MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **72** (2014) 577-588

ISSN 0340 - 6253

Scientific Collaboration Network Based on *MATCH Communications in Mathematical and in Computer Chemistry*[1]

Jiaojiao Yang, Haixing Zhao,[2] Jun Yin

School of Computer, Qinghai Normal University, Xining, Qinghai 810008, P. R. China

(Received September 10, 2013)

Abstract

Using the database of scientific papers published in *MATCH Communications in Mathematical and in Computer Chemistry* from 2000 to 2012, by combining complex network theory and social network analysis, we construct the network of collaboration between scientists. In the network two scientists are connected if they have coauthored one or more papers together. At the same time, we study a variety of statistical properties of the network, including average path length between scientists, degree distribution, clustering coefficient, betweenness, existence and size of a giant component of connected scientists. The result shows that the scientific collaboration network is a non-connected network and presents the network characteristic of scale–free and small–world.

1 Introduction

Complex networks represent an important area of multidisciplinary research, consisting of sets of vertices or nodes joined together in pairs by edges or links, and appear fre-

[1]Research supported by the National Science Foundation of China (No. 61164005), the National Basic Research Program of China (No. 2010CB334708) and the Program for Changjiang Scholars and Innovative Research Team in Universities (No. IRT1068), and the Nature Science Foundation from Qinghai Province (No. 2012-Z-943).

[2]Corresponding author. Email: h.x.zhao@163.com

quently in various technological, social, and biological scenarios [1,7,9,17,21]. These networks include the Internet [10], the World Wide Web [2,8], social networks [18], scientific collaboration networks [13,16], food webs [24], and protein-protein interaction networks [25]. All in all, complex networks are everywhere. Therefore, the qualitative and quantitative research of complex network has become a major theme of today's science. They have been shown to share global statistical features, such as the "small world" [6,26] and the "scale–free" [4] effects, as well as the "clustering" property.

A social network is a set of people or groups, each of which has connections of some kind to some or all of the others. In the language of social network analysis, people or groups are called "actors" and the connections called "ties". Both actors and ties can be defined in different ways depending on the questions of interest. An actor might be a single person, a team, or a company. A tie might be a friendship between two people, a collaboration or common member between two teams, or a business relationship between companies. Social network analysis [19,22] is an analysis method to investigate the social relationship. The scientific collaboration network is a relationship network formatted in the scientific collaboration practice between the researchers, which is suitable to be investigated using the method social network analysis. In fact, social network analysis has produced many results concerning social influence, social groupings, inequality, disease propagation, communication of information.

Based on the complex network theory and social network analysis, in this paper we make a statistical analysis of the network according to the papers published in *MATCH Communications in Mathematical and in Computer Chemistry* from January 2000 to December 2012, and we study the scientific collaboration network and discuss the influence on the evolving network structure and the characteristic of the network.

2 Relevant Characteristics

In this section, we introduce some relevant characteristics of the network as follows [14, 15].

2.1 Average path length

The average path length is one of the most important statistical characteristics of a network. Let $N = (V(N), E(N))$ be a network with vertex set $V(N)$ and edge set $E(N)$, its average path distance [12] is defined as:

$$L(N) = \frac{1}{|V(N)|(|V(N)| - 1)} \sum_{u,v \in V(N)} d(u, v)$$

where $d(u, v)$ is the distance between vertices u and v in N. The maximum value of the distance between any two vertices is called the diameter of the network, denote by $D = Max\{d(i, j)|i, j \in V(N)\}$.

2.2 Clustering coefficient

The clustering coefficient [5,23] of a network is another parameter used to characterize networks. The clustering coefficient of a vertex was introduced into quantify this concept. Given a network $N = (V(N), E(N))$, for each vertex $v \in V(N)$ with degree δ_v, its clustering coefficient $C(v)$ is defined as the fraction of the $\binom{\delta_v}{2}$ possible edges among the neighbors of v that are present in N. More precisely, if E_v is the number of edges between the δ_v vertices adjacent to vertex v, its clustering coefficient is:

$$C_v = \frac{2E_v}{\delta_v(\delta_v - 1)}$$

whereas the clustering coefficient of N, denoted by $C(N)$, is the average of $C(v)$ over all vertices v of N:

$$C(N) = \frac{1}{|V(N)|} \sum_{v \in V(N)} C_v \ .$$

Obviously, $C \in [0, 1]$, when C=0, N has no any triangle, when C=1, any two vertices in the network are directly connected. In completely random network with N vertices, $C \approx N^{-1}$. However, the empirical result shows that the vertices tend to come together in most of the large scale real network, although clustering coefficient is less than 1, but far greater than N^{-1}.

2.3 Degree distribution and betweenness

The degree distribution [3] is also an important statistical characteristic of a network. By definition, the degree of the vertex v is the total number of edges incident from v. We define $p(k)$ to be the fraction of vertices in the network that have degree k. Equivalently, $p(k)$ is the probability that a vertex chosen uniformly at random has degree k. In the network, the degree distribution function $P(k)$ is equal to the proportion of the number of the vertex which the degree is k in the network. In the present study, there are some common degree distributions: One common forms for the degree distribution are exponentials, such as railway networks [20]. Another distribution is the distribution of power-law distribution, that is $P(k) \propto k^{-r}$, where r is called the degree index. Furthermore, completely random networks obey Poisson distribution and completely regular networks obey Delta distribution. In recent years, a large number of the empirical researches showed that the degree distribution of many of the networks is better described by using power law distribution. The degree index r is generally between two to three, such as in the World Wide Web.

Betweenness [11] is a centrality measure of a vertex within a network. The beteweenness of the vertex v is defined as the total number of shortest paths between pairs of authors that pass through vertex v in network. It shows the influence of the vertex in the network.

2.4 Network entropy

The entropy of graph N, denoted by $I(N)$, is a measure of graph structure. Entropy is a term borrowed from information theory. It is a measure of the "amount of information" or "surprise" communicated by a message. The basic unit of information is the bit, so entropy is the number of bits of "randomness" in a network. The higher the entropy, the more random is the graph.

More formally, the "randomness" in network N is the entropy of its degree sequence distribution g'. The unit of measurement of entropy is a bit, so let $I(N)$ be the expected number of bits in g', as follows:

$$I(N) = -\sum_{i=1}^{max} h_i(log_2(h_i)), \quad \text{where} \quad g' = [h_1, h_2, \cdots, h_{max}] \ .$$

3 Collaboration Network Based on MATCH

3.1 Data collection and arrangement

For this study, we collected all papers published in *MATCH Communications in Mathematical and in Computer Chemistry* from 2000 to 2012 inclusive. There are 1000 papers and 961 authors. According to the title of the papers and the authors, we constructed a collaboration network. In this network, a vertex represents an author, the edges connect different authors if there is coauthor relation between them. In this paper, we do not mark the number of coauthor papers between two authors.

The collaboration network is a self-organizing network. Table 1 shows the database information.

Number of authors in a paper	1	2	3	4	5	6	7	8	9
Number of papers	243	381	244	92	22	9	1	5	3
Proportion	23.4	38.1	24.4	9.2	2.2	0.9	0.1	0.5	0.3

Table 1: The information of the database

In science metrology, coauthor rate is an index to measure a subject of coauthor, which refers to the coauthored papers in proportion to the total number of papers. From Table 1, papers with $1 \sim 3$ authors are 86.8% of all papers.

3.2 Analysis of scientific collaboration network

Analysis shows that the scientific collaboration network is highly non-connected. First, we determined the statistics of hyperdegree and degree distribution of vertices. Next, we analyzed the network's characteristics of the giant component from 2000 to 2012, and the second largest connected component's characteristics.

3.2.1 Degree distribution of vertices

In the network, the number of the edges which connecting a vertex to the network is called the degree of the vertex. A vertex's hyperdegree is the number of hyperedges, that is, the number of the papers a author participated in.

In Tables 2 and 3 we give the degree and the hyperdegree of vertices in the network.

degree of vertices	0	1	2	3	4	5	6	7	8	9	10	11
number of vertices	49	203	238	178	87	62	30	31	27	10	10	9
degree of vertices	12	13	14	15	16	17	19	20	25	27	40	97
number of vertices	3	2	4	4	1	2	5	1	2	1	1	1

Table 2: The degree of the vertices in the collaboration network

hyperdegree	1	2	3	4	5	6	7	8	9	10	11	12	13
number of vertices	568	164	89	39	26	20	9	11	7	6	5	4	1
hyperdegree	15	17	18	23	24	27	29	30	35	37	44	79	
number of vertices	1	1	1	1	1	1	1	1	1	1	1	1	

Table 3: The hyperdegree of the vertices in the collaboration network

Figures 1 and 2 show the degree distribution and the hyperdegree distribution. We find that the degree distribution obeys power distribution $P(k) \propto k^{-r}$, where $r \approx 2.45$. The hyperdegree distribution also obeys power distribution $P(k) \propto k^{-r}$, where $r \approx 2.03$.

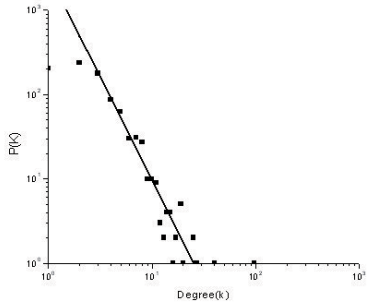

Figure 1: Vertex-degree distribution

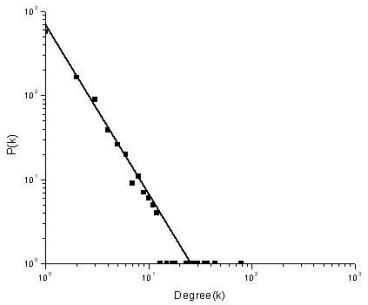

Figure 2: Vertex-hyperdegree distribution

The entropy of the degree and the hyperdegree is 3.110236 and 2.087196, respectively.

3.2.2 Some characteristics of the network

In this section, we analyze some characteristics of the largest component in the collaboration network from 2000 to 2012, including its average path length, diameter of network, and clustering coefficient, the maximum betweenness, given in Table 4.

year	node	$L(N)$	diameter	$C(N)$	maximum betweenness
2000-2001	18	2.765	7	0.613	95.000
2000-2002	26	3.274	6	0.555	181.500
2000-2003	50	3.991	8	0.591	712.000
2000-2004	57	4.002	9	0.577	919.000
2000-2005	88	4.379	10	0.573	2607.167
2000-2006	108	4.558	12	0.585	3483.333
2000-2007	142	4.679	12	0.597	5646.267
2000-2008	201	4.266	11	0.590	73012.699
2000-2009	263	4.678	14	0.599	21793.688
2000-2010	378	4.662	14	0.613	29087.387
2000-2011	408	4.974	16	0.619	49388.363
2000-2012	563	5.687	18	0.597	99128.938

Table 4: Main properties of the largest component

From Table 4, we find that our network has the following characteristics.

(1) With the increasing of the number of vertices in the network, the diameter shows a rising tendency, but the average path length is very small, ranges from 2.765 to 5.687.

(2) The greatest betweenness shows a rising tendency, while the clustering coefficient ranges from 0.555 to 0.619, but it is much more than N^{-1}.

In brief, the network has obvious small world network characteristics, that is smaller average path length and larger clustering coefficient.

3.2.3 Largest component

The scientific collaboration network based on papers from 2000 to 2012 is a non-connected network. In Table 5, we give the number of all its components and their

vertex number.

# vertices	1	2	3	4	5	6	7	9	10	11	12	13	20	563
# components	49	36	26	13	3	6	3	1	1	1	1	1	1	1

Table 5: The number of connect components of the network and their vertex number

The statistics shows that more than 5.09 % of the authors are independent, maybe due to the author's work background or region caused and so on. However, there is also a super research team, the number of the team is 563, and a greater team's number is 20, these big research teams play an important role spreading new knowledge and new thought.

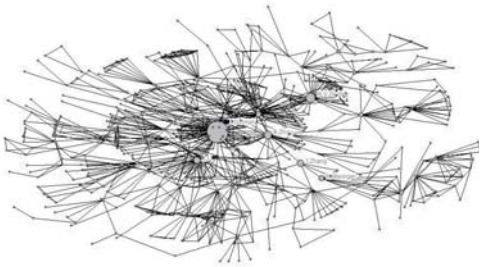

Figure 3: The giant component

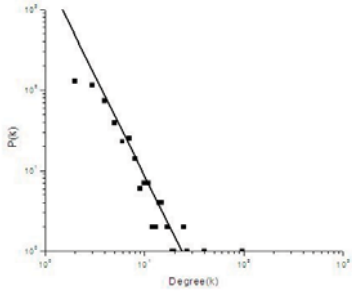

Figure 4: Vertex-degree distribution

The giant component is a connected component with $\delta \times N$ vertices at least, where δ is a positive constant. Figure 3 is the giant component, it contains 563 vertices with $\delta > \frac{1}{2}$, entropy 3.127, average path length 5.687, clustering coefficient 0.597, which is about 336 times than N^{-1}. The result shows the network has obvious clustering features (small world characteristics), the hub vertex's degree (the largest degree of nodes) is 97 and has the largest betweenness 99128.938, which is Gutman, whose influence is the greatest. Second, the degree of is Diudea 40 with the betweenness 45298.160. Then Li's degree is 27, betweenness 10033.384, Zhou's degree is 25, betweenness 29181.273.

These data imply that the authors with large betweenness, such as Prof. Gutman from University of Kragujevac, Prof. Diudea from Babes–Bolyai University, Prof. Xueliang Li from Nankai University and Prof. Bo Zhou from South China Normal University, are well-known scholars and have an effect to spread new knowledge. From Figure 4, we can find this network's degree distribution which also has the obvious scale-free characteristics, power law index $r \approx 2.49$. Thus, the network possesses the characteristics of a scale-free network.

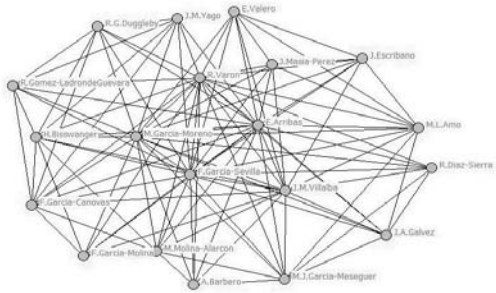

Figure 5: A large component with 20 vertices

Figure 5 is the second largest component with average path length 1.416, clustering coefficient is 0.849, which is about 17 times than N^{-1}. The result shows that the network has obvious clustering features (small world characteristics), the hub vertex's degree of is 19 (R. Varon, F. Garcia–Sevilla, E. Arribas, M. Garcia–Moreno), the largest betweenness is 16.043 (R. Varon, F. Garcia–Sevilla, E. Arribas, M. Garcia–Moreno). In this sub-network, researchers are concentrated comparatively, contacted closely, which plays an important role in spreadin knowledge.

4 Summary

In this paper, we established the network based on the papers published in *MATCH Communications in Mathematical and in Computer Chemistry* from January 2000 to December 2012. By using the tool *SATI* and *UCINET*, we found that the scientific collaboration network is a non-connected network. We established a large number of

statistics for our network, including the average path length, clustering coefficient, degree distribution and betweenness. We note that the distance between pairs of authors in each connected component is very small and the network is "small-world" and "scale-free". Besides, the distributions of these quantities roughly follow a power-law form, although there are some deviations which may be due to the finite time window used for the study.

Acknowledgements. The authors are greatly indebted to Professor Gutman and the referees for their valuable comments and suggestions, which were very helpful for improving the presentation of the paper.

References

[1] R. Albert, A. L. Barabási, Statistical mechanics of complex networks, *Rev. Modern Phys.* **74** (2002) 47–97.

[2] R. Albert, H. Jeong, A. L. Barabási, Diameter of the world–wide web, *Science* **401** (1999) 130–131.

[3] F. M. Atay, T. Biyikoglu, J. Jost, Synchoronization of networks with prescribed degree distributions, *IEEE Trans. Circuits* **53** (2006) 92–98.

[4] A. L. Barabási, E. Bonabeau, Scale-free networks, *Sci. Amer.* **288** (2003) 60-69.

[5] A. Barrat, M.Weigt, On the properties of small-world networks, *Europ. Phys. J. B* **13** (2000) 547–560.

[6] M. Barthelemy, L. A. N. Amaral, Small-world networks: Evidence for a crossover picture, *Phys. Rev. Lett.* **82** (1999) 3180–3183.

[7] S. Boccaletti, V. Latora, Y. Moreno, Complex network: Structure and dynaimcs, *Phys. Rep.* **424** (2006) 175–308.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, *J. Compt. Networks* **33** (2000) 309–320.

[9] S. N. Dorogovstev, J. F. F. Mendes, Evolving of networks, *Adv. Phys.* **51** (2002) 1079–1187.

[10] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, *Comp. Commum. Rev.* **29** (1999) 251–262.

[11] L. C. Freeman, A set of measure of centrality based on betweenness, *Sociometry* **40** (1977) 35–41.

[12] A. Fronczak, P. Fronczak, J. A. Holyst, Average path length in random networks, *Phys. Rev. E* **70** (2004) 056110.

[13] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, The web of human sexual contacts, *Nature* **411** (2001) 907–907.

[14] M. E. J. Newman, Scientific collaboration networks. I. Network construction and fundamental results, *Phys. Rev. E* **64** (2001) 016131.

[15] M. E. J. Newman, Scientific Collaboration Networks. II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E* **64** (2001) 016132.

[16] M. E. J. Newman, The structure of scientic collaboration networks, *Proc. Natl. Acad. Sci. U.S.A.* **98** (2001) 404–409.

[17] M. E. J. Newman, The structure and function of complex networks, *SIAM Review* **45** (2003) 167–256.

[18] M. E. J. Newman, D. J. Watts, S. H. Strogatz, Random graph models of social networks, *Proc. Natl. Acad. Sci. U.S.A.* **99** (2002) 2566–2572.

[19] J. G. Scott, *Social Network Analysis – A Handbook*, Sage, London, 2000.

[20] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, S. S. Manna, Small–world properties of the Indian railway network, *Phys. Rev. E* **67** (2003) 036106.

[21] S. H. Strogatz, Exploring complex networks, *Nature* **410** (2001) 268–276.

[22] S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge Univ. Press, Cambridge, 1994.

[23] D. J. Watts, S. H. Strogatz, Collective dynamics of "small world" networks, *Nature* **293** (1998) 440–442.

[24] R. J. Williams, N. D. Martinez, Simple rules yield complex food webs, *Nature* **404** (2000) 180–182.

[25] S. Wuchty, Scale–free behavior in protein domain networks, *Mol. Biol. Evol* **18** (2001) 1694–1702.

[26] X. S. Yang, Small–world networks in geophysics, *Geophys. Res. Lett.* **28** (2001) 2549–2552.