A Bibliometric Analysis of the Surveys in Blockchain

Hira Ahmad\(^1\), Muhammad Ahtazaz Ahsan\(^1\), and Adnan Noor Mian\(^1\)

\(^1\) Information Technology University (ITU), Punjab, Lahore, Pakistan

**Abstract**—Blockchain has become an emerging technology in the field of computer science and is used for asset security, anonymity, and verifiability, etc. Enormous use cases of blockchain, e.g., for making online payments, storing or sharing of private data in a tamper-proof way has attracted the research community in the past decade. In this regard, a lot of survey papers have been published focusing on the domain-specific exploratory study of applying blockchain in that particular domain. In this paper, we collected those survey papers published and collected 542 unique papers covering the span of 2017 to 2020 published in different venues. We then perform a bibliometric study on those papers and provide interesting insights on different features extracted from that papers. We hope this study will serve as a valuable resource for the blockchain community.

I. INTRODUCTION

Blockchain is evolved as an emerging technology in the field of computer science. Blockchain is a decentralized, peer-to-peer network, in which nodes communicate with each other in a trust-less environment. Blockchain was first coined in 2008 by Satoshi Nakamoto [1]. Blockchain has unique characteristics, which are; (i) making data tamper-proof, (ii) distributed ledger, and (iii) no single point of failure, etc. The main structure of blockchain consists of blocks that are made up of transactions. The blocks are created through a process called mining in which a miner node collects, verifies and, adds new transactions into a block. The blocks are linked together in a chain-like structure through hashing. A variety of blockchain applications ranging from data security [2], data sharing [3], networking [4], and secure authentication [5] have shown its vast applicability in various domains. Blockchain is also applied in other domains of computer science like the internet of things (IoT), healthcare, data security and sharing, fog and cloud computing, artificial intelligence (AI) or access control, etc.

A survey or review paper covers the state-of-the-art research techniques, challenges, opportunities, or future work in the specific area of knowledge. It also provides the taxonomy or categorization of the present work and gives useful insights in the form of future work and challenges faced by the research community. The community takes such papers as a guide to enhance the current work in a particular field. In this way, the research and innovation go side by side for the development of new tools and techniques. To the best of our knowledge, the first survey paper on blockchain was published in 2017, which is almost after 9 years since the blockchain was first introduced. There are 542 survey papers published in the field of blockchain from 2017 to 2020 covering different domain areas. So, a large number of survey papers are currently available which is quite unprecedented. To compare this, the query on Google Scholar’s [6] advanced search about “ad-hoc networks” AND “survey” OR “ad-hoc networks” AND “review” from 2017 to 2020 resulted in approximately 100 survey papers. To investigate the trends in blockchain survey paper, we developed some research questions (RQs), which are described as follows. **RQ1:** What is the overall trend in survey paper publication since 2017 and what percentage of them are published in different venues like IEEE, Elsevier, etc. **RQ2:** What is the frequency of references, citations, and paper size (in terms of number of pages) are present in all survey papers. **RQ3:** In which domain, most of the blockchain-related survey papers are written. **RQ4:** Which countries and authors are writing the most number of survey papers and can we identify the collaboration network between different authors. This motivates us to perform a bibliometric analysis on the survey or review papers published in the blockchain domain to answer the above defined RQs.

A bibliometric analysis is a quantitative research evaluation approach that assesses the previous relevant scientific works based on quantitative indicators. In other words, a bibliometric study is performed to analyze and visualize the development of a scientific field. For example, an evaluation of publications on the metrics of citations, authors, countries, etc. is conducted for the bibliometric study. The main purpose of performing a bibliometric study is to recognize the state of the art of the scientific field in a specific area. We perform bibliometric analysis on the surveys or review papers published in the blockchain field of computer science. We collect the published and bibliometric data of 542 papers that are published in different venues and under different publishers.

**Organization of the Paper:** The rest of the paper is organized as follows. We provide a brief overview of related work in Section II. We describe data collection and preprocessing methods in Section III. We present bibliometric analysis in Section IV. Finally, we conclude our paper in Section V.

II. RELATED WORK

Firdaus et al. [7] used the term “blockchain” for collecting blockchain-related articles. Authors extracted 1119 articles during 2013 to 2018 from the Scopus [8] and discovered that; (i) the future trend will be of solving IoT security issues, (ii) blockchain will be mostly used in healthcare, (iii) USA, China, and Germany have the most number of publications in...
blockchain, (iv) Singapore and Switzerland have fewer publications and many citations, (v) higher research collaborations mean higher the publications except for Canada, India and, Brazil, (vi) the keyword analysis showed that blockchain is used in various fields of research. Similarly, Guo et al. [9] has obtained 3826 articles for blockchain published during 2013 – 2020 from Web of Science (WoS) [10], and used CiteSpace [11] and VOSviewer [12] to extract publication trends, top-cited authors, highly cited journals, most-cited references, authors’ network, top-productive countries and institutions, and emerging trends of blockchain. Zeng et al. [13] selected Ei Compendex [14] and China national knowledge infrastructure [15] for blockchain-related literature to extract literature between January 2011 and September 2017. From both sources of literature, they analyzed the most productive authors and institutes, collaboration patterns among authors and institutes, and the emerging topics.

Smart contracts (SC) are simple programs that are stored on a blockchain and execute only when predetermined requirements are met. They automate the execution of a contract so that all members can be instantly sure of the outcome, without time loss. Ante [16] searched the term smart contract from WoS for analyzing 468 articles having 20,188 references in them which have 15,714 unique papers referenced in them. This work has applied exploratory factor analysis for co-citation analysis to recognize six groups of research that are (i) blockchain networks development, (ii) blockchain and smart contracts for IoT, (iii) smart contract security, standardization, and verification, (iv) smart contracts and blockchain for the disruption of industries, (v) challenges of smart contracts, and (vi) smart contracts and law. The work [17] has claimed to be the first general bibliometric study of bitcoin literature that collected 1162 papers during 2012 to 2019 from WoS and they have found the leading authors, main research areas, countries with most publications, most productive authors, research clusters, and leading authors.

Ante et al. [18] analyzed 166 articles from WoS on the energy sector with blockchain, used exploratory factor analysis to find six research streams that are (i) energy market reform and change, (ii) blockchain for security and data sharing, (iii) energy management in scalable systems and smart grids, (iv) information sending across networks and its applications, (v) peer-to-peer energy micro-grids and, (vi) blockchain technology potential. Social network analysis is applied to these streams to find the relationships and dependencies among them. The results showed that there were more than 71.6% of variance among the above mentioned streams.

Müßigmann et al. [19] analyzed the articles from 2016 to January 2020 on the domain of logistics and supplychain managament (LSCM) along with blockchain technology (BCT). The dataset was collected from 10 databases including Scopus, Google Scholar (GS) and from WoS, then applied refined data collection process which made the articles count to 613. Authors then performed statistical analysis on affiliations and collaborations between different authors, highlighted the keywords, and also performed a citation and a co-citation network analysis that helped to divide the existing work into five classes as (i) theoretical sense-making of BCT in LSCM, (ii) testing and conceptualizing blockchain applications, (iii) digital supplychain management, (iv) technical design of BCT applications for real-world LSCM applications, (v) framing BCT in supplychains.

Moosavi et al. [20] performed bibliometric analysis on the articles and book chapters collected from Scopus for application of blockchain in supplychain to find out the important studies that let them define the supplychain areas and additional integrated technologies, main research groups, institutions, and, countries. Tandon et al. [21] selected 586 articles from Scopus covering the domain of management in blockchain and included 72 countries, 273 journals, 1016 organizations, and 1284 authors. Their findings are based on blockchain applications in particular managerial areas, e.g., finance and supplychain management. In their research work, they recognized four sub-categories of research as (i) policy and management, (ii) enablement of blockchain in management, (iii) multi-domain deployment, and (iv) incompetence of bitcoin.

For blockchain in the IoT domain, Kamran et al. [22] conducted a bibliometric study on the dataset containing 151 articles extracted from WoS [10]. The authors analyzed the yearly trends of publications, keyword analysis, the highest average citations per year, and the top listed venues. Anjum et al. [23] identified useful insights by performing a bibliometric study on blockchain and healthcare domain. They identified yearly trends of publications, highest publications by authors, institutes, countries, and publishers from all over the world where the data was collected from Scopus from January 2020 to March 2020.

Our work is different from the other bibliometric works in a way that we have limited our scope to only surveys and reviews papers for the time span of four years (2017 to 2020). We have covered different publishing venues and collected our dataset. Then we performed bibliometric analysis to look into the trends in writing surveys in blockchain that can assist the researchers to visualize the domain diversity.

### III. DATA COLLECTION AND PRE-PROCESSING

To collect our dataset, we scrapped the data from Google Scholar (GS). For this purpose, we used Google Scholar’s advanced search option and queried the phrase which consists of the word “Blockchain” AND “Survey” OR “Blockchain”

### TABLE I: Different attributes, their types and brief description that are scrapped for performing bibliometric analysis.

| Attribute name     | Type of Attribute | Description                                                                 |
|--------------------|-------------------|-----------------------------------------------------------------------------|
| Paper Title        | string            | The title of the published paper.                                            |
| Venue              | string            | The venue on which the paper is accepted for publication.                   |
| Type of Venue      | string            | The venue is a conference, journal, or an ArXiv.                            |
| Year               | integer           | Year in which the paper is published.                                       |
| Citation count     | integer           | Number of citations received by a paper.                                    |
| References count   | integer           | Number of reference papers that are cited in a paper.                       |
| Author names       | string            | Names of first author and all other authors in the paper.                   |
| Country            | string            | Country of the first author and all other authors in the paper.             |
| Paper size         | integer           | Length of the paper in number of pages.                                     |
AND “Review”. The purpose of choosing GS is because it is free of cost, and also we can get a higher percentage of publications and citations over all the fields which are far greater than Scopus and WoS. In the field of computer science, GS has almost all the citations [24] of Scopus and WoS, which makes it suitable over other mentioned databases. We selected those papers which are published from 2017 to 2020. After applying our query, we downloaded each paper and collected different attributes of each paper as shown in Table I. These attributes include paper title, venue in which the paper is published, type of paper (conference, journal, or not published but available online on arXiv), year in which the paper was published, count of citations a paper received till December 2020, count of references (number of papers cited in particular survey or review paper), all authors, their country and institute, paper size (number of pages) and, the published text (from abstract till conclusion).

After collecting all data, we converted all portable document format (pdf) files into text documents and applied basic pre-processing techniques like removing all tables, images, header/footer, authors and institute information, references and, biography details. After that, we removed all extra spaces from the text files and copied the values of all important attributes into a separate file.

IV. BIBLIOMETRIC ANALYSIS

Yearly statistics are shown in Fig. 1. It can be seen that there is an increase in the number of papers from 2017 to 2020. This shows the maturity of the blockchain and its wider applicability in other domains. It can be seen from Fig. 1 that there are more number of journal papers (318) than the other two classes, i.e., conference and not published. This shows that more of the conferences generally do not accept survey or review papers.

A. Publication Type

The publication type refers to the type of paper which is adopted by authors and publishers. Commonly, there are three main types of publications, which include: (i) conference, (ii) journal and, (iii) not published but available online aka preprint or ArXiv version of the paper. We obtained year-wise statistics based on these three types of publications. The results are shown in Fig. 1. There are very few papers published in 2017 and, there are no such papers which are not published in any venue. But in the next years, there is an increase in the number of journal publications. This result shows immense interest in blockchain applicability and a wider range of applications.

B. Publication Venue

In our collected dataset, we analyzed that among all papers, 96% papers are published in some venues and 4% papers are not published but available online, i.e., in preprint form. The distribution of overall published papers under the venue is shown in Fig. 2. It can be seen that the maximum number of papers (30%) are published in IEEE (both conference and journal). Also, Springer, Elsevier, MDPI, ArXiv, ACM and, other venues have 14%, 11%, 5%, 4%, 3% and, 34% distribution of papers, respectively. In “others” category, we combine all those venues, which have less than 5 published papers.

In Fig. 2, we have seen that are 163 out of 542 papers are published in “IEEE”, among which 93 are conference and 70 papers are journal papers. “IEEE access” has the most number of papers (47) as compared to the other venues. “IEEE communications surveys & tutorials” (ComST) has 7 papers, and IEEE Internet of Things Journal has 5 papers.

C. References

References are those papers that are cited in a particular article. As a thumb rule, more number of references in a paper show an exhaustive literature review. Results for the number of references are shown in Fig. 3, where $x$ – axis represents the count of papers cited in a paper (number of references used) and $y$ – axis shows the number of papers which have used those references. It can be inferred from the results that 73% citations are in the range of 10 to 100. Also, 58% of the papers in data contains more than 40 references. There is only one article that has a maximum of 461 references.

D. Citations

A paper gets cited when other authors refer to it in their own papers. In our collected data, citations are recorded till
December 2020. The results based on citations are explained as follows. There are 160 out of 542 papers that are not cited by any article. On the other side, there is only one paper that has 1243 citations. Further trends are shown in Fig. 4 where $x$-axis shows the number of citations a paper received and $y$-axis represents a total number of papers that fall under that number (frequencies of a particular citation). The overall trend can be viewed as a decrease in the number of papers that are highly cited. The dense area is circled and shown in the plot as represented in the box of in Fig. 4 which is showing that most of the papers are less cited.

E. Paper Size

Statistical analysis for the size of each paper published is presented in Fig. 6. Approximately 29% of papers are short papers, i.e., the number of pages is in the range of 2 to 7. Any paper having $8 \leq$ paper size $\leq 15$ is considered as medium-sized papers and 32% of papers belong to this category. The rest of the survey papers (208) are large-sized papers having page size greater than 15.

F. Statistics on classification of surveys in blockchain

Blockchain has been applied in many domains due to its vast applicability in healthcare, IoT, and smart cities, etc. Also, some papers are published to cover different components or features of blockchain like smart contracts (SC), blockchain security, consensus or fog/mobile/quantum (FMQ) computing, etc or use of blockchain with machine learning, artificial intelligence (AI), etc. We have also classified the surveys into two main categories, named, blockchain application scenarios and blockchain in other domains. The results for these statistics are shown in Fig. 7. In the first set, we have covered the main domains on which the blockchain is applied. The categorization of the papers is achieved by using the title of the papers. We removed stop words like is, of, to, on, for, etc. from the title strings and kept the keywords. For the sake of simplicity, we also removed the words like blockchain, surveys and, review from the title string just because they are redundant and available in each title string. Further, we converted each word into lowercase and also applied stemming (a process in natural language processing for obtaining the stem word, also called root word).

The results for the first set are shown in Fig. 7(a). It can be seen that IoT has the most number of papers (58) as well as the most diverse sub-domain areas as shown in the legend of Fig. 7(a). Second highest number of survey papers are published in healthcare domain (38). There are more survey papers on IoT security than healthcare security. Similarly, the results for the second category of classification...
are shown in Fig. 7(b). In this set, we have further classified the keywords representing another domain of computer science or a component of a blockchain. For example, a domain can be in fog/mobile/quantum (FMQ) computing and a component can be consensus or scalability of blockchain, etc. Interestingly, it can be seen that most of the papers are on the “issues and challenges” of the blockchain. The second most number of surveys are on consensus in blockchain (a blockchain method to maintain the unified state of the ledger in a decentralized system). The most diverse surveys (having more sub-domains) are conducted in the field of “Smart contracts (SCs)”.

We identified those domains on which the researches have been conducted from 2017 to 2020. In 2017, the researchers have focused on the consensus mechanisms. In 2018, the researchers have focused on the area of blockchain in IoT more than another topic. In 2019 the researchers mainly focused on blockchain-based IoT applications and their security and in 2020, the researchers again focused more on the blockchain-based IoT applications. We can conclude that blockchain-based IoT applications have been a major research area in recent years.

G. Countries, Authors and their Collaboration

1) Statistics on country-wise publications: The results for country-wise publication are shown in Fig. 8. Most authors (first authors) are found to belong from India and China, with the publication of 97 and 79, respectively. The total number of unique countries which have published their survey papers is found to be 68. It is also observed from the dataset that 25 countries have published only one paper showing that almost 37% of the total countries have only one survey paper published while India and China have published 33% of the total published papers.
TABLE II: Statistics on authors and their collaboration

| Attribute                                         | Count |
|--------------------------------------------------|-------|
| Total authors in survey papers                   | 1682  |
| Unique first authors                             | 449   |
| Unique other authors (other than first authors)  | 1221  |
| Total collaborations between pair of authors      | 3065  |

Fig. 9: Author collaboration network

2) Author’s Collaboration: We have identified the collaboration between different authors. To find the collaboration between authors, we have assumed that if any two authors are present in the same papers, they have collaborated with each other in that. To perform this, first of all, we have identified all first authors. As a standard practice, all papers must have mentioned the name of the first author. Then, we collected the names of other authors as well. Using the set properties of union, we combined all authors and gave them a unique identifier. The overall statistics on the number of authors and the number of collaborations between them are provided in Table II.

To create an author collaboration network, we used Gephi [25], which is an open source software for visualizing different kinds of networks. The results for the author’s collaboration are shown in Fig. 9. It can be seen from Fig. 9 that most of the authors (97.5% of the 1682 authors) have collaborated with each other only in one paper. The links for the one time collaboration are shown in light pink color. There are 2.33% authors who have collaborated in two papers and 0.13% of the authors who have collaborated in three papers. The maximum number of collaborations is 4, i.e., the authors have collaborated together in 4 paper (as shown in red color) which is 0.03% of the total authors. Author “Neeraj Kumar” has published a maximum of 7 research papers.

H. Domains of highly cited papers

We have reported the statistics on top 10 most cited papers, their domain, and the number of citations received till December 2020. The results are shown in Table III. It can be seen from that highly cited paper is on the “challenges and opportunities” of the blockchain domain.

![Word cloud](image)

Fig. 10: All key words occurred in all years (2017 to 2020).

I. Topic Keywords

We extracted the keywords from the title string of all papers to identify the key domains or topics. We have removed all possible stop words like on, to, for, etc. to collect possible keywords in all papers. Also, we have combined the words which belong to same domain, e.g., internet-of-things and all of its variants to “iot”. Similarly, the keywords like patient, medical, electronic medical records are replaced with “healthcare”. The word cloud representation is shown in Fig. 10. A word cloud is a pictorial representation of showing frequency of a specific word. In a word cloud, the word size represent word frequency, i.e., the larger the word, higher the frequency. For example, “iot” occurred in the topic string of most of the survey papers with 85 occurrences. Similarly, “security” occurred 58 times in the topic strings. It can be inferred from Fig. 10 that most of the topic keywords are representing the same phenomenon as seen in Section IV-F, i.e., the more prominent keywords belong to the same classes of surveys.
V. Conclusion

We have conducted a bibliometric analysis, processed, and examined 542 survey papers on the topic of blockchain from 2017 to 2020. We have identified the publications with respect to the publication type, publication venue, references, citations, paper size, different areas of knowledge, year, countries, authors, and their collaboration. We found out that there are more journal surveys than a conference each year. Almost 58% of the papers contain more than 40 references and 73% of the papers have referenced the papers in the range of 10 to 100. Almost 30% of survey papers have no citations. There are 157, 177 and, 208 short sized, medium-sized and large-sized papers, respectively. A more diverse and maximum number of survey papers are published in IoT. Statistics on countries show that 68 countries have published the survey papers and 33% of the countries have only published one research paper. We also found out that there are 1682 total authors and almost 97.5% of the authors have only one research paper.

REFERENCES

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” Decentralized Business Review, p. 21260, 2008.
[2] D. Shrier, W. Wu, and A. Pentland, “Blockchain & infrastructure (iden- tity, data security),” Massachusetts Institute of Technology-Connection Science, vol. 1, no. 3, pp. 1–19, 2016.
[3] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, “Medblock: Efficient and secure medical data sharing via blockchain,” Journal of medical systems, vol. 42, no. 8, pp. 1–11, 2018.
[4] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. Leung, “Blockchain and machine learning for communications and networking systems,” IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 1392–1431, 2020.
[5] H.-W. Kim and Y.-S. Jeong, “Secure authentication-management human-centric scheme for trusting personal resource information on mobile cloud computing with blockchain,” Human-centric Computing and Infor- mation Sciences, vol. 8, no. 1, pp. 1–13, 2018.

[6] [Online]. Available: https://scholar.google.com
[7] A. Firdaus, M. F. Ab Razak, A. Feizollah, I. A. T. Hashem, M. Hazim, and N. B. Anuar, “The rise of “blockchain”: bibliometric analysis of blockchain study,” Scientometrics, vol. 120, no. 3, pp. 1289–1331, 2019.
[8] A. N. Guz and J. Rushchitsky, “Scopus: A system for the evaluation of scientific journals,” International Applied Mechanics, vol. 45, no. 4, p. 351, 2009.
[9] Y.-M. Guo, Z.-L. Huang, J. Guo, X.-R. Guo, H. Li, M.-Y. Liu, S. Ezzed- dine, and M. J. Nkeli, “A bibliometric analysis and visualization of blockchain,” Future Generation Computer Systems, vol. 116, pp. 316–332, 2021.

[10] C. Analytics, “Web of science,” 2017.
[11] C. Chen, “Citescapes ii: Detecting and visualizing emerging trends and transient patterns in scientific literature,” Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 359–377, 2006.
[12] N. J. Van Eck and L. Waltman, “Software survey: Vosviewer, a computer program for bibliometric mapping,” scientometrics, vol. 84, no. 2, pp. 521–538, 2010.
[13] S. X. Ni, Y. Yuan, and F.-Y. Wang, “A bibliometric analysis of blockchain research,” in 2018 IEEE intelligent vehicles symposium (IV). IEEE, 2018, pp. 102–107.

[14] [Online]. Available: https://www. elsevier.com/solutions/engineering-village/content/compendex
[15] [Online]. Available: https://global.cnki.net/kns/brief/result.aspx?dbprefix=SCDB
[16] L. Ante, “Smart contracts on the blockchain–a bibliometric analysis and review,” Telematics and Informatics, p. 101519, 2020.
[17] I. Meredia-Solà and A. F. Bariviera, “A bibliometric analysis of bitcoin scientific production,” Research in International Business and Finance, vol. 50, pp. 294–305, 2019.