2018 IEEE Intelligent Vehicles Symposium (IV)
Changshu, Suzhou, China, June 26-30, 2018

A Bibliometric Analysis of Blockchain Research

Shuai Zeng*†, Xiaochun Ni*†,Yong Yuan*†,*Senior member, IEEE*, Fei-Yue Wang*†‡, *Fellow, IEEE*
*The State Key Lab for Management and Control of Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
† Qingdao Academy of Intelligent Industries, Shandong 266000, China.
‡Research Center of Military Computational Experiments and Parallel System, National University of Defense, China
(e-mails: {shuai.zeng, xiaochun.ni,yong.yuan(Corresponding Author),feiyue.wang}@ia.ac.cn)

***Abstract*—Taking Ei Compendex (EI) and China National Knowledge Infrastructure (CNKI) databases as the literature sources, this paper presented a bibliographic analysis of the blockchain-related literature between January 2011 and September 2017. For each literature source, we built a separate dataset. Authors' productivity and collaboration, affiliation of authors and collaboration amongst institutions were analyzed using techniques of social networks analysis on both datasets. According to the results, the CNKI authors/institutes raised their productivity and outperformed the EI authors/institutes since 2016. However, the EI authors/institutes show a higher level than the CNKI authors/institutes in collaboration. We also summarized the hot topics on the EI dataset using textual analysis and discovered researchers have shifted their attention from Bitcoin itself to the blockchain technology underlying it.**

Keywords: Blockchain, Bibliometric Analysis, Scientific Collaboration, Topic Analysis

I. INTRODUCTION

Generally, a trusted third party is needed to process electronic payments between individuals. To solve this problem, Satoshi presented a peer-to-peer electronic payment system based on cryptographic proof instead of trust, which is named Bitcoin [1]. Bitcoin is starting to come into its own as a digital currency, but the blockchain technology behind it could prove to be much more significant [2].

Blockchain is a decentralized transaction and data management technology. Currently, It has become an emerging field of research and practice [3]–[6]. Blockchain technology is considered in the Peak of Inflated Expectations phrase by research powerhouse Gartner Inc. in their 2016 report Hype Cycle for Emerging Technologies. And in their 2017 report, blockchain have moved significantly along the Hype Cycle. The reason for the interest in blockchain is its central attributes including security, anonymity and data integrity without any third-party trusted systems, and therefore it creates interesting research areas, especially from the perspective of technical challenges and limitations [7].

In order to better understand the state of the art of blockchain-related research, it is important to identify top-tier researchers, institutes and collaboration amongst them as well as hot topics in blockchain. To address these questions, we aim to make a bibliometric analysis on relevant papers related to blockchain.

In this paper, we choose two typical scientific databases as the retrieval sources, i.e., the Ei Compendex (EI) and the China National Knowledge Infrastructure (CNKI). EI is the broadest and most complete engineering literature database available in the world, which is a product offered by Elsevier Engineering Information. It provides a global view of peer reviewed and indexed publications with over 20 million records from 77 countries across 190 engineering disciplines. CNKI is a famous online publishing platform in China. Articles published in CNKI are mostly in the China Integrated Knowledge Resources Database (CIKRD), which provides over 90% of China knowledge resources, including journals, dissertations, newspapers, proceedings, and so on. Therefore with these two knowledge sources, it is available to help reveal the status and trends of blockchain research in a comprehensive perspective.

The rest of this article is organized as follows: Section II describes the collection of our dataset and the methods we used to interpret our dataset. Section III shows the analysis results of these data through various bibliometric procedures. Finally, Section IV summarizes our major findings and discuss our further research.

II. DATA AND METHODOLOGY

In this section, we introduce the data we collected and the methodology we adopted to process and analyze the data in this paper.

A. Dataset

We searched the EI database with the search terms blockchain and bitcoin, and searched the CNKI database with the search term blockchain. We did not use bitcoin as a search term for the CNKI database, because most

Corresponding author: Yong Yuan (email: yong.yuan@ia.ac.cn).
This work is partially supported by NSFC (71472174, 71232006, 61533019, 61233001, 71402178, 71702182) and the Early Career Development Award of SKLMCCS (Y6S9011F4E, Y6S9011F4H, Y6S9011F52).

of its searching results are nonacademic publications. As only research papers are considered in this research, news reports are excluded from the searched results. Roughly, this covers a period of 7 years, as the first piece of blockchain research was published in 2011. Furthermore, we crawled citations and downloads from the CNKI websites for each publication from the CNKI database. We did not choose Web of Science as one of our literature sources, since more than 80% of blockchain-related publications in WoS are overlapped by the EI database. Through the above procedures, we obtained two datasets for this research —EI and CNKI. The EI dataset is the metadata for publications from the EI database. And the CNKI dataset is the metadata for publications from the CNKI database. Both datasets include title, authors, author affiliations, issue date, and keywords of each publication.

For bibliometric analysis, the data should be preprocessed as follows. Authors are identified based on their full names and affiliations. We try to align authors with multiple affiliations so that their contributions would not be underestimated. Note that when evaluating contributions of institutions, multiple affiliations were still considered separately [8].

B. Methodology

Except for straightforward counting numbers of papers, authors, or institutes, we analyze the collected datasets in various ways, which are described as follows.

1) Social network analysis: Coauthorship relations directly reflect collaborations between scientists [9] [10]. And citation relations reflect relationships between pairs of cited papers and citing papers [11] [12]. In this paper, we examine collaborations in the blockchain research field on both datasets at two levels: individual researcher level and institution level. Therefore we constructed two types of coauthorship networks correspondingly. For each type, there are two networks: the EI network and the CNKI network. For researcher-level coauthorship networks, nodes represent authors. Two authors are connected if they have co-authored papers. For institution-level coauthorship networks, nodes represent institutions, and links connect institutions if they are affiliated with coauthors. Besides, we examine paper-level citations on the CNKI dataset and constructed a citation network.

We used Cytoscape [14] to visualize and analyze these networks. In coauthorship networks, the node degree represents numbers of collaborators. Thus the average number of neighbors, clustering coefficients and the number of isolated nodes may offer insights representing whether the authors tend to work alone or collaborate in groups [9]. The average shortest path length and the network diameter give the expected and largest steps for knowledge diffusion between two connected nodes [13]. In citation networks, the node degree represents number of citations for each paper.

2) Textual analysis: A list of word frequencies was generated to compare the use of title words in different time periods. Given titles of blockchain papers, we were able to study researchers' common interests. We divided the EI data into groups by the published year and counted the word frequencies annually. At last, we gained a list of the most frequent words for each year.

III. RESULTS

The statistics of publications in both datasets are listed in the Table I. In our EI dataset, there are 473 papers, 1063 authors and 622 institutes. And in our CNKI dataset, there are 497 papers, 765 authors and 423 institutes. The number of papers in both two datasets are close, whereas the EI dataset has a larger number of authors than the CNKI dataset. Another finding is that the proportion of isolated authors (e.g. authors who are not co-author with anybody) in the CNKI dataset is higher than in the EI dataset. This demonstrates that the EI authors may have a higher incentive than the CNKI authors in collaboration.

For the EI dataset, on average, each author is involved with 1.35 papers, and each institution publishes 1.3 papers. And for the CNKI dataset, each author is involved with 1.15 papers, and each institution publishes 1.58 papers.

TABLE I
STATISTICS OF PUBLICATIONS IN BOTH DATASETS

	EI	CNKI
Papers	473	497
Authors	1063	765
Isolated Authors	67	229
Average Papers for Authors	1.35	1.15
Max Papers for Authors	10	6
Institutes	622	423
Isolated Institutes	264	177
Average Papers for Institute	1.3	1.58
Max Papers for Institutes	25	9

Figure 1 depicts the total number of papers on a yearly basis. Both the EI and the CNKI papers have shown obviously ascending trends. However, their growth patterns are quite different. The number of the EI papers keeps steadily increasing while the number of the CNKI papers does not change significantly until 2016. In fact, the CNKI authors/institutes raised their productivity and exceeded the EI authors/institutes since 2016.

A. Productivity Analysis

Table II reports the most productive authors in the EI dataset. Karame, Ghassan O. (NEC Laboratories Europe, Heidelberg) published the largest number of papers according to paper counts, followed by Wattenhofer, Roger (ETH, Zürich) and Decker, Christian (ETH, Zürich). These top authors have been involved in more than six papers. Furthermore, we notice that five

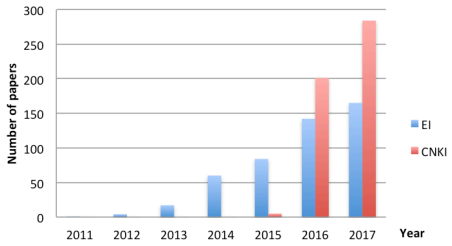

Fig. 1. Papers published by year

of them are from newly established blockchain-oriented research institutes, including ETH, Initiative for CryptoCurrencies and Contracts (IC3) and Commonwealth Scientific and Industrial Research Organisation (CSIRO) 's Data61, which indicates a great power that cannot be ignored. Besides, five top authors are from Europe and four are from the U.S.. This clearly shows the leading role of both Europe and the U.S. in a global context of blockchain research.

TABLE II
MOST PRODUCTIVE AUTHORS (EI)

Name	Affiliation	Counts
Karame, Ghassan O.	NEC Laboratories Europe, 69115 Heidelberg, Germany	10
Wattenhofer, Roger	ETH, Zürich, Switzerland	8
Decker, Christian	ETH, Zürich, Switzerland	8
Miller, Andrew	IC3, Ithaca, United States; UMD, College Park, United States	8
Kiayias, Aggelos	University of Athens, Athens, Greece	7
Clark, Jeremy	Concordia Institute for Information Systems Engineering, United States	7
Weber, Ingo	Data61, CSIRO, Sydney, Australia; School of Computer Science and Engineering, UNSW, Sydney, Australia	6
McCorry, Patrick	School of Computing Science, Newcastle University, Newcastle upon Tyne, United Kingdom	6
Zohar, Aviv	Microsoft Research, Silicon Valley, Mountain View, CA, United States	6
Shi, Elaine	IC3, Ithaca, United States; Cornell University, Ithaca, United States	6

Table III reports the most productive authors order by paper counts in the CNKI dataset. Yi Qin (Asia Pacific International Core of Excellence, Deloitte) published the largest number of papers, followed by Xuemai Yu (Baoquan.com) and Xuchuan Wu (Institute of Finance, People's Bank of China). These top authors have been involved in more than three papers, which is lower than the EI top authors on average. Although most of the top authors are still from universities and traditional research institutions, the best ranked two authors are from service companies. We also notice that there is no research institute specializing in blockchain according to the list. It seems that Chinese researchers have suffered from lagged responses while the industrial community have deeply involved in the Challenges. According to the citation counts and download counts, Yong Yuan and Fei-Yue Wang(Institute of Automation, Chinese Academy of Sciences) published papers with the largest number of citations (144) and downloads (18966).

TABLE III
MOST PRODUCTIVE AUTHORS (CNKI)

Name	Affiliation	PC	CC	DC
Yi Qin	Asia Pacific International Core of Excellence, Deloitte	6	1	601
Xuemai Yu	Baoquan.com	5	1	653
Xuchuan Wu	Institute of Finance, People' s Bank of China	4	2	748
Gaoying Cui	Jiangsu Electric Power Research Institute	4	4	1386
Bin Li	North China Electric Power University	4	4	1386
Bin Qi	North China Electric Power University	4	4	1409
He Li	Chinese People's Insurance	4	0	547
Yin Cao	Energy Blockchain Laboratory	4	3	745
Yong Yuan & Fei-Yue Wang	Institute of Automation, Chinese Academy of Sciences	3	144	18966
Wei-Tek Tsai	Beihang University	3	1	1705

PC: Paper Counts, CC: Citation Counts, DC: Download Counts.

Table IV shows the top ten institutions according to the paper counts in the EI dataset. In this list, the highest ranked institute ETH has published 25 papers, more than twice as the second-place NEC Laboratories. Note that we don't label the country of NEC laboratories and Microsoft Research as they are both transnational corporations whose publications are from different nations/regions. Among the top ten institutions, four are from Europe, which indicates the importance of Europe in blockchain research. Other productive institutions are in the U.S, China and Australia.

Table V shows the top ten institutions according to the paper counts in the CNKI dataset. Industrial and Commercial Bank of China (ICBC) and North China Electric Power University (NCEPU) are tied for the most productive institutions, both having 9 papers. We also notice that three institutes are nationalized banks, which suggests that Chinese state-controlled financial systems have attached importance to blockchain technology.

We further construct a paper-level citation network for the CNKI dataset. In this network, each node represents

TABLE IV
MOST PRODUCTIVE INSTITUTES (EI)

Name	Counts
ETH, Zürich, Switzerland	25
NEC Laboratories	11
Computer Science Department, Cornell University, Ithaca, NY, United States	11
Beihang University, Beijing, China	10
Newcastle University, Newcastle Upon Tyne, United Kingdom	9
University of Athens, Athens, Greece	9
Data61, CSIRO, Sydney, Australia	8
University College London, London, United Kingdom	8
Microsoft Research	8
Stanford University, California, United States	7

TABLE V
MOST PRODUCTIVE INSTITUTES (CNKI)

Name	Counts
Industrial and Commercial Bank of China	9
North China Electric Power University	9
Agricultural Bank of China	8
China Academy of Information and Communications Technology	8
Tsinghua University	8
Institute of Finance, People's Bank of China	8
Chinese Academy of Social Sciences	7
Beijing University of Posts and Telecommunications	7
Beihang University	7
Peking University	7

a paper and each edge represents a citation relation from one node to another. Besides, there is a positive correlation between node sizes and citation counts.

The network is shown in Fig. 2, which includes 348 nodes and 691 edges with 6 connected components. There is no isolate node in this network, as every paper has cited at least one paper in the CNKI dataset. The giant component contains 335 nodes, illustrating that most papers are connected by citation relationships. Several highly cited papers are labeled in the network. The paper with largest citation counts (144) is published by Yong Yuan and Fei-Yue Wang in 2016. It is also the key-player in the giant component.

B. Collaboration Patterns

Fig. 3 and Fig. 4 show the paper distributions for researchers and institutions. As for the EI papers, coauthor is prevalent as more than 83.5% of the papers in the EI dataset having more than one authors. And most collaborations take place within the same research institute (54.5%) or between two institutes (28.8%). However, most CNKI papers have only one author (57.9%) and one institute (74.6%), which shows low collaboration.

We further construct researcher-level and institute-level coauthorship networks and conduct topological

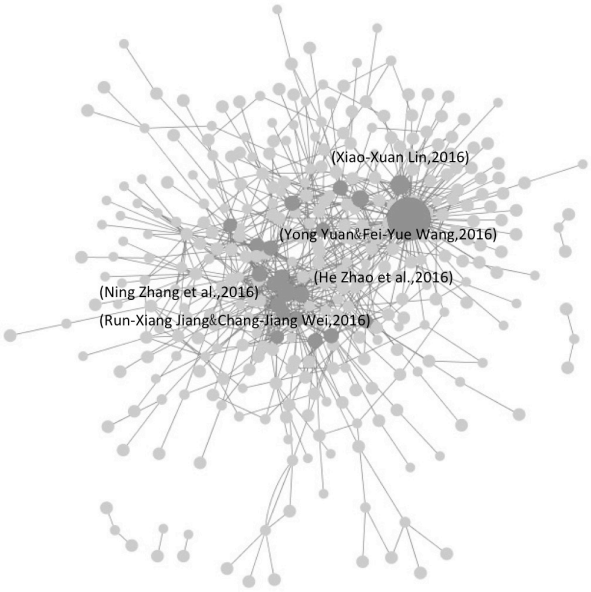

Fig. 2. Paper-level Citation network (CNKI)

analysis on these networks to obtain more detailed analysis results. In these networks, each node represents an author/institute and each edge represents coauthorship relations between authors/institutes. There is a positive correlation between node sizes and paper counts, as well as edge widths and collaboration counts. We provide visualization of these networks(see Fig. 5, Fig. 6, Fig. 7, and Fig. 8) and label the most prolific authors/institutes. Note that isolated nodes aren't shown for convenience.

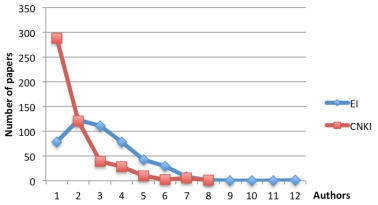

Fig. 3. Paper distributions on authors.

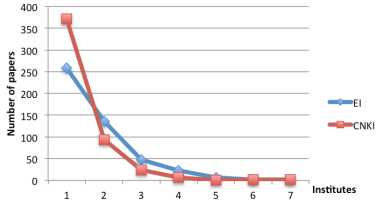

Fig. 4. Paper distributions on institutions

1) Researcher-Level Coauthorship: The researcher-level coauthorship network for the EI dataset is shown in Fig. 5, which includes 1063 nodes, 1723 edges, 232 connected components and 67 isolated nodes. The giant component contains 94 nodes. In this coauthorship network,

the average degree is 3.46 and the clustering coefficients is 0.789. However, within the giant component, the average degree is 5.83 and the clustering coefficients is 0.804. Miller, Andrew has the largest number of collaborators (degree = 23). The authors with the second largest degree (19) are Shi, Elaine and Saxena, Prateek. They are all key-players in the giant component. Two hundred and thirty-seven pairs of authors collaborate with each other more than once. The most frequent cooperation happens between Wattenhofer, Roger and Decker, Christian (8).

The researcher level coauthorship network for the CNKI dataset includes 765 nodes, 680 edges, 174 connected components and 229 isolated nodes (see Fig. 6). There are numerous small components in this network, whose sizes are smaller than the EI components in general. The giant component contains only 18 nodes. Obviously, the EI authors show a higher level than the CNKI authors in collaboration. The average degree of the nodes is 2.537 and the clustering coefficients is 0.591 in this network. Both the two parameters in the CNKI network are smaller than in the EI network. Bin Li, Bin Qi, and Gaoying Cui are authors with the largest number of collaborators (degree = 14). These three authors are all in the giant components. Only twenty-five pairs of authors collaborate with each other more than once. The most frequent cooperation happens between Bin Li and GaoYing Cui (4).

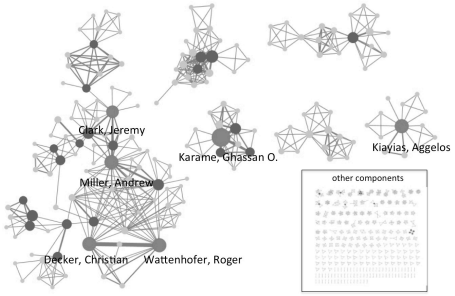

Fig. 5. Connected components in the researcher-level coauthorship network (EI)

2) Institute-Level Coauthorship: The institution-level coauthorship network for the EI dataset includes 622 nodes, 538 links, 85 components and 264 isolated nodes (see Fig. 7). In this network, the average degree is 2.52 and the clustering coefficients is 0.519. The largest component contains 93 institutions (degree=3.591) and the second largest component contains 46 institutions (degree=3.87). ETH, Zürich is with the largest degree (18) and Beihang University, Beijing is with the second largest degree (17). The former is the key-player in the largest component and the latter is in the second largest component. Forty pairs of institutions collaborate with each other more than once. The most frequent cooperation happens between NEC Laboratories and ETH, Zürich, Switzerland (7).

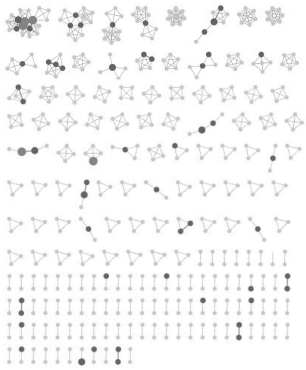

Fig. 6. Connected components in the researcher-level coauthorship network (CNKI)

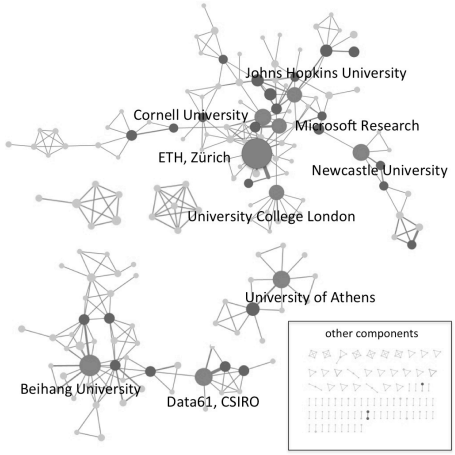

Fig. 7. Connected components in the institute-level coauthorship network (EI)

The institution-level coauthorship network for the CNKI dataset includes 423 nodes, 212 links, 95 components and 177 isolated nodes (see Fig. 8). In this network, the average degree is 1.724 and the clustering coefficients is 0.352. Both the two parameters in the CNKI network are smaller than in the EI network as well. Besides, the giant component contains only 8 institutions. The institution with the largest degree (7) is China Electric Power Research Institute, which is in the giant component. Only nineteen pairs of institutions collaborate with each other more than once. The most frequent cooperation happens between China Electric Power Research Institute and NCEPU (4). Therefore, the EI institutes also show a higher level than the CNKI institutes in collaboration.

C. Research Topics

We divided the EI data into three phases based on papers' published year: 2011-2015, 2016 and 2017. Popular research topics of each phase are shown in Fig. 9 based on the frequencies of words in paper titles.

We noticed that the topic Bitcoin maintained a dominant position between 2011 and 2015. And since 2016,

Fig. 8. Connected components in the institute-level coauthorship network (CNKI)

blockchain is emerging as a new hot research topic and sharing the status as a leading topic with Bitcoin. Finally in 2017, blockchain defeat all others, and Bitcoin turned into a trivial one. This demonstrates that more and more researchers have shifted their attention from Bitcoin itself to the blockchain technology behind it.

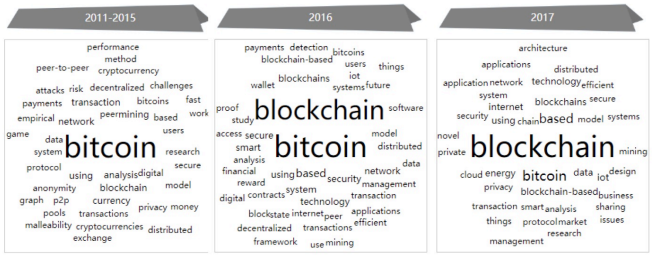

Fig. 9. Research topics in different phases

IV. CONCLUSIONS

In this paper, a bibliographic analysis of blockchain research is presented. We took EI and CNKI as the literature sources and obtained blockchain-related literature published between January 2011 and September 2017. For each literature source, we established a separate dataset. By analyzing papers, authors or institutes, we list the most productive authors and institutes.

Furthermore, we construct four coauthorship networks and one citation network based on the two datasets. Social network analysis methods are applied on these networks to analyze authors' productivity and collaboration, affiliation of authors and collaboration amongst institutions. According to the results, the CNKI authors/institutes raised their productivity and outperformed the EI authors/institutes since 2016. We found that coauthorship is more popular in the EI dataset than in the CNKI dataset, which suggests that the EI authors/institutes has a higher level than the CNKI authors/institutes in collaboration.

We also summarized the popular research topics on the EI dataset using textual analysis and discovered the topic blockchain has defeated Bitcoin so as to risen to the first place since 2017, which indicates that researchers have shifted their attention from Bitcoin itself to the blockchain technology behind it.

In the future, we plan to make a more comprehensive and in-depth analysis based on more network parameters, e.g. betweenness centrality indicating the importance of a given node in knowledge diffusion [14].

ACKNOWLEDGMENT

This work is partially supported by NSFC (71472174, 71232006, 61533019, 61233001, 71402178) and the Early Career Development Award of SKLMCCS (Y6S9011F4E, Y6S9011F4H, Y6S9011F52).

References

[1] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008): 28.
[2] Swan, Melanie. "Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc., 2015.
[3] Y. Yuan, F.Y. Wang. "Blockchain: The state of the art and future trends." Acta Automatica Sinica 42.4 (2016): 481-494.
[4] Y. Yuan, T. Zhou, A. Zhou, Y. Duan, F.Y. Wang. "Blockchain technology: from data intelligence to knowledge automation." Acta Automatica Sinica, 2017, 43(9):1485-1490.
[5] Y. Yuan, F.Y. Wang. "Parallel Blockchain:Concept, Methods and Issues." Acta Automatica Sinica, 2017, 43(10): 1703-1712. doi: 10.16383/j.aas.2017.c170543
[6] Y. Yuan, F.Y. Wang. "Towards blockchain-based intelligent transportation systems." Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on. IEEE, 2016.
[7] Yli-Huumo J, Ko D, Choi S, Park S, Smolander K."Where Is Current Research on Blockchain Technology? — A Systematic Review." PLoS ONE 11(10): e0163477. https://doi.org/10.1371/journal.pone.0163477
[8] Li, Linjing, et al. "Research collaboration and ITS topic evolution: 10 years at T-ITS." IEEE Transactions on Intelligent Transportation Systems 11.3 (2010): 517-523.
[9] M. E. J. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 64, no. 1, p. 016 131, Jun. 2001.
[10] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration," Proc. Nat. Acad. Sci. U.S.A., vol. 101, no. 1, pp. 5200–5205, Apr. 2004.
[11] Liang, Yicong, Qing Li, and Tieyun Qian. "Finding relevant papers based on citation relations." International Conference on Web-Age Information Management. Springer, Berlin, Heidelberg, 2011.
[12] Liu, Haifeng, et al. "Context-based collaborative filtering for citation recommendation." IEEE Access 3 (2015): 1695-1703.
[13] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top., vol. 64, no. 1, p. 016 132, Jun. 2001.
[14] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T., "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Research 2003 Nov; 13(11):2498-504