Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 74, No. 3 (2008) 471–482
DOI: 10.1007/s11192-007-1587-1

Locating active actors in the scientific collaboration communities based on interaction topology analyses

YICHUAN JIANG[a,b]

[a] *School of Automation, Southeast University, Nanjing (P. R. China)*
[b] *Department of Social Informatics, Kyoto University, Kyoto (Japan)*

While implementing a large-scale research project, it is necessary to appoint some principle scientists, and let each principle scientist lead a research group. In a scientific collaboration community, different scientists perform different roles while they implement the project, and some scientists may be more active than others; these active scientists often undertake the role of leadership or key coordinator in the project. Obviously, we should assign the role of principle scientists onto those active actors in the communities. In this paper, we present the model and algorithms for locating active actors in the community based on the analyses of scientists' interaction topology, the actors with high connection degrees in the interaction topology can be considered as active ones. Finally, we make some case studies for our model and algorithms.

Introduction

Nowadays, some research projects are always very large and can't be performed by single scientist; in the research community, the scientists always need to collaborate with each other for a common research topic. Moreover, some projects may even be performed through international scientific cooperation [1]. Therefore, the scientists often build up some collaboration communities according to their research topics [2]. In the community, the scientists can be also divided into some research groups; and a research group is considered to be a system and the scientists in this system [3].

Received April 27, 2007

Address for correspondence:
YICHUAN JIANG
School of Automation, Southeast University, Si Pai Lou 2#, Nanjing 210096, P. R. China
E-mail: jiangyichuan@yahoo.com.cn

0138–9130/US $ 20.00

Scientific Collaboration Community (SCC) is an academic interest group, in which there are many scientists or labs who collaborate for a research topic [4, 5]. In a scientific collaboration community, cooperation will enable scientists to solve the problems that cannot be solved by individual one. To implement cooperation among agents, there is a significant demand for scientists to interact with each other effectively. A scientist can profit from the research of other scientists, as well as benefit to other scientists. Moreover, collaboration can save the research resource and reduce the redundant research endeavors.

Indeed, collaboration is often considered as one of the key concepts in current scientific research communities. When some academic interest group wants to perform a large project, the first step is to allocate the project to some scientists, which is called *project assignment*. The main problem in the project assignment is that no scientist is versatile, and each scientist has different capabilities and can only fulfill a subset of the projects. Therefore, we should make an effective project assignment, i.e., get effective project-scientists mapping strategy. The main goal of the project assignment is to maximize the overall research result of the project as well as save the research costs.

In a scientific collaboration community, different scientists perform different roles in the collaboration, and some scientists may be more active than others. These active scientists often undertake the role of leadership or coordinator in the project. Obviously, the key issue in the research project assignment is the location of principle scientists (i.e., active actors in the community).

The ability to detect the active actors in the scientific collaboration community could clearly have practical applications. Being able to locating active actors in these communities could help us to understand the key tach in the communications among scientists. For example, in the Major State Basic Research Development Program of China (973 Program), one project is always assigned neither to a single scientist nor all the scientists in the scientific community [6]. However, in the ever assignment of Chinese 973 Program, the projects were sometimes assigned to the inactive scientists, which influenced the progress efficiency of the project. I think that we should assign the project to the active scientists in the community, and each of those active scientists can organize a team (sub-group) to undertake one subject of the project. By doing this, the research project can be performed well.

Then how to locate the active actors in the scientific collaboration community? Obviously, the activeness of a scientist is expressed by the interactions between him and other scientists in the community. The more frequently he interacts with other scientists, the more active he is in the community. There are many types of interactions among scientists, such as research collaboration, project negotiation, paper co-authorship, paper citation, interests discussing, and so on. In this paper, we mainly consider the project collaboration interactions among scientists in the community.

The implementation of a project often results from the interactions occurred at many levels and across various sectors, and the interactions among research scientists form the network topology [8]. The analysis of social interaction network is a method to explore the social phenomenon [11, 12]. In this paper, by analyzing the interaction topology, we can select the actors that are the most "between" entities in the science collaboration community. The scientists with the highest "betweenness" can be considered as the active ones in the community and can be assigned as the principle scientists in a large project.

Interaction topology among scientists

In the scientific communities, the scientists interact with each other by different types of relations, such as project cooperation, paper co-authorship or citation, research meeting, and so on. For simplicity, in this paper we only consider the project cooperation relation between scientists.

To describe the project cooperation relations among scientists in the communities, we present the concept of *Scientist Cooperation Relation Graph* (SCRG).

Definition 1. *Scientist Cooperation Relation Graph* (SCRG): SCRG = (*V,E*), where:

– $V = \{v_1, v_2, \cdots, v_n\}$, where v_i denotes scientist i;

– $E \subseteq V \times V$; $E = \{e_1, e_2, \cdots, e_n\}$, where $e_i = (v_u, v_v)$ denotes the project cooperation relation between scientist v_u and scientist v_v.

From the example of SCRG in Figure 1, we can see that s_1 cooperates with s_2, s_3, and s_4; s_2 cooperates with s_1, s_3 and s_5, etc.

We can use the *Scientist Cooperation Relation Matrix* to represent the information of cooperation relations in the community.

Definition 2. *Scientist Cooperation Relation Matrix* (SCRM): SCRM = $[r_{ij}]$, where:

$$r_{ij} = \begin{cases} 1) \text{ the weight of the edge, if there is a direct cooperation between } s_i \text{ and } s_j \\ 2)\ 0, \quad \text{if } i = j \\ 3)\ \infty, \text{ otherwise} \end{cases} \tag{1}$$

Therefore, the SCRM representation of the Figure 1 is shown as Figure 2.

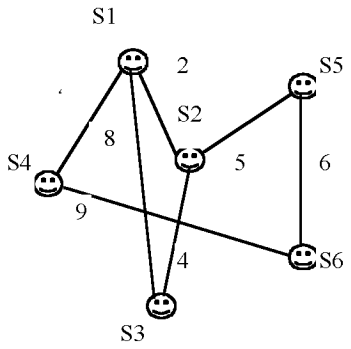


Figure 1. An example of SCRG

$$\begin{bmatrix} 0 & 2 & 8 & 4 & \infty & \infty \\ 2 & 0 & 4 & \infty & 5 & \infty \\ 8 & 4 & 0 & \infty & \infty & \infty \\ 4 & \infty & \infty & 0 & \infty & 9 \\ \infty & 5 & \infty & \infty & 0 & 6 \\ \infty & \infty & \infty & 9 & 6 & 0 \end{bmatrix}$$

Figure 2. SCRM of the example in Figure 1

In the collaboration of scientists, they make cooperation according to the SCRG. The cooperation will cost some resources, which can be called as *cooperation cost.* Cooperation cost between two scientists denotes that the cost they will spend during their cooperation for the research activity. The cooperation cost often includes the one of cooperative development, the one of resource negotiation, the one of travel, the one of communication, and so on. For simplicity, in this paper, we only consider the cooperation cost as an abstract concept.

In Figure 1, we can see that the cooperation cost between s_1 and s_2 is 2, the one between s_2 and s_3 is 4, the one between s_1 and s_3 is 8. Therefore, if s_1 wants to make a cooperative research with s_3, it will be more economical that s_1 cooperates with s_2 firstly

and then s_2 cooperates with s_3. Therefore, in the scientific collaboration community, if two scientists want to make cooperative research, they should find *a shortest cost path* between them, and the cooperative scientists will make cooperation and communication according to such cooperation path.

We can use the concept of cooperation cost matrix to denote the cooperation cost among scientists in the community.

Definition 3. *Scientist Cooperation Cost Matrix* (SCCM): SCCM = $[c_{ij}]$, where c_{ij} is the total cooperative cost of the shortest cost path from scientist i to j.

Therefore, the SCCM of the example in Figure 1 is shown as Figure 3.

$$\begin{bmatrix} 0 & 2 & 6 & 4 & 7 & 13 \\ 2 & 0 & 4 & 6 & 5 & 11 \\ 6 & 4 & 0 & 10 & 9 & 15 \\ 4 & 6 & 10 & 0 & 11 & 9 \\ 7 & 5 & 9 & 11 & 0 & 6 \\ 13 & 11 & 15 & 9 & 6 & 0 \end{bmatrix}$$

Figure 3. SCCM of the example in Figure 1

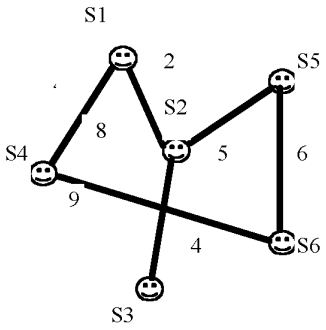
In [9] we have used the agent communication topology graph to denote the set of agent communication paths. By referring it, now we provide the concept of scientist interaction topology graph to denote the set of scientist cooperation paths in the collaboration community.

Definition 4. *Scientist Interaction Topology Graph* (SITG): SITG denotes the topology graph that is composed of scientist cooperation paths in scientific collaboration community. Since the scientist communication often goes along the shortest cost path, SITG is composed of the shortest cost paths between scientists.

SITG = (V", E"), where:

- if the S_i cooperates with S_j, then S_i, $S_j \in V$";
- if the scientist S_i cooperates with S_j, then the edges along the shortest cost path between S_i and S_j are attributed to E".

We can compute the SITG of the example in Figure 1, shown as Figure 4. From Figure 4, we can see that all of the cooperation paths are shown in the SITG.



a)

$$\begin{bmatrix} 0 & 1 & \infty & 1 & \infty & \infty \\ 1 & 0 & 1 & \infty & 1 & \infty \\ \infty & 1 & 0 & \infty & \infty & \infty \\ 1 & \infty & \infty & 0 & \infty & 1 \\ \infty & 1 & \infty & \infty & 0 & 1 \\ \infty & \infty & \infty & 1 & 1 & 0 \end{bmatrix}$$

b)

Figure 4. The SITG and the matrix representation of the example in Figure 1

The SITG can be represented by a matrix, which can be defined as:

Definition 4. Matrix representation of SITG is defined as: $W[w_{ij}]$, where:

$$w_{ij} = \begin{cases} 1)\ 1, \text{ if there is a direct link between } s_i \text{ and } s_j \\ 2)\ 0, \text{ if } i = j \\ 3)\ \infty, \text{ otherwise} \end{cases} \quad (2)$$

Therefore, we can give the matrix representation of the SITG in Figure 4 (a), shown as Figure 4 (b).

Locating active actors

Now, we will address how to locate the active actors in the community; after the active actors are located, they can be assigned as principle scientists. In the scientist collaboration community, the active actors are the scientists that are most "between" among the cooperations in the science collaboration community. Therefore, we should select the vertexes with high "*betweenness*" in the SITG as the active actors.

First proposed by Freeman [10], *Vertex betweenness* has been studied in the past as a measure of the centrality and influence of nodes in networks, and the *betweenness* centrality of a vertex *i* is defined as the number of shortest paths that pass through it. It is a measure of the influence of a node over the flow of information between other nodes [7], which was introduced to investigate the participation rate of any node in the network.

Obviously, from the description of the concept of SITG, we can simply select the vertexes with the high connection degrees in the SITG as active actors. The vertexes with high degrees in the SITG are the most *betweenness* within the cooperation flow in the scientific collaboration community.

Now we can design the algorithms for computing the SITG and locating the active actors (i.e., the principle scientists in a project). The model includes three steps:

1) Compute the shortest cost paths between different scientists in the community;

2) Combine those shortest cost paths to form the SITG;

3) Select the vertexes with high connection degrees in the SITG, and then we can locate those vertexes as the active actors, i.e. principle scientists.

We can appoint the active actors in the community as the principle scientists for a project, and each scientist can take charge of a sub-project.

The algorithms for the 1) 2) and 3) in our model are shown as Algorithm 1, Algorithm 2, and Algorithm 3.

Algorithm 1. Compute the shortest cost path between all scientists in the community.

```
/*let Scientist cooperation relation matrix (SCRM) is [r_ij], c_ij denotes
the total cooperative costs of the shortest cost path between s_i and s_j,
path_ij denotes the shortest cost path between s_i and s_j. */
  1) for (int i=1; i<=n; i++)
        for (int j=1; j<=n; j++)
           { c_ij = r_ij;
            if (c_ij <max)
            path_ij = [i] + [j];}
  2) for (int k=1; k<=n; k++)
        for (int i=1; i<=n; i++)
              for (int j=1; j<=n; j++)
                   if (c_ik + c_kj < c_ij)
                      { c_ij = c_ik + c_kj;
                        path_ij = path_ik + path_kj;}
```

Algorithm 2. Computing the SITG.

```
1) SITG = {};
2) Compute the shortest cost path between all scientists in the
community;
 3) for (int i=1; i<=n; i++)
     for (int j=1; j<=n; j++)
             SITG = SITG + path_ij.
```

Let SITG is represented by a matrix $[w_{ij}]$, then we can use Algorithm 3 to select the m highest active actors as principle scientists.

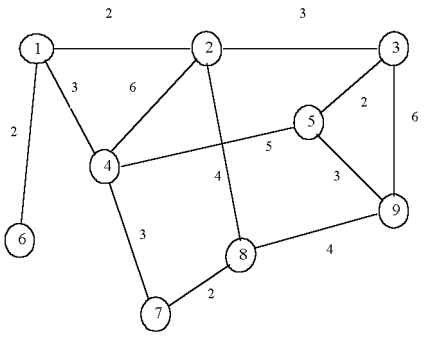
Algorithm 3. Locating the active scientists in the SITG.

```
1) Scientists={1, 2, …, n}; a=0;
2) for (int i=1; i<=n; i++)
         int C[i]=0;
3) for (int i=1; i<=n; i++)
       for (int j=1; j<=n; j++)
           if  w_ij = = 1 then C[i]++;
4) for (int x=1; x<=m; x++)
       {result={0}; result-degree=0;
        for (int i=1; i<=n; i++)
           if C[i]>result-degree
              {result={i}; a=i; result-degree=C[i];}
        C[a]=0;
        Scientists=Scientists-result;
        Output (result);}
```

Case studies and demonstrations

Now we give some cases and demonstrate how our model can work in those cases. Let Figure 5 (a) is a scientist collaboration community, its *Scientist Cooperation Relations Matrix* (SCRM) representation is shown as Figure 5 (b). Now according to Algorithms 1 and 2, we can compute the shortest cost paths between different scientists, and combine the shortest cost paths to form the *Scientist Interaction Topology Graph* (SITG), shown as Figure 5 (c). The matrix representation of SITG is shown as Figure 5 (d).

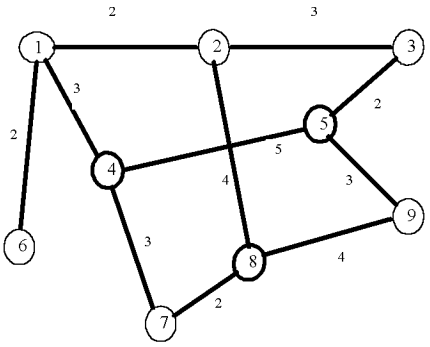
Now, if there is a large project P which can be evenly divided into three sub-projects: P_1, P_2 and P_3. In the assignment of the three sub-projects, at first we should locate three principle scientists. According to Algorithm 3, the active actors with high connection degree in Figure 5 (d) should be selected as principle scientists. Therefore, S_4, S_5 and S_8 can be assigned as the principle scientists, and other scientists can select the principle scientist with the minimum collaboration cost to form a sub-group.



a)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
s_1	0	2	∞	3	∞	2	∞	∞	∞
s_2	2	0	3	6	5	∞	∞	4	∞
s_3	∞	3	0	∞	2	∞	∞	∞	6
s_4	3	6	∞	0	5	∞	3	∞	∞
s_5	∞	∞	2	5	0	∞	∞	∞	3
s_6	2	∞	∞	∞	∞	0	∞	∞	∞
s_7	∞	∞	∞	3	∞	∞	0	2	∞
s_8	∞	4	∞	∞	∞	∞	2	0	4
s_9	∞	∞	6	∞	3	∞	∞	4	0

b)



c)

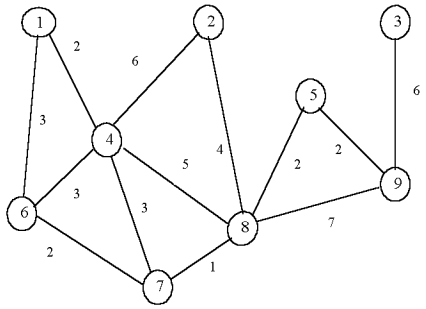
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	degree
s_1	0	1	∞	1	∞	1	∞	∞	∞	3
s_2	1	0	1	∞	∞	∞	∞	1	∞	3
s_3	∞	1	0	∞	1	∞	∞	∞	∞	2
s_4	1	∞	∞	0	1	∞	1	∞	∞	3
s_5	∞	∞	1	1	0	∞	∞	∞	1	3
s_6	1	∞	∞	∞	∞	0	∞	∞	∞	1
s_7	∞	∞	∞	1	∞	∞	0	1	∞	2
s_8	∞	1	∞	∞	∞	∞	1	0	1	3
s_9	∞	∞	∞	∞	1	∞	∞	1	0	2

d)

Figure 5. Case one

At last, the three research groups are: $\{S_4, S_1, S_6\}$, $\{S_5, S_3, S_9\}$, $\{S_8, S_2, S_7\}$. The total collaboration costs are made up of two parts: the collaboration cost within the sub-group and the one among the principle scientists. The assignment scheme and the collaboration cost are shown as No. 1 in Table 1.

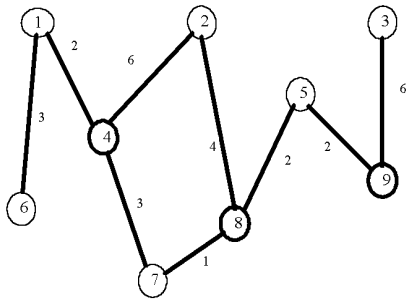
Then we can change the cooperation relations of the 9 scientists, shown as Figure (6) and Figure (7). We also use our model to locate the active actors in the community, and make the project assignments. The ultimate results are shown as No. 2 and No. 3 in Table 1. From the results, we can see that our results can achieve the minimum collaboration cost for the research projects.



a)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
s_1	0	∞	∞	2	∞	3	∞	∞	∞
s_2	∞	0	∞	6	∞	∞	∞	4	∞
s_3	∞	∞	0	∞	∞	∞	∞	∞	6
s_4	2	6	∞	0	∞	3	3	5	∞
s_5	∞	∞	∞	∞	0	∞	∞	2	2
s_6	3	∞	∞	3	∞	0	2	∞	∞
s_7	∞	∞	∞	3	∞	2	0	1	∞
s_8	∞	4	∞	5	2	∞	1	0	7
s_9	∞	∞	6	∞	2	∞	∞	7	0

b)



c)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	degree
s_1	0	∞	∞	1	∞	1	∞	∞	∞	2
s_2	∞	0	∞	1	∞	∞	∞	1	∞	2
s_3	∞	∞	0	∞	∞	∞	∞	∞	1	2
s_4	1	1	∞	0	∞	∞	1	∞	∞	3
s_5	∞	∞	∞	∞	0	∞	∞	1	1	2
s_6	1	∞	∞	∞	∞	0	∞	∞	∞	1
s_7	∞	∞	∞	1	∞	∞	0	1	∞	2
s_8	∞	1	∞	∞	1	∞	1	0	∞	3
s_9	∞	∞	1	∞	1	∞	∞	∞	0	2

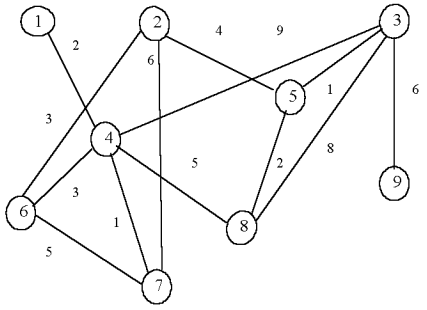
d)

Figure 6. Case two

Table 1. The project assignment schemes for the three cases

No.	Project assignment scheme			Collaboration cost	Total collaboration cost
	Sub-project	Principle scientist	Members		
1	P_1	S_4	S_1, S_6	8	(8+5+6)+(5+5+7)[1] = 36
	P_2	S_5	S_3, S_9	5	
	P_3	S_8	S_2, S_7	6	
2	P_1	S_4	S_1, S_6	7	(7+5+8)+(4+4+8) = 36
	P_2	S_8	S_2, S_7	5	
	P_3	S_9	S_5, S_3	8	
3	P_1	S_4	S_1, S_7	3	(3+3+10)+(6+7+4) = 33
	P_2	S_2	S_6	3	
	P_3	S_5	S_8, S_3, S_9	10	

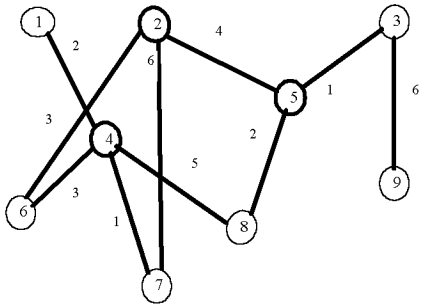
[1](5+5+7) denotes the collaboration cost among the three principle scientists.



a)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
s_1	0	∞	∞	2	∞	∞	∞	∞	∞
s_2	∞	0	∞	∞	4	3	6	∞	∞
s_3	∞	∞	0	9	1	∞	∞	8	6
s_4	2	∞	9	0	∞	3	1	5	∞
s_5	∞	4	1	∞	0	∞	∞	2	∞
s_6	∞	3	∞	3	∞	0	5	∞	∞
s_7	∞	6	∞	1	∞	5	0	∞	∞
s_8	∞	∞	8	5	2	∞	∞	0	∞
s_9	∞	∞	6	∞	∞	∞	∞	∞	0

b)



c)

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	degree
s_1	0	∞	∞	1	∞	∞	∞	∞	∞	1
s_2	∞	0	∞	∞	1	1	1	∞	∞	3
s_3	∞	∞	0	∞	1	∞	∞	∞	1	2
s_4	1	∞	∞	0	∞	1	1	1	∞	4
s_5	∞	1	1	∞	0	∞	∞	1	∞	3
s_6	∞	1	∞	1	∞	0	∞	∞	∞	2
s_7	∞	1	∞	1	∞	∞	0	∞	∞	2
s_8	∞	∞	∞	1	1	∞	∞	0	∞	2
s_9	∞	∞	1	∞	∞	∞	∞	∞	0	1

d)

Figure 7. Case three

Conclusion

In this paper we have presented a new method for large scale research project assignment. At first we make analyses to the interaction topology of scientists in the scientific collaboration community, the shortest cost paths between all scientists form the *Scientist Interaction Topology Graph* (SITG). From the SITG, we can locate the actors with high degrees as the principle scientists. At last, the large scale research project can be divided into some sub-projects, and we can assign the sub-projects to these principle scientists and let them take charge of the sub-groups.

Nowadays, the research project allocation is always qualitative, our model can make such thing be quantitative and measurable. Moreover, with our model, we can make the collaboration cost of the research project be reduced.

However, the interaction relation of scientists is simple in this paper which only considers the cooperation cost and the cooperation cost is also only an abstract concept. In our future works, we will investigate more kinds of interaction relations of scientists, and make the model be more comprehensive and be well suited into practical situations. Moreover, optimal cooperation may take place between two (or more) scientists who individually do not necessarily have the best set of skills/technical expertise simply because there exists a high level of trust/familiarity between them. Therefore, in our future work, we also will consider such exceptional situation in the project allocations.

References

1. LECLERC, M., GAGNÉ, J., International scientific cooperation: The continentalization of science, *Scientometrics*, 31 (3) (1994) : 261–292.
2. MELIN, G., PERSSON, O., Studying research collaboration using co-authorships, *Scientometrics*, 36 (1996) : 363–377
3. KRETSCHMER, H., Cooperation structure, group size and productivity in research groups, *Scientometrics*, 7 (1-2) (1985) : 39–53.
4. NEWMAN, M. E. J., The structure of scientific collaboration networks, *PNAS*, 98 (2) (2001) : 404–409.
5. NEWMAN, M. E. J., Coauthorship networks and patterns of scientific collaboration, *PNAS*, (101) (1) (2004) : 5200–5205.
6. *Profile of 973 Program*. http://www.973.gov.cn/English/Index.aspx
7. GIRVAN, M., NEWMAN, M. E. J., Community structure in social and biological networks, *Proc. Natl. Acad. Sci*, 99 (2002) : 8271–8276.
8. WAGNER, C. S., L. LEYDESDORFF, Mapping the network of global science: comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, 1 (2) (2005) : 185–208
9. JIANG, Y. C., XIA, Z. Y., ZHANG, S. Y., An adaptive adjusting mechanism for agents distributed blackboard architecture, *Microprocessors and Microsystems*, (Elsevier Science), 29 (1) (2005) : 9–20.
10. FREEMAN, L., A set of measures of centrality based upon betweenness. *Sociometry*, 40 (1977) : 35–41.
11. GRANOVETTER, M. S., The strength of weak ties, *The American Journal of Sociology*, 78 (6) (May, 1973) : 1360–1380.
12. BURT, R. S., *Structural Holes:The Social Structure of Competition*. Cambridge, Massachusetts, Harvard University Press,1992.