# Mining the Potential Collaborative Relationships Based on the Author Keyword Coupling Analysis and Social Network Analysis

Peng Yufang[1] Gu Dongxiao[2] Shi Jin[1]

[1] *laisitianshi@163.com*
School of Information Management, Nanjing University, Nanjing (China)

[2]*dongxiaogu@yeah.net*
School of management, HeFei University of Technology, Anhui (China)

[3]*shijin@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing (China)

## Abstract

This study aims at explore and discover the potential collaborative relationship for big data privacy and security literature. We collected data from Web of Science and EI - Engineering Village 2 and analysed the data by using social network analysis, author keyword coupling analysis, TF-IDF, co-word analysis and cluster analysis. We find 1,380 collaborative articles from the total number of 1,645 papers. However, 90.92% of all authors published one paper, so it also means that it is valuable to research the potential relationships. The integrity of the entire network is relatively small, interpersonal communication is not too close, but it has a significant potential space for collaboration. Finally, we find many authors have the potential collaborative relationships.

## Conference Topic

Social network analysis; Knowledge discovery and data mining; Co-occurrence analysis; Author keyword coupling analysis; Potential collaboration relationships; big data privacy and security

## Introduction

Bertalanffy, L. von (1951) proposed the General system theory, it regards the organization as a kind of system, which refers to the existence of the system elements of the inflow and outflow, and thus its components change and replacement, that system and the outside world has energy and material exchange. Pfeffer, J., and Salancik, G. (1978) proposed Resource Dependence theory, pointed out that any organization and individual in order to survive on the need to extract resources from the surrounding environment, the need to interdependence with the surrounding environment, the interaction can achieve the goal. Fan Zhiying was based on the above theories to explain the necessity of scientific research cooperation (Fan, 2015). In addition, Nowak, MA also (2006) believed that “Humans are the champions of cooperation: From hunter-gatherer societies to nation-states, cooperation is the decisive organizing principle of human society.” Capozzalo, G et al. see collaboration as a source of strength (Capozzalo, 1991). Therefore, it is valuable and important to study the cooperation of researchers.

In recent years, many researchers have made some achievements in the area of collaborative research. For example, collaboration the context of co-location (Wener, Woodgate, 2016), influence the collaborative factors (Verdecho, Alfaro-Saiz, Rodríguez-Rodríguez, 2011), collaborative demand forecasting (Dong, Huang, Sinha, Xu, 2014), collaborative relationships in general practice projects (Walker, Adam, 1998), partners selection for enterprise-university-institute cooperation (Bi, 2008) etc. Above all, most articles paid more attention to discuss collaborative relationships in some specific practice. But very few papers about mining potential partnerships, for example, explore the potential cooperative relationship (Sun, Hou, 2014; Chen, Zheng, 2013).

This study, social network analysis and author keyword coupling analysis were used to analyse and mining the potential collaborative relationships. “Big data privacy and security” was taken as a case. Because, big data privacy and security is very hot topic, and very few documents are about cooperation research, especially mining its potential collaborative relationships. The primary purposes of our study are two-folds: *(i)* Learn about the collaborative status of big data privacy and security field. *(ii)* Mining the potential collaborative relationships, it possibly to promote their cooperation.

This paper tackles this line of research, and it is structured as follows: The next Section presents analytical approaches, data collection and data processing. Then collaborative status of big data privacy and security field is analysed by social network analysis. After that, the potential collaborative relationships are mined by author keyword coupling analysis. Finally, the main conclusions are highlighted.

## Methods

Social network analysis has its roots in the work of Kurt Lewin (1936). An author collaboration network is an important particular type of social networks and has been extensively applied to determine the structure of scientific collaborations and the status of individual authors (Garfield, E., 1979). Therefore, social network analysis method was adopted to reveal the collaborative status in big data privacy and security. Social network analysis (SNA) measures aspects of social and community relationships to understand relationships between people. It is made up of nodes of individuals, groups, organizations, etc. and the tie in one or more types of interdependencies, which included kinship, social contacts, shared visions, etc., among numerous other aspects of human relationships (Stanley & Faust, 1994). Freeman defined social network analysis as an organized paradigm for research (Freeman, 2004). Ucinet (Borgatti, Everett & Freeman, 2002) wsa used to uncover the author collaboration network and Netdraw (Borgatti, 2002) was used to visualize it, including the following aspects: number of nodes, the number of ties, average degree, network density, average distance, degree centrality, betweenneess centrality, closeness centrality, etc.

Author keyword coupling analysis is a method that analyzes the relationships between the authors by using the coupling strength of the authors-keywords (Liu, Zheng, 2011). The specific steps include: ① Summarize the keyword set that appears in all of the published papers of all authors, K = { k1, k2, k3, k4,……, kn } (N is the total number of keywords) and the frequency of each keyword. ② Respectively, we Statistics the keyword set of each author from their published papers. Ki = { ki1, ki2, ki3, ki4,……, kim } ( M is the author i has the total number of keywords ) and the frequency of each keyword. ③ Calculate the weight of each keyword by using TF-IDF (Christopher, Prabhakar, Hinrich, 2010) ④ Identify the keywords that are shared by each of the two authors and calculate the similarity of the author's research content according to author-keyword coupling strength. ⑤ Sort the similarity in descending order, according to the similarity of the size of the excavation of the potential cooperation between the authors.

## Data Collection and Data Processing

Data for this study was collected from Web of Science and EI - Engineering Village 2, the former is a comprehensive citation index database, the latter offers access to 12 engineering literature and patent databases providing coverage from a wide range of trusted engineering sources. From EI - Engineering Village 2, the search string was "(((“big data”) WN KY) AND ((security* or privacy*) WN KY)), the search time was 2007-2016.6.3, PM 5:07, we obtained 1,557 papers. From Web of Science, we collected data from Web of Science Core Collection the search string was “Topic: ((“big data”) and (security* or privacy*)),” the search time was 2007-2016.6.3, PM 5:27, we obtained 662 papers. We output these two sets of data, through excel, we mixed data of two databases together, we received 2,219 articles. However, there are some duplicate data, author’s name and author’s address is vacant, we removed these data. In the end, we obtained 1,645 papers.

Name disambiguation. C# program was made by us to differentiate the same name. The concrete steps are as follows: (1) Judge by collaborators: set the minimum comparable parameters, where the number of the collaborators is at least 1. Then we can rule comparison: ① compared to the name of the partner, if the number of partners and the name of the collaborators are the same, or if the collaborators of the other party is a subset of the other party, the judgment is the same author. ② use the address of the collaborators and the number of duplicate partners, to a certain percentage, (such as the number of different standards do not have the same), such as three there are two repeated, five there is three repeat, and so determine the same author. (2) Judge by the author's address: ① the address of the normalization (case, punctuation, etc.), if the author's address is the same, they are the same author. ② If an address is the subset of the another one, it may be the same address, and the other is not precise enough. If the two addresses of the front part of the matching degree to meet a certain percentage (according to the number of different numbers of characters), then if they have any one of the same names is the same author. However, they have no collaborators with each other, and we use accurate address analysis: If they have the same zip code, they are the same author; if one has a zip code, another does not have a zip code, when we compare them, we will remove the zip code, etc. Finally, all the same, name was sentenced to different authors, the end of a number to distinguish, for example, name_1, name_2, etc. As a result, we obtained 4,739 authors. And then we continue to observe the accuracy of the data and found that the correct rate is 98%, some author address abbreviation and foreign people in the name of the suffix or more of the other symbols, easy to misjudge the different authors, and then manually clean the data again.

Finally, we obtained the active the number of authors is 4,704, we removed the same authors, and we collected 4,704 authors. And we set up two tables; one contains author’s name, title, keywords, Supplementary Keywords, author’s address, year and so on; the other contains author’s name, author’s address, title, the number of published papers, the number of collaborative times and so on. As a result, we finally obtained the number of 1,645 articles from 4,704 authors.

## Results and Discussion

### Collaborative papers analysis

Fig. 1 shows the number of papers and the number of collaborative authors. By statistics, big data privacy and security research started from 2009. There is only one published paper in 2009 and did not produce a cooperative relationship. The real collaborative relationships started from 2010. Fig. 1, different colors represent the relevant data for various years. The collaborative status became more frequently from 2010 (Note: 2016 data only collected in the first half year), especially since 2013, the cooperative situation was in the rapid growth. On the one hand, researchers began to pay more attention to the big data privacy and security since 2013. On the other hand, the rapid development of big data leads to the more threats of the privacy and security

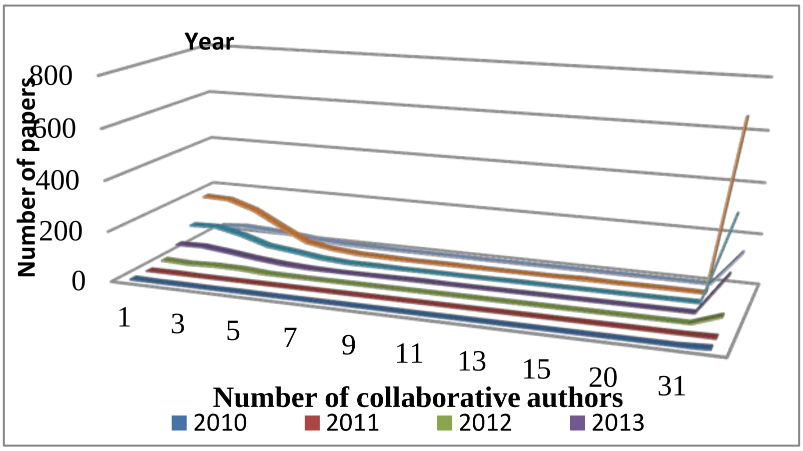
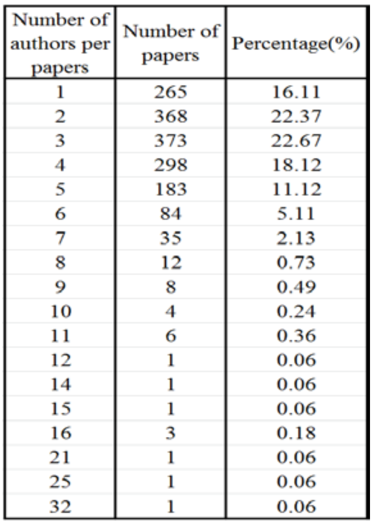


**Fig. 1 Number of papers and Number of collaborative authors**

Table 1 shows the number of authors per paper and number of papers. Through further analyzing the number of collaborative papers, we found 1,645 articles in total, of which 1,380 of them are collaborative documents. To study the collaborative status, we used the following two leading indicators (Liu, Zhang, 2010): *(i)* the author collaboration degree is the rate of the total number of authors and the total number of papers. *(ii)* The author collaboration ratio is the ratio of the total number of collaborative papers and the total number of papers. Therefore, according to our study, in the big data field, the author collaboration degree is 2.86 coauthors per paper, and the author collaboration ratio is 83.89% of collaborative papers, which indicate a relatively high overall level of cooperation in this field. Our study shows that the collaborative documents are in an increasing trend year after yea. Moreover, we found cumulative 4,471 coauthors from the total cumulative number of 4,704 authors (i.e., 95.05% of coauthors).

**Table 1 Number of authors per paper and Number of papers**



| Number of authors per papers | Number of papers | Percentage(%) |
|---|---|---|
| 1 | 265 | 16.11 |
| 2 | 368 | 22.37 |
| 3 | 373 | 22.67 |
| 4 | 298 | 18.12 |
| 5 | 183 | 11.12 |
| 6 | 84 | 5.11 |
| 7 | 35 | 2.13 |
| 8 | 12 | 0.73 |
| 9 | 8 | 0.49 |
| 10 | 4 | 0.24 |
| 11 | 6 | 0.36 |
| 12 | 1 | 0.06 |
| 14 | 1 | 0.06 |
| 15 | 1 | 0.06 |
| 16 | 3 | 0.18 |
| 21 | 1 | 0.06 |
| 25 | 1 | 0.06 |
| 32 | 1 | 0.06 |

Through a further quantitative analysis of all the 1,645 papers, 265 are single-author papers, which mean that 16.11% of articles have no author collaboration. In the remaining 1,380 collaborative documents, the percentages of two-author papers, three-author papers, four-author papers, five-author papers, and six-author papers, and seven-author papers, respectively, 22.37%, 22.67%, 18.12%, 11.12%, 5.11%, 2.13% (Table 1). Through this analysis, one can conclude that the collaboration ratio is very high and remain relatively stable. Above all, big data privacy and security research is from 2009. Since 2013, the collaborative status is in good condition.

## Author collaboration network analysis

### *Measure index of network cohesion analysis*

By statistics, there are 4,277 authors published one article (90.92%), that is, many authors just begin to research in the field. This area is currently not mature enough, and there is more room for development. If we can help them find some authors who study the same area, it will promote his or her research achievements. There are 311 authors published two articles, 77 authors published three papers, etc. To be able to demonstrate the current cooperation status in the big data privacy and security, we selected not less than two authors of the paper for the study, a total of 427 (9.08%).

Fig.3 illustrates author collaboration network extracted from Web of Science and EI - Engineering Village 2 during a period from 2007 to 2016. Firstly, we applied C ++ program to build 427 authors co-occurrence matrix to give 427 * 427 adjacency matrix. Then we squared the resulting matrix import UCINET6, and, finally, Netdraw was applied to visualize the author collaboration network. The nodes are colored based on their components. Degree centrality measures the size of labels of the name. Its Degree centrality measures the size of a node. If one node has a bigger degree, it means that the node has higher degree centrality than others, and it is more important in the network (Wang, Li, Chen, 2012). Wiring thickness is determined by the strength of the co-occurrence, that is, the total number of two nodes is significant if the connection is thicker. As Fig. 3 shows, this network is not overall connected, and it contains some large subnetworks.

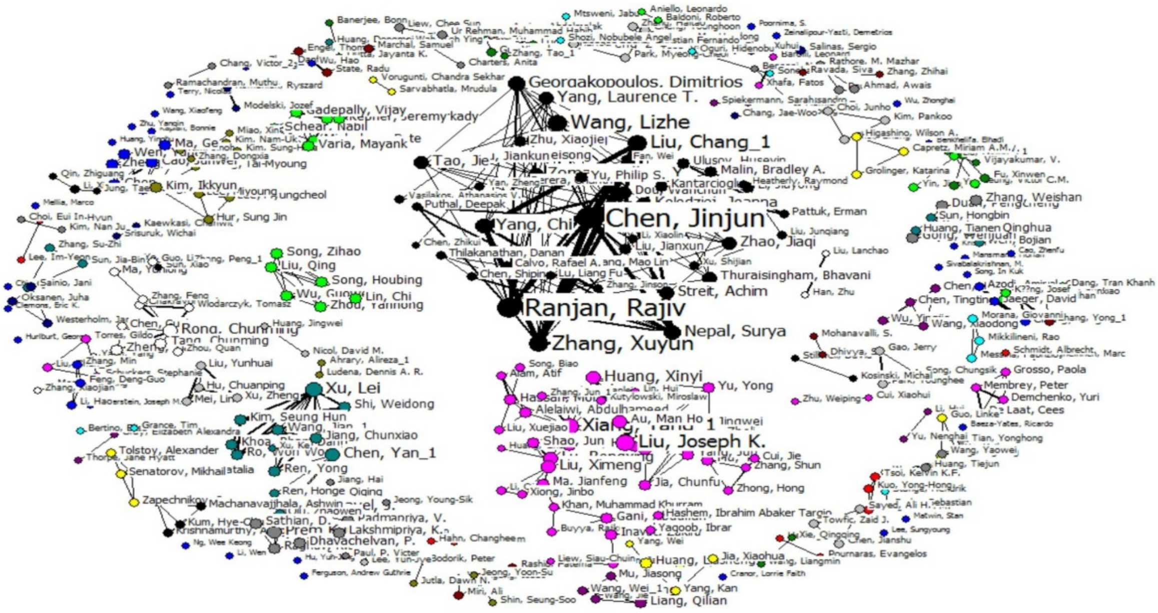


**Fig. 2 Author collaboration network of big data privacy and security from 2007-2016**

In this section, to further know the network cohesion of author collaboration network, we discussed the number of nodes, the number of ties, average degree, network density, average distance, components, connectedness, fragmentation, Avg Distance, etc. (Table2). Table 2 indicates general statistics of author collaboration network. This author collaboration network is composed of 427 nodes (authors) and 1034 ties, which includes 162 connected components. An average degree is an average number of collaborators per author who has the number of direct partners. The average degree of this author collaboration network is 2.422. That means that between the observed periods, a given author has approximately two authors collaborate in a published paper in privacy and security of big data.

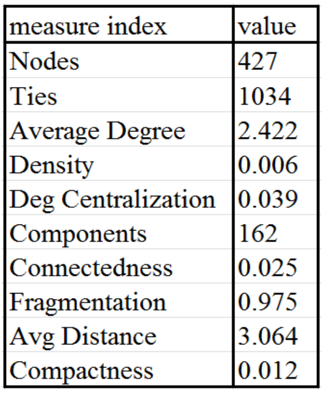
This study analyzes the tightness degree of the overall network, mainly detailed analysis the density, Avg Distance and Degree Centralization. The density of a graph is the proportion of actually existing lines and possible lines in the figure. When the actual number of relationships is closer to the total amount of all possible connections in the network, the density of the network is big and vice versa. Density is used to indicate whether the relationship between the actors is close. The density value is between 0 and 1. If the density value is closer to 1, the whole network will have the greater tightness degree and integration degree. If a network has high density, it will have a good interactive, easy sharing and dissemination of knowledge (Wang, 2008). We observed that the network has a density of 0.006，it is very low. If the group has close relationships, it will have much collaboration. And it is easy to transfer information. However, if the groups' relationship is very alienated, it will often lead to information barrier, a small degree of collaboration, etc. (Chen, Ceng, Xing, 2007). The whole network is not too close to the exchange of information, or interpersonal communication is, and the entire network integrity is little.

Average distance is the average shortest distance between all members of other members in the network. If the average distance is big, it will mean that the nodes of the network have the large span. And it has the lower cohesion. The average distance of this network is 3.064. It means that the network between any two authors needs only through the average for 3.1 authors can communicate with each other, with small world effect (Watts, Strogatz, 1998). In a sense, the network is a smooth flow of information, interpersonal communication quickly, and facilitate the exchange of the existence of the network.

Degree Centralization measures the extent to which the entire network gathers to the center and can be used as an estimate of whether the network relies on a small number of actors. The network's Deg Centralization is 0.039; it is shallow. It shows that there is no apparent network concentration and concentricity in the whole cooperative network. On the contrary, the "eccentricity" is very large, which demonstrates that the collaboration among the authors is very scattered.

Moreover, through Connectedness (0.025), Fragmentation (0.975), Compactness (0.012), all of what show that the network is not a complete network. It means that many authors have a little exchange with each other. However, it also means that enormous potential opportunities for authors to collaborate. It is possible to find the new things when further collaboration.

**Table 2 Measure index of network cohesion**



| measure index | value |
|---|---|
| Nodes | 427 |
| Ties | 1034 |
| Average Degree | 2.422 |
| Density | 0.006 |
| Deg Centralization | 0.039 |
| Components | 162 |
| Connectedness | 0.025 |
| Fragmentation | 0.975 |
| Avg Distance | 3.064 |
| Compactness | 0.012 |

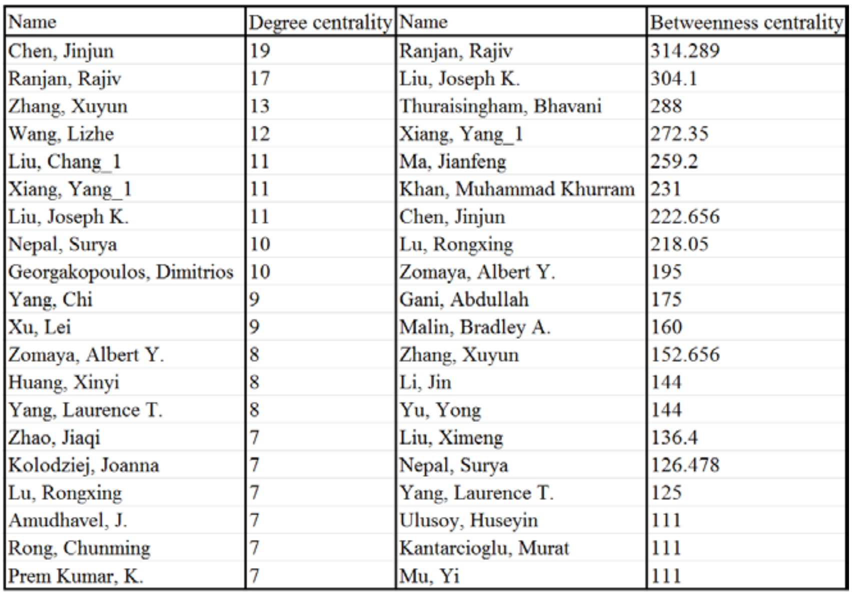
In general, the integrity of the entire network is relatively small, interpersonal communication is not too close, low collaboration degree, but the average distance from the network point of view, the entire network is easy to communicate, there is a lot of room for cooperation.

### *Centrality analysis of the Network*

According to social network analysis, centrality represents the power and status and influence the distribution of authors. If one node has higher centrality degree, it will be at the core. It can control and influence the activities of other actors in the network. On the contrary, if one node has the lower centrality degree, it will be in the marginal position, and rarely participate in interactive communication, and the impact of other nodes is minimal. The most commonly used are degree centrality and betweenness centrality (Freeman, 1977; Freeman, 1979; Freeman, , 2000; Fu, Niu, Wang, et al., 2009). Degree centrality is the ability of an actor to associate with other actors. If an actor is in direct contact with many other actors, the actor is at the center and has a greater "power" in the network. Betweenneess centrality is the extent to which the actor controls the resources. If an actor is on a path between many other two points, the actor can be considered an important place because he or she can control the communication between the other two actors. If one actor takes more positions in the network, it will represent a high betweenneess centrality, that is, more actors need to be able to contact others through him or her.

Table 3 provides the list of authors with the higher degree centrality and betweenness centrality (top 20). We find degree centrality is not correctly matched with betweenness centrality. Some authors have a higher degree centrality, but betweenness centrality is relatively small, and vice versa. For example, Zhang, Xuyun's degree centrality is the highest (3)，but his betweenness centrality is relatively low (152.656); Liu, Joseph K.'s betweenness centrality (304.1), his degree centrality (11), etc. Finally, we find some authors have higher degree centrality and higher betweenness centrality: Chen, Jinjun; Ranjan, Rajiv; Zhang, Xuyun; Wang, Lizhe; Liu, Chang_1; Xiang, Yang_1 et al. (Table 3). They play an important role for information exchange and network collaboration in the network.

**Table 3 Top authors with respect to degree centrality and betweenness centrality**



| Name | Degree centrality | Name | Betweenness centrality |
|---|---|---|---|
| Chen, Jinjun | 19 | Ranjan, Rajiv | 314.289 |
| Ranjan, Rajiv | 17 | Liu, Joseph K. | 304.1 |
| Zhang, Xuyun | 13 | Thuraisingham, Bhavani | 288 |
| Wang, Lizhe | 12 | Xiang, Yang_1 | 272.35 |
| Liu, Chang_1 | 11 | Ma, Jianfeng | 259.2 |
| Xiang, Yang_1 | 11 | Khan, Muhammad Khurram | 231 |
| Liu, Joseph K. | 11 | Chen, Jinjun | 222.656 |
| Nepal, Surya | 10 | Lu, Rongxing | 218.05 |
| Georgakopoulos, Dimitrios | 10 | Zomaya, Albert Y. | 195 |
| Yang, Chi | 9 | Gani, Abdullah | 175 |
| Xu, Lei | 9 | Malin, Bradley A. | 160 |
| Zomaya, Albert Y. | 8 | Zhang, Xuyun | 152.656 |
| Huang, Xinyi | 8 | Li, Jin | 144 |
| Yang, Laurence T. | 8 | Yu, Yong | 144 |
| Zhao, Jiaqi | 7 | Liu, Ximeng | 136.4 |
| Kolodziej, Joanna | 7 | Nepal, Surya | 126.478 |
| Lu, Rongxing | 7 | Yang, Laurence T. | 125 |
| Amudhavel, J. | 7 | Ulusoy, Huseyin | 111 |
| Rong, Chunming | 7 | Kantarcioglu, Murat | 111 |
| Prem Kumar, K. | 7 | Mu, Yi | 111 |

## Mining the Potential Collaborative Relationships

In this section, We still chose the study object that authors who published no less than 2 articles, 427 authors. We used author keyword coupling analysis to mining the potential collaborative relationships, that is, a hidden relationship that researchers have the potential cooperation because of the similarity of their research content but have not yet collaborated with each other.

### *Author-Keyword sets Extraction*

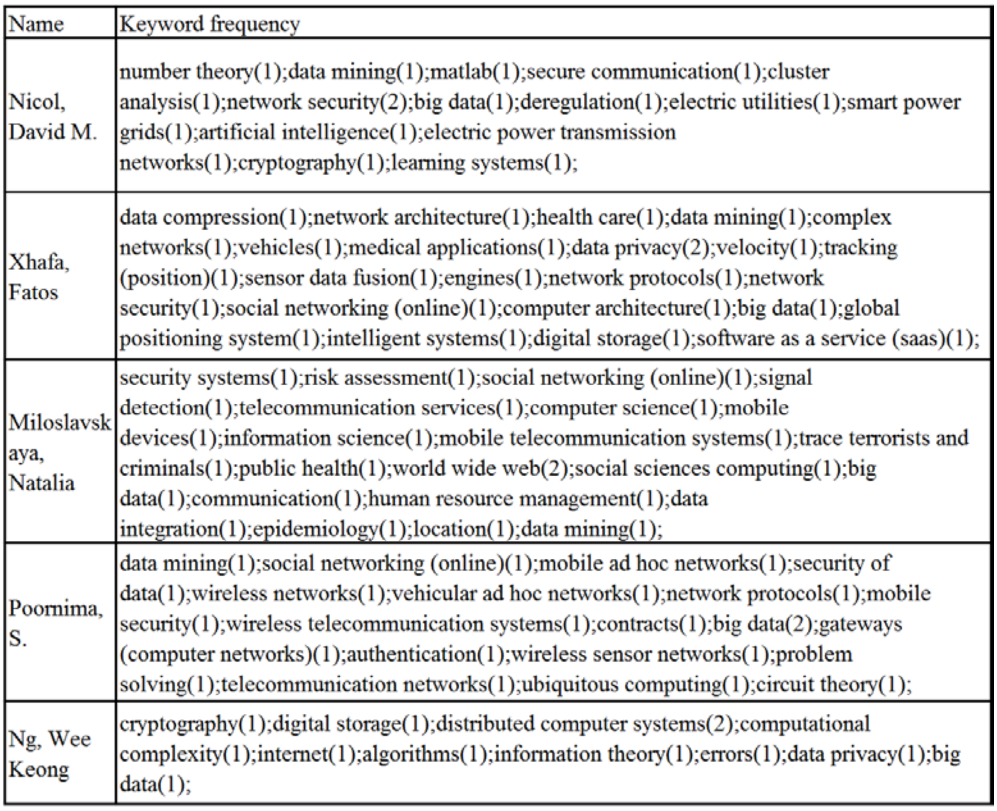
2476 keywords from 1645 papers, and the frequency of each keyword (Table 4).

**Table 4 Number of Keywords and their frequency (partly)**

| keyword | frequency | keyword | frequency |
|---|---|---|---|
| big data | 677 | information management | 121 |
| data privacy | 291 | data mining | 119 |
| cloud computing | 250 | mobile security | 116 |
| digital storage | 209 | security systems | 93 |
| security of data | 148 | distributed computer systems | 87 |
| data handling | 146 | social networking (online) | 87 |
| network security | 145 | algorithms | 84 |
| cryptography | 132 | artificial intelligence | 80 |
| internet | 122 | access control | 64 |

In 1645 papers, we collected 998 keywords from the authors who published no less than two articles by the program and compute the keywords sets of per author, as the Table 5 shows.

**Table 5 Keywords sets of per author (partly)**



| Name | Keyword frequency |
|---|---|
| Nicol, David M. | number theory(1);data mining(1);matlab(1);secure communication(1);cluster analysis(1);network security(2);big data(1);deregulation(1);electric utilities(1);smart power grids(1);artificial intelligence(1);electric power transmission networks(1);cryptography(1);learning systems(1); |
| Xhafa, Fatos | data compression(1);network architecture(1);health care(1);data mining(1);complex networks(1);vehicles(1);medical applications(1);data privacy(2);velocity(1);tracking (position)(1);sensor data fusion(1);engines(1);network protocols(1);network security(1);social networking (online)(1);computer architecture(1);big data(1);global positioning system(1);intelligent systems(1);digital storage(1);software as a service (saas)(1); |
| Miloslavskaya, Natalia | security systems(1);risk assessment(1);social networking (online)(1);signal detection(1);telecommunication services(1);computer science(1);mobile devices(1);information science(1);mobile telecommunication systems(1);trace terrorists and criminals(1);public health(1);world wide web(2);social sciences computing(1);big data(1);communication(1);human resource management(1);data integration(1);epidemiology(1);location(1);data mining(1); |
| Poornima, S. | data mining(1);social networking (online)(1);mobile ad hoc networks(1);security of data(1);wireless networks(1);vehicular ad hoc networks(1);network protocols(1);mobile security(1);wireless telecommunication systems(1);contracts(1);big data(2);gateways (computer networks)(1);authentication(1);wireless sensor networks(1);problem solving(1);telecommunication networks(1);ubiquitous computing(1);circuit theory(1); |
| Ng, Wee Keong | cryptography(1);digital storage(1);distributed computer systems(2);computational complexity(1);internet(1);algorithms(1);information theory(1);errors(1);data privacy(1);big data(1); |

*Author Similarity Analysis*

In this section, we used TF-IDF to find author similarity. TF-IDF is a statistical method that is used commonly to compute weights for information retrieval and information exploration. This study takes into account the frequency of keywords co-occurrence in each author's literature collection and the distribution of all authors' collections and can explore the authors of similar research contents. TF is keywords in the author of the literature focused on the absolute frequency or relative frequency. In this paper, the relative frequency is used as the value of TF. The formula is $n_{kj}tf_{i,j}=n_{i,j}/\sum_{k} n_{k,j}$. In additon, $n_{i,j}$ is the absolute frequency of the keyword i is in the author j keyword set. $\sum_{k} n_{kj}$ is the sum of the frequency of all the keywords of the author j. IDF is used to measure the distribution of keywords in all authors' collections. The formula is $idf_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$, |D| is the total number of authors published papers. $|\{\ j: t_i \in d_j\}|$ is the total number of papers that contain the keyword i. Keywords weight calculation formula : TF-IDF=W=$tf_{i,j}$*$idf_i$. The author's similarity formula : $S_{xy}=\sum_{n} W_i * W_j$, $S_{xy}$ is the similarity between author X and author Y ; $W_i$ is the weight of the keyword i is in the author X, $W_j$ is the weight of the keyword i is in the author Y, n is the number of keyword of author X and author Y.

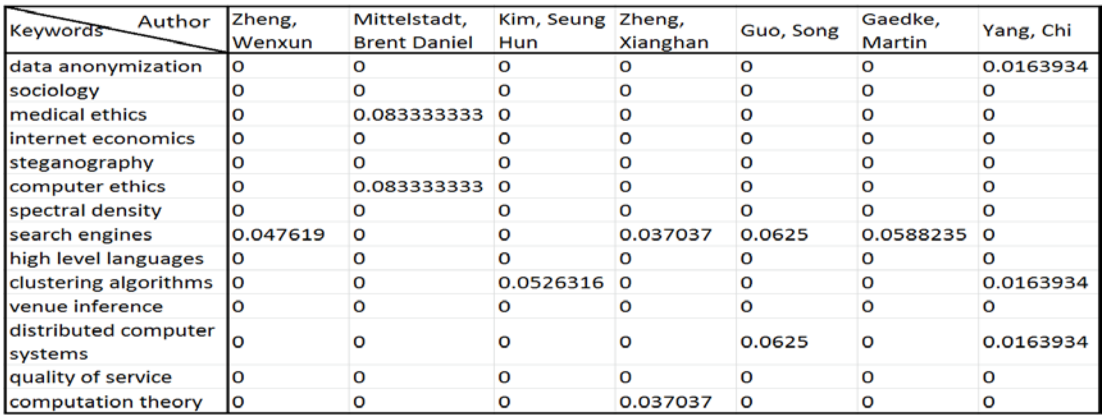
We collect authors who published the number of papers is no less than 2. Then we collect 427 authors, and 998 keywords of theirs. We compute the frequency of keywords, TF, IDF, and TF-IDF by making program, as theTable 6 shows.

**Table 6 Frequency of keywords, TF, IDF, and TF-IDF (partly)**

| Keyword | Frequency | TF | IDF | TF-IDF |
|---|---|---|---|---|
| infrared radiation | 1 | 0.03125 | 0.03125 | 0.0009766 |
| life log | 3 | 0.5 | 0.5 | 0.25 |
| bioinformatics drug discovery | 2 | 0.4 | 0.4 | 0.16 |
| virtual storage | 9 | 1.3333333 | 1.3333333 | 1.7777778 |
| design of experiments | 5 | 0.3125 | 0.3125 | 0.0976563 |
| measurement errors | 1 | 0.0714286 | 0.0714286 | 0.005102 |
| strategic | 2 | 0.4 | 0.4 | 0.16 |
| unstructured data | 4 | 0.6428571 | 0.6428571 | 0.4132653 |
| data framework | 2 | 1 | 1 | 1 |

According to the formula : TF-IDF=W=tfi,j*idfi. By comparing each author's keyword set with the whole keyword set. If the author's keyword set contains someone keyword in the total keyword set, the corresponding matrix element value is its weight. Otherwise, it is 0, we build 427*998 Author-Keyword weight matrix, as the Table 7 shows.

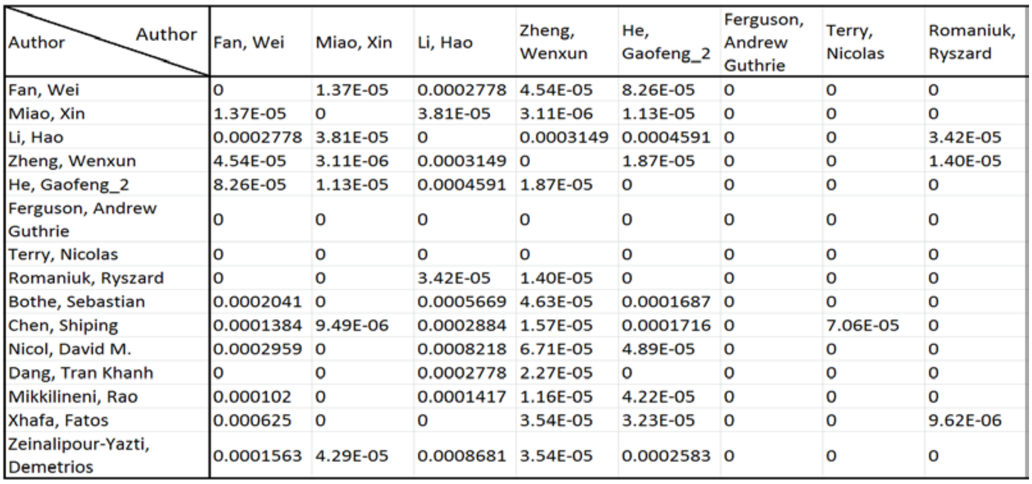
**Table 7 Author-Keyword weight matrix (partly)**



| Keywords \ Author | Zheng, Wenxun | Mittelstadt, Brent Daniel | Kim, Seung Hun | Zheng, Xianghan | Guo, Song | Gaedke, Martin | Yang, Chi |
|---|---|---|---|---|---|---|---|
| data anonymization | 0 | 0 | 0 | 0 | 0 | 0 | 0.0163934 |
| sociology | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| medical ethics | 0 | 0.083333333 | 0 | 0 | 0 | 0 | 0 |
| internet economics | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| steganography | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| computer ethics | 0 | 0.083333333 | 0 | 0 | 0 | 0 | 0 |
| spectral density | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| search engines | 0.047619 | 0 | 0 | 0.037037 | 0.0625 | 0.0588235 | 0 |
| high level languages | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| clustering algorithms | 0 | 0 | 0.0526316 | 0 | 0 | 0 | 0.0163934 |
| venue inference | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| distributed computer systems | 0 | 0 | 0 | 0 | 0.0625 | 0 | 0.0163934 |
| quality of service | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| computation theory | 0 | 0 | 0 | 0.037037 | 0 | 0 | 0 |

According to the formula : $S_{xy}=\sum_{n} W_i * W_j$, depending on the sum of the weights of each keyword corresponding to each author in the above figure, we can get the similarity between the author and the author. Finally, we build the 427*427 Author- Author weight matrix, as the Table 8 shows.

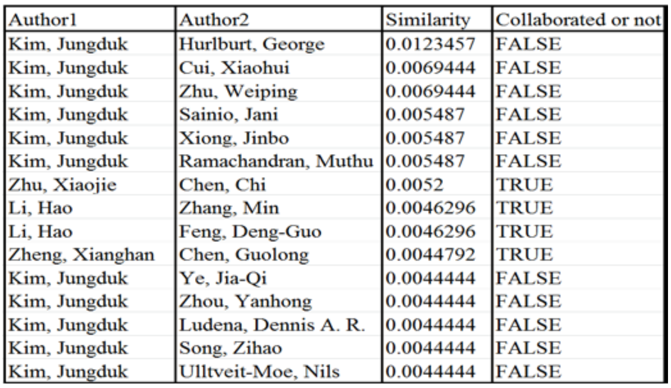
**Table 8 Author- Author weight matrix (partly)**



| Author<br>Author | Fan, Wei | Miao, Xin | Li, Hao | Zheng, Wenxun | He, Gaofeng_2 | Ferguson, Andrew Guthrie | Terry, Nicolas | Romaniuk, Ryszard |
|---|---|---|---|---|---|---|---|---|
| Fan, Wei | 0 | 1.37E-05 | 0.0002778 | 4.54E-05 | 8.26E-05 | 0 | 0 | 0 |
| Miao, Xin | 1.37E-05 | 0 | 3.81E-05 | 3.11E-06 | 1.13E-05 | 0 | 0 | 0 |
| Li, Hao | 0.0002778 | 3.81E-05 | 0 | 0.0003149 | 0.0004591 | 0 | 0 | 3.42E-05 |
| Zheng, Wenxun | 4.54E-05 | 3.11E-06 | 0.0003149 | 0 | 1.87E-05 | 0 | 0 | 1.40E-05 |
| He, Gaofeng_2 | 8.26E-05 | 1.13E-05 | 0.0004591 | 1.87E-05 | 0 | 0 | 0 | 0 |
| Ferguson, Andrew Guthrie | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Terry, Nicolas | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Romaniuk, Ryszard | 0 | 0 | 3.42E-05 | 1.40E-05 | 0 | 0 | 0 | 0 |
| Bothe, Sebastian | 0.0002041 | 0 | 0.0005669 | 4.63E-05 | 0.0001687 | 0 | 0 | 0 |
| Chen, Shiping | 0.0001384 | 9.49E-06 | 0.0002884 | 1.57E-05 | 0.0001716 | 0 | 7.06E-05 | 0 |
| Nicol, David M. | 0.0002959 | 0 | 0.0008218 | 6.71E-05 | 4.89E-05 | 0 | 0 | 0 |
| Dang, Tran Khanh | 0 | 0 | 0.0002778 | 2.27E-05 | 0 | 0 | 0 | 0 |
| Mikkilineni, Rao | 0.000102 | 0 | 0.0001417 | 1.16E-05 | 4.22E-05 | 0 | 0 | 0 |
| Xhafa, Fatos | 0.000625 | 0 | 0 | 3.54E-05 | 3.23E-05 | 0 | 0 | 9.62E-06 |
| Zeinalipour-Yazti, Demetrios | 0.0001563 | 4.29E-05 | 0.0008681 | 3.54E-05 | 0.0002583 | 0 | 0 | 0 |

Logo 427 author already exist the cooperative relationships, identified as TRUE, otherwise FALSE, as the Table 9 shows

**Table 9 427 Author- Author similarity**

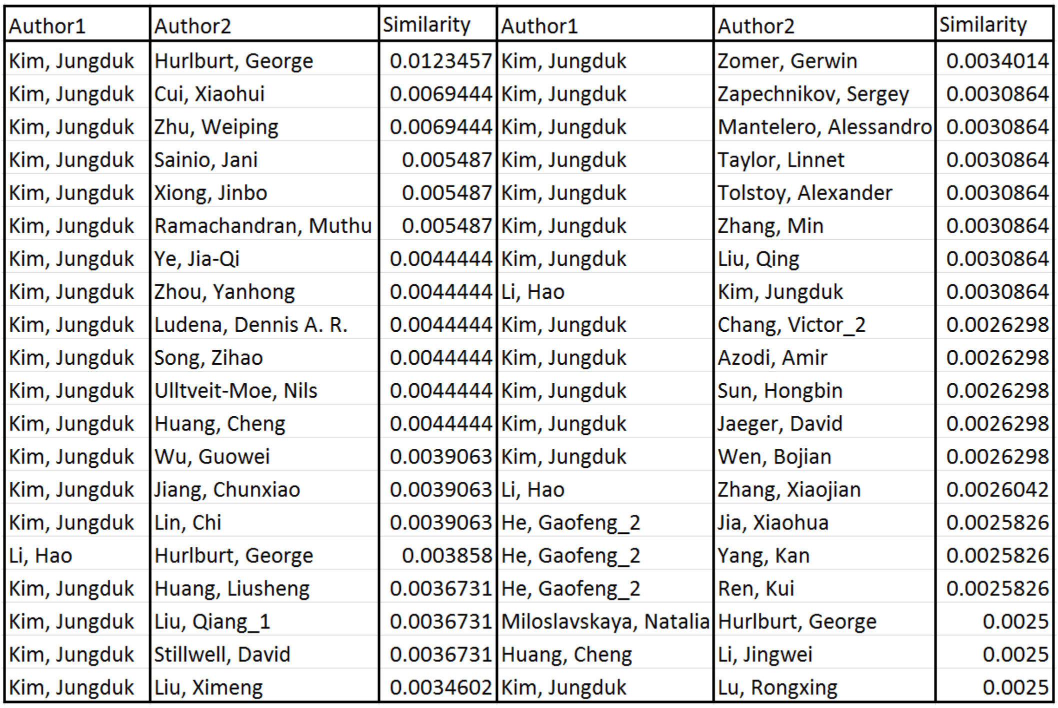


| Author1 | Author2 | Similarity | Collaborated or not |
|---|---|---|---|
| Kim, Jungduk | Hurlburt, George | 0.0123457 | FALSE |
| Kim, Jungduk | Cui, Xiaohui | 0.0069444 | FALSE |
| Kim, Jungduk | Zhu, Weiping | 0.0069444 | FALSE |
| Kim, Jungduk | Sainio, Jani | 0.005487 | FALSE |
| Kim, Jungduk | Xiong, Jinbo | 0.005487 | FALSE |
| Kim, Jungduk | Ramachandran, Muthu | 0.005487 | FALSE |
| Zhu, Xiaojie | Chen, Chi | 0.0052 | TRUE |
| Li, Hao | Zhang, Min | 0.0046296 | TRUE |
| Li, Hao | Feng, Deng-Guo | 0.0046296 | TRUE |
| Zheng, Xianghan | Chen, Guolong | 0.0044792 | TRUE |
| Kim, Jungduk | Ye, Jia-Qi | 0.0044444 | FALSE |
| Kim, Jungduk | Zhou, Yanhong | 0.0044444 | FALSE |
| Kim, Jungduk | Ludena, Dennis A. R. | 0.0044444 | FALSE |
| Kim, Jungduk | Song, Zihao | 0.0044444 | FALSE |
| Kim, Jungduk | Ulltveit-Moe, Nils | 0.0044444 | FALSE |

Delete all the authors who have collaborated, according to the Author-Author similarity, we chose top 40 potential collaborative authors, as Table10 shows. Kim, Jungduk focuses on the data handling and complex networks; Hurlburt, George focuses on data processing; digital storage; mobile network security; Cui, Xiaohui focuses on wireless telecommunication signal systems; cloud computing and digital storage, etc. The similarity between these authors can analyze from their research contents.

Through author keyword coupling analysis, it can help the authors to find some authors who research the similar contents, to provide possible cooperation way for each author in big data privacy and security. Also, In a sense, it also promotes the development of big data privacy and security.

**Table 10 top 40 of the potential collaborative relationships**



| Author1 | Author2 | Similarity | Author1 | Author2 | Similarity |
|---|---|---|---|---|---|
| Kim, Jungduk | Hurlburt, George | 0.0123457 | Kim, Jungduk | Zomer, Gerwin | 0.0034014 |
| Kim, Jungduk | Cui, Xiaohui | 0.0069444 | Kim, Jungduk | Zapechnikov, Sergey | 0.0030864 |
| Kim, Jungduk | Zhu, Weiping | 0.0069444 | Kim, Jungduk | Mantelero, Alessandro | 0.0030864 |
| Kim, Jungduk | Sainio, Jani | 0.005487 | Kim, Jungduk | Taylor, Linnet | 0.0030864 |
| Kim, Jungduk | Xiong, Jinbo | 0.005487 | Kim, Jungduk | Tolstoy, Alexander | 0.0030864 |
| Kim, Jungduk | Ramachandran, Muthu | 0.005487 | Kim, Jungduk | Zhang, Min | 0.0030864 |
| Kim, Jungduk | Ye, Jia-Qi | 0.0044444 | Kim, Jungduk | Liu, Qing | 0.0030864 |
| Kim, Jungduk | Zhou, Yanhong | 0.0044444 | Li, Hao | Kim, Jungduk | 0.0030864 |
| Kim, Jungduk | Ludena, Dennis A. R. | 0.0044444 | Kim, Jungduk | Chang, Victor_2 | 0.0026298 |
| Kim, Jungduk | Song, Zihao | 0.0044444 | Kim, Jungduk | Azodi, Amir | 0.0026298 |
| Kim, Jungduk | Ulltveit-Moe, Nils | 0.0044444 | Kim, Jungduk | Sun, Hongbin | 0.0026298 |
| Kim, Jungduk | Huang, Cheng | 0.0044444 | Kim, Jungduk | Jaeger, David | 0.0026298 |
| Kim, Jungduk | Wu, Guowei | 0.0039063 | Kim, Jungduk | Wen, Bojian | 0.0026298 |
| Kim, Jungduk | Jiang, Chunxiao | 0.0039063 | Li, Hao | Zhang, Xiaojian | 0.0026042 |
| Kim, Jungduk | Lin, Chi | 0.0039063 | He, Gaofeng_2 | Jia, Xiaohua | 0.0025826 |
| Li, Hao | Hurlburt, George | 0.003858 | He, Gaofeng_2 | Yang, Kan | 0.0025826 |
| Kim, Jungduk | Huang, Liusheng | 0.0036731 | He, Gaofeng_2 | Ren, Kui | 0.0025826 |
| Kim, Jungduk | Liu, Qiang_1 | 0.0036731 | Miloslavskaya, Natalia | Hurlburt, George | 0.0025 |
| Kim, Jungduk | Stillwell, David | 0.0036731 | Huang, Cheng | Li, Jingwei | 0.0025 |
| Kim, Jungduk | Liu, Ximeng | 0.0034602 | Kim, Jungduk | Lu, Rongxing | 0.0025 |

## Conclusion and future work

We collected data of 1645 from papers from the Web of Science and EI - Engineering Village 2. Firstly, we analyze the collaborative status by using social network analysis. We find that since 2010, the cooperative situation tends to be fine. Also, through the author collaboration degree is 2.86 coauthors per paper and the author collaboration ratio is 83.89% of collaborative papers, they also mean the network collaboration is well. However, from the density, we find authors do not usually communicate with each other. But there is a big room for them to collaborate with each other. Moreover, 90.92% of authors from 4,277 just start to research the big data privacy and security. Secondly, therefore, we think it is meaningful to mining the potential relationships between them. Then we analysis the authors' similarity by using author keyword coupling analysis, and find some authors have great potential collaborative relationships. Through this paper, it can help authors learn about his or her potential collaborative relationships. In a sense, it can promote the development of big data privacy and security and improve the author's research achievements. There are some inadequacies in this study, such as we just use TF-IDF to find the similarity authors. After this, we will continue to explore new algorithm to find the potential collaborators.

## Acknowledgments

The authors are grateful to anonymous referees and editors for their invaluable and insightful comments. This work is supported in part by the National Natural Science Foundation of China (Grant No. 61100197), the Jiangsu Province Graduate Education Innovation Project (Grant No.KYLX15_0025) and the program B for Outstanding PhD candidate of Nanjing University (Grant No.201702B042).

## Reference

Fan ZY. (2015). A Theoretical Analysis of the Necessity of Cooperation in Production, Teaching and Research. Basic Construction in Chinese, 5, 331. Retrieved October 15, 2017 from: http://www.chinaqking.com/yc/2015/510520.htm.

Pfeffer, J., and Salancik, G. (1978). The External Control of Organizations: A Resource Dependence Perspective, New York: Harperand Row.

Bertalanffy, L. von. (1951). General system theory-A new approach to unity of science (Symposium), Human Biology, 23, 303-361.

Nowak MA. (2006). Five Rules for the Evolution of Cooperation, science, 314(805), 1560-1563.

Wener P, Woodgate RL. (2016). Collaborating in the context of co-location: a grounded theory study, Bmc Family Practice, 17 (1), 30.

Verdecho MJ, Alfaro-Saiz JJ, Rodríguez-Rodríguez R. (2011). A Review of Factors Influencing Collaborative Relationships. Springer Berlin Heidelberg, 362 (2), 535-542.

Dong Y, Huang XW, Sinha K.K, Xu KF. (2014). Collaborative Demand Forecasting: Toward the Design of an Exception-Based Forecasting Mechanism. Journal of Management Information Systems, 31 (2), 245-284

Walker R, Adam J. (1998). Collaborative relationships in general practice projects, Australian Health Review A Publication of the Australian Hospital Association, 21 (4), 203-20.

Bi D. (2008). Research on the partners selection for enterprise-university-institute cooperation, Northeastern university in Chinese, 3-67.

Sun HF, Hou W. (2014). Improved TFIDF algorithm in mining the potential cooperative relationship, New Technology of Library and Information Service in Chinese, 30 (10) , 84-92.

Chen WJ, Zheng Y. (2013). Mining Potential Cooperative Relationships Based on the Author Keyword Coupling Analysis. Journal of Intelligence (in Chinses), 32 (5): 127-131.

Lewin, K. (1936). Principles of Topological Psychology. Martino Fine Books. doi:10.1037/10019-000

Garfield, E. (1979). Citation indexing-It’s theory and application in science, technology, and humanities. New York: John Wiley & Sons.

Stanley, W., & Faust, K. (1994). Social network analysis: Methods and applications. New York: Cambridge University Press.

Freeman, L. C. (2004). The development of social network analysis. Vancouver, BC Canada: ΣP Empirical Press.

Borgatti, S.P., Everett, M.G. & Freeman, L.C. (2002). Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Borgatti, S.P. (2002). NetDraw Software for Network Visualization. Analytic Technologies: Lexington, KY. Retrieved October 15, 2017 from: https://sites.google.com/site/netdrawsoftware/home.

Liu ZH., Zheng YN. (2011). On Research Specialty Evolution Mapping and It’s Application. Journal of the China Society for Scientific and Technical Information, 30(11), 5291-5296.

Liu Z.H, Zhang Z.Q. (2010). Author Keywords Coupling Analysis and Empirical Study. Information Technology in Chinese, 29(2), 268 -275.

Wang XF., Li X., Chen GR. (2012). Introduction to Network Science. Higher Education Press in China, 158-159.

Wang D. (2008). An empirical study of network structure analysis in co-authorship. Information Science in Chinese, 11.

Chen XD., Ceng YY., Xing DP. (2007). The research on social networks of collaborative learning – A case study of a class of AFC. Open Education Research in Chinese, 13(6), 67-71.

Watts D.J., Strogatz S.H. (1998). Collective dynamics ”small world” networks. Nature(393), 440-442.

Freeman, L. (1977). A set of measures of centrality based on betweenness.Sociometry, 40(1), 35–31.

Freeman, L. (1979). Centrality in social networks. 1. Conceptual clarification. Social Networks, 1, 215–239.

Freeman, L. (2000). Visualizing social networks. Journal of Social Structure, 1(1), 1–15.

Fu, Y., Niu, W.Y., Wang Y.L., et al. (2009). Co-Author Network Analysis in the scientific field - Take "Science Research Management" (2004-2008) as an example, Research Management in Chinese, 30(3), 41-46.