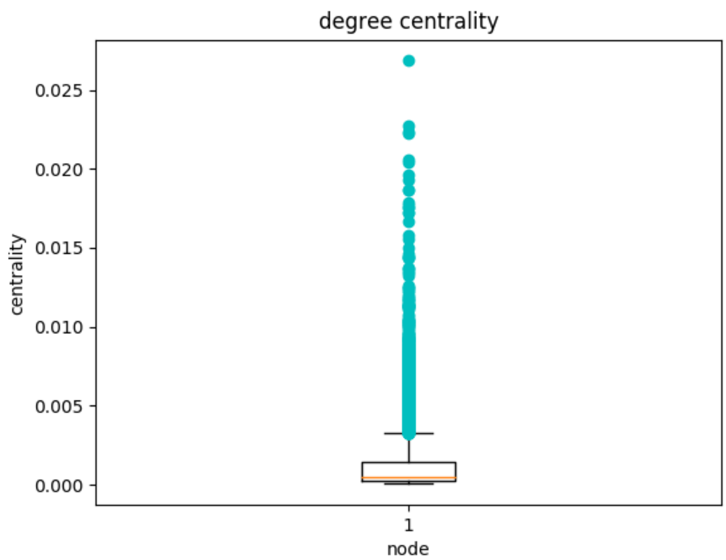

Fig10 boxplot of degree centrality

Look at the Fig10, we can see a node have the biggest degree. Seventy-five percent nodes are only 0.002 degree. In the Astro Physics field, there is a powerful person. Five percent of people are also have important effect.

3.2 Betweenness Centrality

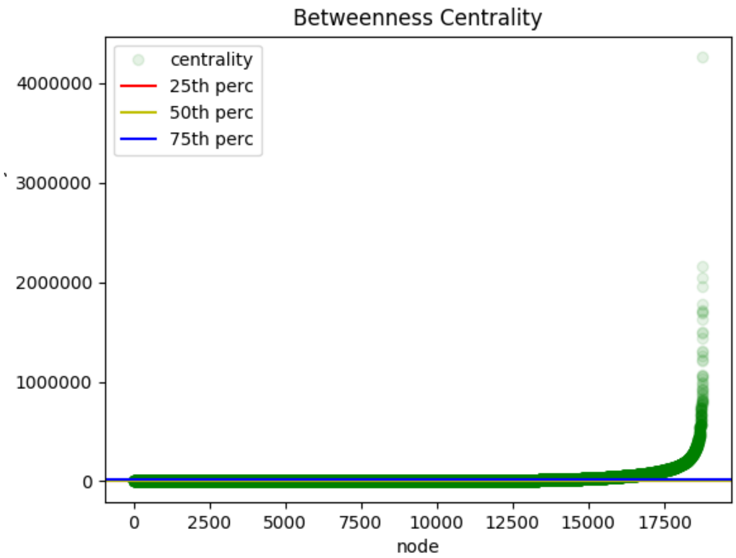

Fig11 plot of betweenness centrality

Look at the Fig11, we can find that there is one node particularly obvious have high betweenness centrality. It means that the node is connected with a lot of nodes.

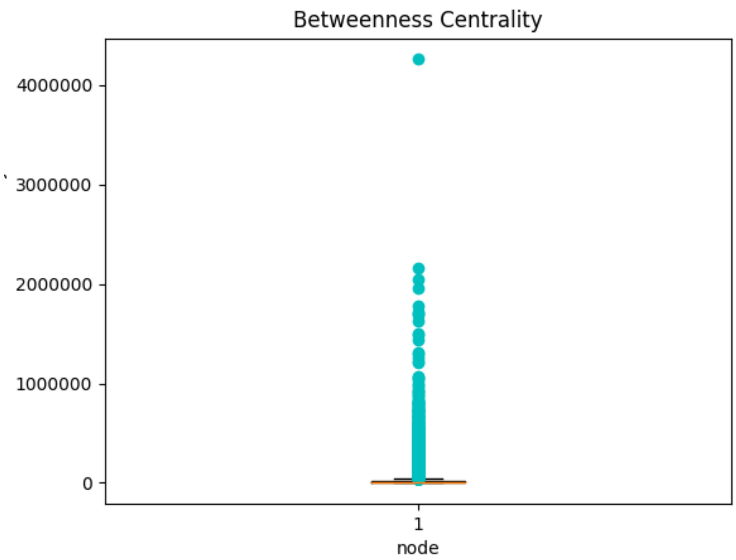

Fig12 boxplot of betweenness centrality

Look at the Fig12, we can find that there are many edges that pass through this node.

3.3 Closeness Centrality

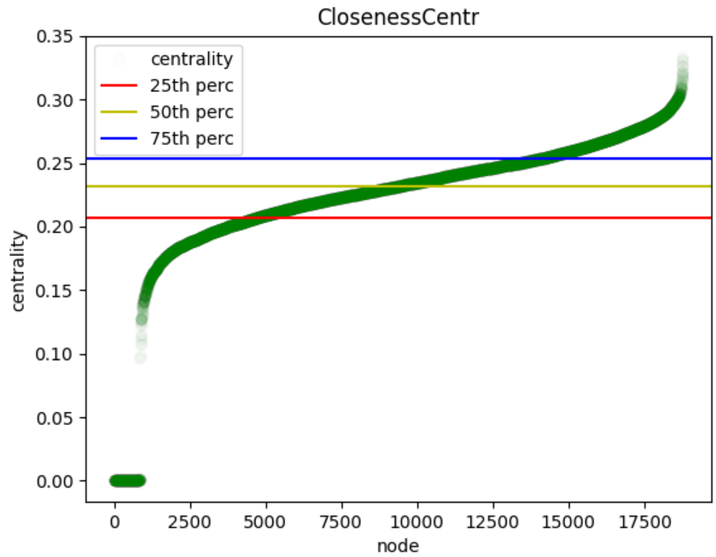

Fig13 plot of closeness centrality

In a connected graph, closeness centrality of a node is a measure of centrality in a network, calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus the more central a node is, the closer it is to all other nodes.

Look at the Fig13, we can find that there are a few nodes are independent. Closeness of most nodes is similar with each other and is 0.224.

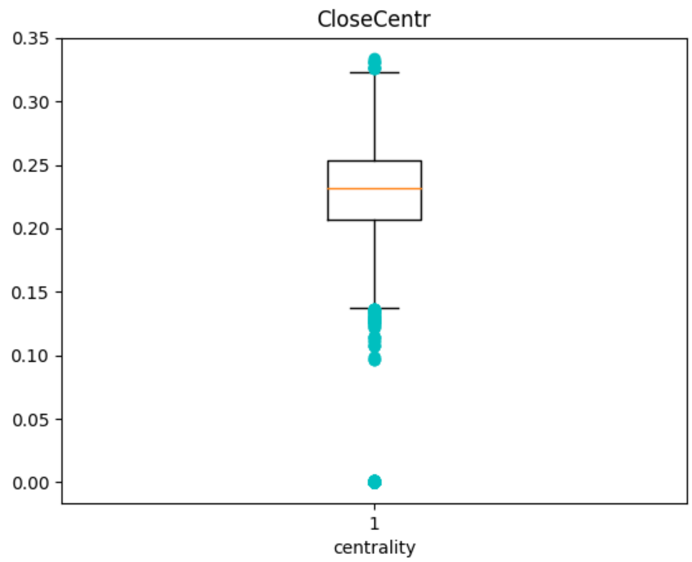

Fig14 boxplot of closeness centrality

Look at the Fig14, we can find that there are outliers in the bottom. And a few nodes have high centrality. Closeness centrality of most nodes are in the range of [0.20, 0.25].

3.4 Farness Centrality

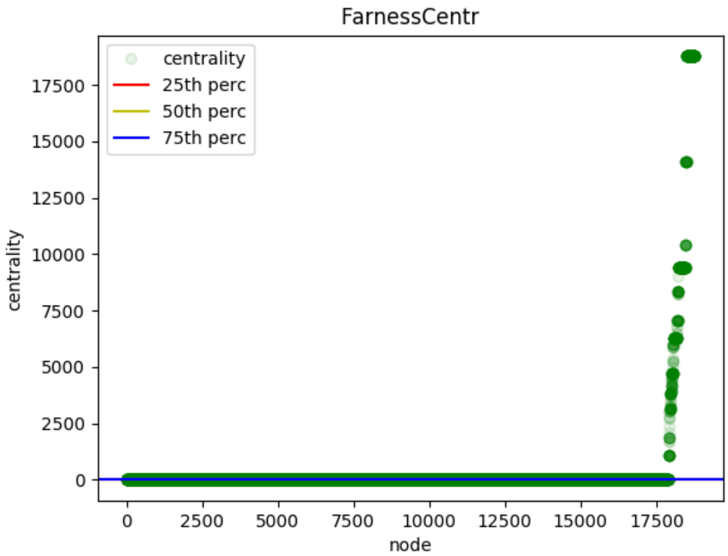

Fig15 plot of farness centrality

Look at the Fig15, we can find that when nodes more than 17500, the closeness centrality start rapidly growth. It means that there are 1500 nodes are far away to each other. The farness centrality of several nodes is very high.

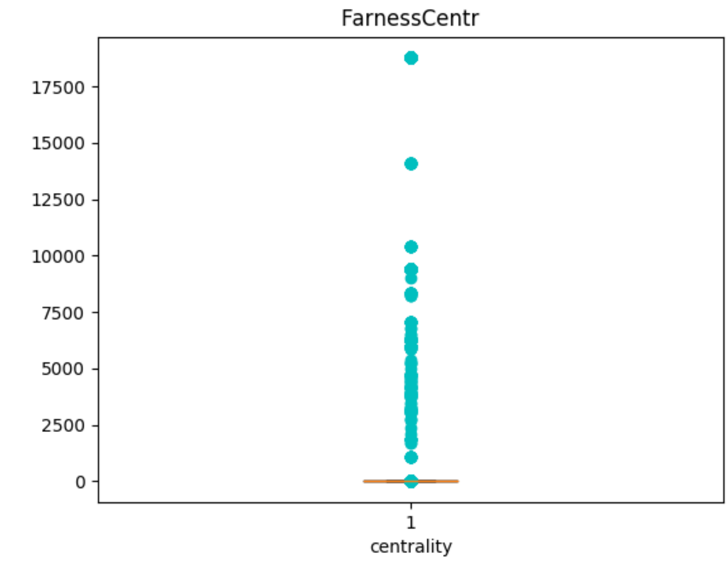

Fig16 boxplot of farness centrality

Look at the Fig16, we can find that not more than ten nodes have high farness centrality. It means that they are far away from the center. Most of nodes are closely connected.

3.5 PageRank

PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank is a way of measuring the importance of website pages.

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

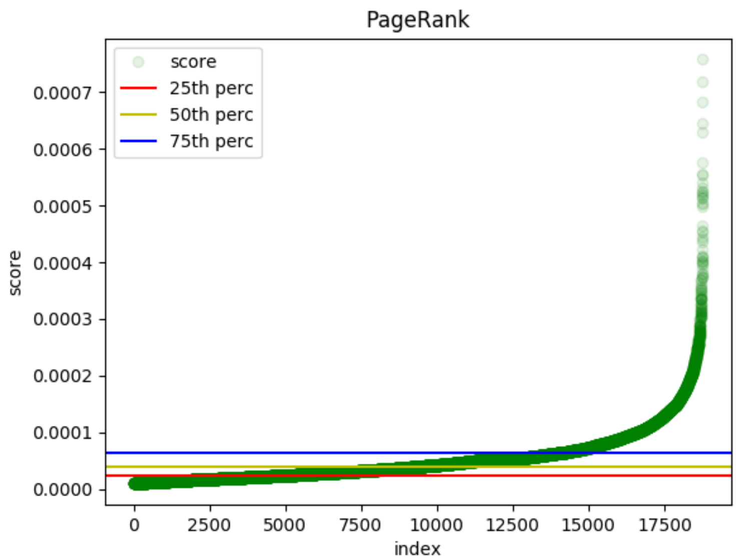

Fig17 plot of PageRank

Look at the Fig13, the importance of PageRank increases sharply from 17500 nodes. It means that one node has links with a lot of nodes. And other four nodes also have high links.

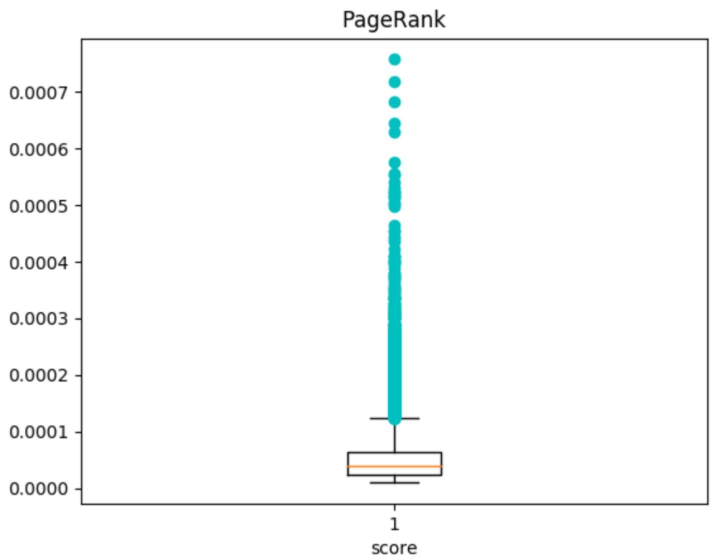

Fig18 boxplot of PageRank

Look at the Fig14, five nodes have high links. High join nodes are average 0.0004. Low join nodes are average 0.00004. And scores of high join nodes are much larger than the scores of low join nodes.

3.7 Hits

Hyperlink-Induced Topic Search is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages.

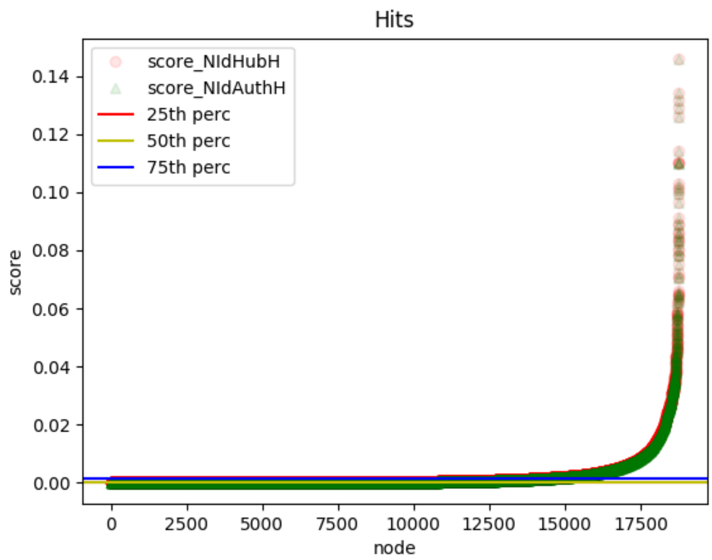

Fig19 plot of Hits

Look at the Fig15, we can find that the hub and authority corresponds to the same nodes. A good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. In the Fig14, the scores of Hits are more than the scores of PageRank in same nodes.

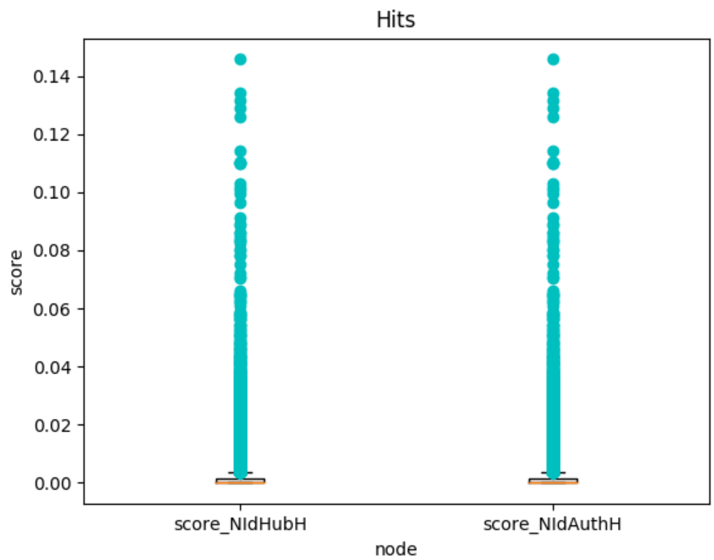

Fig20 boxplot of Hits

The scheme therefore assigns two scores for each page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

Look at the Fig16, we can find that one node has highest score and four nodes have high score. But scores of most of nodes are near zero.

4 Node2vec

4.1 Basic Principle

Node2vec and word2vec are essentially using the connections between adjacent nodes. Nodes in the network generally have two similar measures: 1. Content similarity, 2. Structural similarity. Main content of similarity is the similarity between the adjacent nodes, and the structure of those points are not necessarily adjacent, may separate far away, and this is why in this paper, the combination of DFS and BFS to choose the cause of the neighbor nodes.

```
('model["1086"] = ', array([-0.07231235,  0.30187324,  0.15698448, -0.15184267,  0.13261543,
       -0.2028288 , -0.00092828,  0.23297971,  0.5835233 ,  0.18113875,
        0.00360208, -0.27757767, -0.20351695,  0.21937127,  0.33168945,
       -0.1989391 , -0.10118499,  0.04425671,  0.5110656 , -0.03082806,
       -0.3282296 ,  0.46545663,  0.584056  ,  0.68339854, -0.02179752,
        0.02773618, -0.04883273,  0.29201907,  0.27497014,  0.26552483,
        0.13083765,  0.03766041, -0.06655931,  0.00774123,  0.3578017 ,
        0.3174469 , -0.07187765, -0.22457123, -0.13630633, -0.25600925,
        0.43299937,  0.06920531,  0.0430029 ,  0.47103477, -0.14694612,
        0.33643577,  0.30708006,  0.28828743,  0.38008878, -0.15936306,
        0.44067132,  0.05655835,  0.09670754,  0.3929237 ,  0.51134366,
        0.20138296,  0.30725598, -0.1227631 ,  0.59151894,  0.06068979,
        0.09399573, -0.28530663,  0.05356124,  0.00981973, -0.4056933 ,
       -0.03983407,  0.01103353, -0.34046656,  0.47976226, -0.08119454,
        0.20140804,  0.15261738,  0.37530705, -0.08408635,  0.15389547,
        0.425926  ,  0.5406564 ,  0.22249147,  0.40187216, -0.2802542 ,
        0.41896084,  0.652109  ,  0.41568854,  0.19311278,  0.14623548,
       -0.02400118, -0.1091477 ,  0.1537478 ,  0.24489312,  0.08220242,
        0.32757688, -0.07194504, -0.00977237, -0.1756546 ,  0.51491123,
       -0.30826002,  0.11970973,  0.09582325,  0.16487767,  0.14876683,
       -0.33440945, -0.2985143 ,  0.1633799 , -0.5911379 , -0.37326443,
        0.15551566, -0.20442714, -0.218923  ,  0.0769992 ,  0.37222695,
       -0.01871694, -0.13152102,  0.66716725, -0.24377088,  0.4577827 ,
        0.14409073,  0.30436084,  0.08994904, -0.15860948, -0.8434897 ,
       -0.08197689, -0.00487631, -0.11615458,  0.15346287,  0.36246732,
       -0.47474745,  0.13751188, -0.55509245], dtype=float32))

Process finished with exit code 0
```

Fig21 the vector representation of the 1086 node

4.2 Calculate Similar Nodes

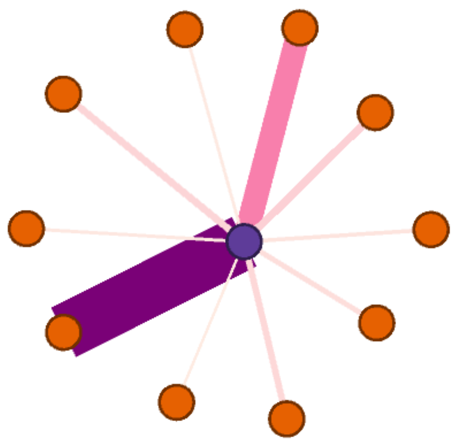

Fig22 similar nodes of 53213

Source,Target,Weight
53213,21718,0.887
53213,37290,0.786
53213,76749,0.733
53213,71856,0.73
53213,73007,0.728
53213,64296,0.724
53213,122294,0.72
53213,6288,0.718
53213,62821,0.717
53213,89093,0.716

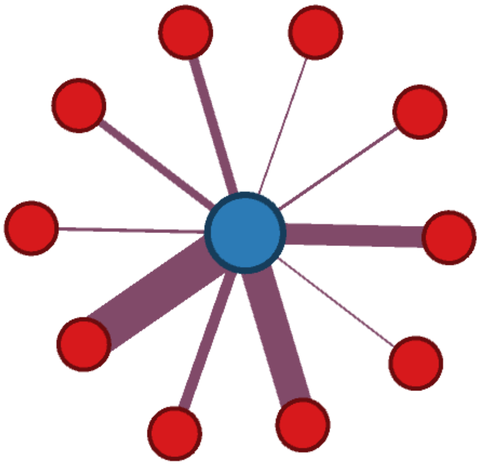

Fig23 similar nodes of 62821

Source,Target,Weight
62821,90402,0.812
62821,77959,0.794
62821,37290,0.778
62821,53577,0.762
62821,48871,0.76
62821,60308,0.757
62821,1187,0.75
62821,122294,0.75
62821,46589,0.748
62821,43470,0.748

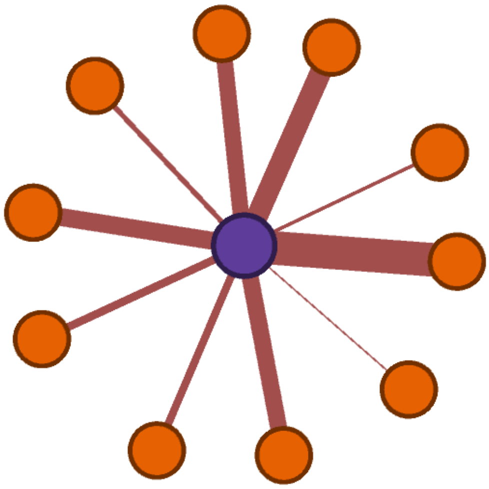

Fig24 similar nodes of 38109

Source,Target,Weight
38109,94235,0.85
38109,118342,0.812
38109,23986,0.801
38109,34608,0.798
38109,89732,0.796
38109,90128,0.77
38109,117443,0.769
38109,131993,0.761
38109,91619,0.756
38109,75223,0.75

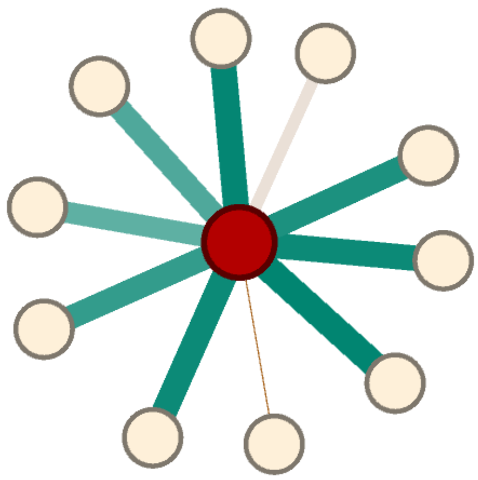

Fig25 similar nodes of 1086

Source,Target,Weight
1086,4685,0.819
1086,3598,0.818
1086,725,0.816
1086,4501,0.816
1086,51101,0.812
1086,2627,0.806
1086,38608,0.799
1086,8745,0.795
1086,2167,0.748
1086,5066,0.694

5 Conclusion

Degree centrality is the most direct index to characterize the centrality of nodes in network analysis. The greater the node degree of a node means the higher the degree centrality of the node is, the more important the node is in the network.

This paper measures centrality through degree, betweenness, closeness and farness. The centrality of nodes is evaluated by comparing PageRank and hits algorithms.

In this paper, node2vec algorithm uses adjacent nodes to represent a node, thus computing the similarity of different nodes.

References

[1] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.