A Cross-Disciplinary Review of Blockchain Research Trends and Methodologies: Topic Modeling Approach

Muhammad Nauman Shahid
National University of Singapore
nauman@comp.nus.edu.sg

Hahn Jungpil
National University of Singapore
jungpil@nus.edu.sg

Abstract

Given the increasing interest in blockchain technology, we present a large-scale cross-disciplinary literature analysis of research on the blockchain using topic modelling with the goal of identifying the major research trends, research methodologies, and fruitful areas for further research. In particular, the analysis focuses on abstracting out research trends from relevant terms and topics related to the research disciplines of Business, Computer Science, Economics, Social Sciences, Engineering, Healthcare, and Law. A total of 2,125 articles published between 2008 to up until early 2019 in academic journals and conferences were analyzed. Results of our analysis reveal that research is bipartite between practical and research domains, with academic research on blockchain not clearly aligning with organizational and social benefits. Also, we found – 1) few inter-disciplinary publications, and 2) a small number of studies that use surveys, experiments, and case studies as their research method. Our findings also reveal that research on Blockchain in the social sciences and law is still in the embryonic stage, thus making it essential to develop more direct research efforts for Blockchain to thrive in all research disciplines.

1. Introduction

Blockchain technology has received exceptional attention in both business and academic circles as supporters argue that it constitutes the foundation for truly trust-free economic transactions due to its unique technological characteristics [1]. The blockchain is a decentralized, and immutable digital record system that is shared among many independent parties and can be updated only by their consensus [2]. It acquired fame as the underlying technology for Bitcoin that upraised its expansion to other functional applications making it the most trending technology that has the potential to disrupt various intermediary services.

The use cases of blockchain are not well understood [1]. On the one hand, researchers are drawing parallels between blockchain technology and, for example, the bubble memory regarding its groundbreaking impact on social and business circles, recalling that bubble memory is short-lived to the prospects linked with it [3]. On the other hand, the compatibility issues with existing technologies in different business and functional contexts have perpetuated a variety of hearsay about the potential usefulness among domain experts. The paucity of knowledge and interdisciplinary nature of fundamental concepts further exacerbates the realization of its usefulness. We argue that alignment of interdisciplinary research can improve usage clarity by reducing cross-disciplinary limitations of knowledge that impede blockchain's expansion. Therefore, the objective of this paper is to examine the existing cross-disciplinary body of literature on blockchain and report current research trends and research methods.

By drawing on extant interdisciplinary academic literature published between 2008 to early 2019, we seek to organize the findings to address our research question: What is the current state of research in blockchain in different research disciplines, what correlations exist between the content, research topics, research methods, and how can blockchain research purposefully be advanced?

To achieve this objective, we retrieved 2,125 academic articles by searching with the keyword "blockchain" on six major databases (i.e., the ACM Digital Library, IEEE Xplore, JSTOR, Science Direct, Scopus, and Web of Science). We then used unsupervised clustering to interpret topics and use those topics as an anchor to discuss interesting temporal patterns and correlations among articles, research disciplines, and different research methodologies used in the blockchain literature. In particular, we show how research on the blockchain has changed over time within each research discipline, what inter-topic correlations and temporal relationships exist in the content and their interdisciplinary significance, as well as what future research trends can be established.

The remainder of the paper is structured as follows. First, we discuss related work that reviewed blockchain literature. Second, we describe the process of data collection and literature analysis. Third, we deduct research topics from the literature and interpret trends.

URI: https://hdl.handle.net/10125/64237
978-0-9981331-3-3

Lastly, we discuss correlations and relationships between content, research disciplines, and research methodologies applied in blockchain research.

2. Related work

The emergent nature of blockchain and its practical suitability has aggravated difficulties of understanding potential research constructs. [4] presented a systematic literature review of 41 peer-reviewed articles published up until 2015. However, 80% of the articles in their corpus examined the usage of blockchain as a protocol for the Bitcoin cryptocurrency. They extracted several features from the abstracts and classified the literature into five primary topics: security; wasted resources; usability, privacy and smart contracts; cryptocurrencies; and trustworthiness. Their findings point to uneven focus of the literature on the aspects of usability and wasted resources.

There are several limitations in their review. First, they used cryptocurrencies as an anchor to discuss the technical perspectives of the blockchain protocol. Given that using cryptocurrency as an application of blockchain may prove a good starting point for such a review, however, specific technical issues of the protocol – such as privacy, security, performance, and scalability – limits the reliability of their findings from the evolutionary aspects of the blockchain protocol. Second, the flexibility of protocol to use cases beyond cryptocurrencies may have different implications for different research domains such as economics, law, and business. Therefore, our review analyzes blockchain research from multiple research disciplines and reports a broader perspective.

[5] presented a research framework using multidisciplinary content analysis on a corpus of 69 articles. They used main databases (i.e., Web of Science, IEEE Xplore, AIS Electronic Library, and Science Direct) to identify relevant research articles. They focused beyond the technical aspects of the blockchain protocol by identifying conceptual papers that discuss technology for humans, organizations, and markets. However, this review was conducted at a time when a substantial number of non-technical research papers – such as in social science, law, economics, and business disciplines – were in review stages; therefore, we argue that publication of articles outside the context of computer science skyrocketed most recently and given the dynamic development of blockchain protocol in recent years, a fresh review is needed. Additionally, the predominant outlets of computing research used in their review do not fully encompass research trends across different disciplines. Also, parts of the developed research questions were formed by help from blockchain developers of computing origin which creates room for bias in interpreting research from other non-technical research disciplines from a technical perspective. That said, during that time, the focus of research remained largely on cryptocurrencies (i.e., Bitcoin) due to its prominence as the dominant application of blockchain. This makes it further challenging to clearly identify the research intersection across different research disciplines given high levels of knowledge paucity at the time. Therefore, the findings of that review are somewhat analogous to [4].

The extant reviews of the literature have used highly manual systematic and structured techniques [4, 5, 6, 7] that may not be suitable for examining a large collection of articles that focus on the literature across multiple disparate disciplines. Additionally, the highly manual approach is costly, resource and labour-intensive, and requires excessive efforts to develop themes especially in the absence of domain experts who can effectively interpret the meaning of derived themes.

Many use citation analyses for literature reviews [8] to infer author-specific influences in a given research domain, however, it is criticized for its inability to capture synthesis of the content [9]. Also, when the amount of the academic literature exceeds a manageable size, manual techniques and citation analysis become impractical. Therefore, we adopt a more practical unsupervised topic modelling approach to uncover trends from a large corpus of blockchain research across multiple disciplines. In particular, we use Latent Dirichlet Allocation (LDA) [10] to inductively identify topics from the corpus of research articles. LDA is a widely used topic modelling technique that aims to annotate large sets of documents with thematic information [10, 11], and it can be used to automatically extract topics that can summarize the underlying themes in these documents [12]. This approach has been recently used in the management literature to identify risk types from risk disclosure texts [13] and to analyze leadership themes from corporate vision statements. In fact, LDA has been increasingly used for semantic analysis in Information Systems research.

3. Materials and method

3.1 Literature selection

To collect the relevant publications for review, we set our focus on finding the articles that used the keyword “Blockchain”, between 2008 and early 2019. We searched the databases of the ACM Digital Library, IEEE Xplore, JSTOR, Science Direct, Scopus, and Web of Science for articles published in conferences and journals. In sum we found approximately 3,285 publications and to ensure the quality of selected articles, we discarded all duplicate entries. For entries published both in a conference and a journal, a similarity comparison was performed using simple spreadsheet functions. The articles that showed 99% similarity to another article in our corpus were manually examined

and we retained the journal entry. In the end, we also performed a full manual check on all entries to ensure the quality and retention of single entry of an article. Working papers, papers in workshop proceedings, and articles published in regional languages were also discarded to keep only published academic research in scholarly journals and conference proceedings.

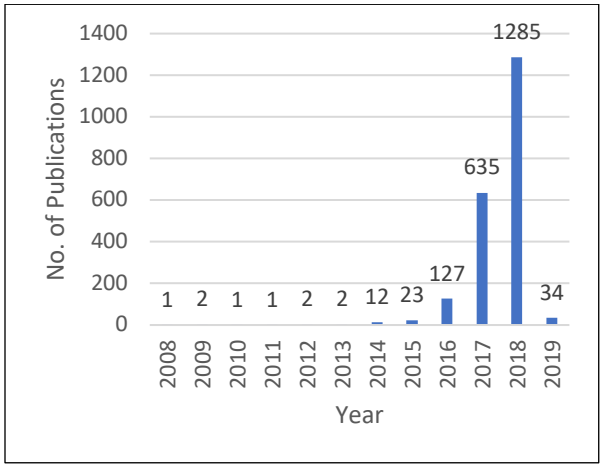

Figure 1. Distribution of publications from 2008-2019

The final database consisted of features such as the authors' name, the title of the paper, keywords, abstract, reference type (e.g., conference or journal), . We concatenated titles and abstracts of all retained articles and created a corpus of 2,125 entries in which each row representing a combination of title and abstract called a "document". The length of words in each document after concatenation was kept at original to ensure the preservation of the context. Figure 1 shows the yearly distribution of publications from 2008 to 2019.

3.2 Data analysis procedure

We used Latent Dirichlet Allocation (LDA) on the corpus to interpret the content of each entry [10]. To develop our LDA algorithm from scratch, we selected python's open-source Anaconda [14] distribution due to its ease of use and facility to develop and execute the LDA algorithm in the browser using Jupyter [15] notebook. The LDA algorithm outputs topics that represent words based on their probability of relevance to each topic. LDA supports two types of transformation for the selection of words – 1) the bag of words (BoW) model, and 2) the term frequency-inverse document frequency (TF-IDF) model. We choose the TF-IDF model because it has all the features of the BoW model as a preset and it is believed that the quality of its output is superior to the BoW model [16].

[5] has shown a high number of computer science publications on the topics of blockchain that discursively signals repetition of similar words. The occurrence of similar words during the execution of the LDA algorithm may obscure the probability of inclusion of less frequent words. Therefore, we set the exclusion criteria to 70% in the parameters of the LDA algorithm for frequent words (e.g., Blockchain, Bitcoin, Cryptocurrency etc.) to ensure the balanced treatment to less frequent words[1]. We also set the values of the parameter "chunk size" to 10% of the corpus size to allow loading of the full dataset in the memory to ensure optimal results during each recursive execution.

Table 1. Identification tags for research disciplines

Discipline	Tags
Business	Ledger, Volatility, Finance, Marketing, Management
Computer Science	Information, Software, Privacy, Security
Economics	Tokenomics, Currency, Finance, Economy
Engineering	Software, Energy, Scalability, Digital
Healthcare	Medicine, Nursing, Health
Law	Regulate, Tort, Legislature, Privacy
Social Science	Sociology, Public, Relationship, Culture, Society

To associate each publication to a research discipline based on its relevance, we used two-step criteria. First, based on the name of the journal and conference, we manually assigned it to the relevant research disciplines. Second, for publications where relevance is difficult to comprehend manually or that may belong to two different disciplines (e.g., Business and Information Systems, Engineering and Computer Science), we used relevance criteria set out by Google's academic graph. Google assigns tags to each publication based on multiple factors such as the relevance of journal or conference to a research discipline, the occurrence of specific keywords in the content, authors, and affiliations to academic institutions (see examples in Table 1). We created a dictionary of tags by searching keyword "blockchain" in Google's graph and collated all tags associated with resultant publications. In the next step, we determined the frequency of occurrence of a tag for each research discipline. Later, we used TF-IDF to assign each entry in our corpus to a research discipline where the similarity index was greater than 70%. For entries that have similarity index lower than 70%, we assigned them to research disciplines after

[1] LDA uses two parameters "no_below" and "no_above" to allow flexibility of removing certain words that reach specific threshold. We selected a value of "40" as the minimum inclusion criteria for a keyword. The model returned a fair coherence value of 31.8% at 40, below which the coherence values are too low and word to topic coherence is difficult to interpret.

manually examining the content of each publication. Table 1 shows an example of tags from our dictionary and their relevance to each research discipline. Ultimately, we abstracted seven categories of research disciplines – i.e., Business (Biz.), Computer Science (C.Sc.), Economics (Econ.), Engineering (Engg.), Healthcare (HC.), Law, and Social Science (S.Sc.).

In the final step, we performed a qualitative analysis of post-LDA results. We discuss our preliminary findings in the subsequent section.

4. Results

In this section, we discuss our preliminary results based on the outcome of LDA topics, LDA topic keywords, and interrelating temporal trends and research methodologies. To determine the research methodologies applied in different studies, we manually examine the content and classified them in Table 3.

4.1. Interpretation of general trends from LDA topics

Table 2 shows eight LDA topics represented by keywords and classification fitness score. Each keyword has a high to low probability from left to right in each row. The inter-topic coherence value of our model with eight topics is 32% which is reasonable as the value of exclusion of frequent words is set to 70%, which inevitably increases the perplexity of the model. We tested our model on the same corpus to examine its ability to correctly classify the words and it returned 87.9% classification fitness score. Classification fitness score shows a model's ability to correctly classify a corpus. A higher score shows a good model.

The results show 87.9% prominence for Topic 8 (Computing, Security, Data) as shown by the classification fitness score in Table 2 and publication trends 2015 onwards in Figure 2. The latent analysis of the probability of words and their associations to 87% of the publications signal research focuses on the computer science and design aspects of blockchain and its protocol features. Topic 2 (Protocol, Services, Develop) is closely associated with Topic 8 because researchers used assessment-based techniques and fuzzy methodologies to clarify the fitness of blockchain protocol to existing organizational processes and services.

Figure 3 shows a surge in a number of publications from 2015 onwards in all protocol level aspects of blockchain. In particular, a low (approx. 284) number of publications on user-level aspects show a decline in the year 2019. Perhaps, aspects of data security and efficiency aspects of transactions remained a focus of research under Topic 8. However, aspects of privacy from the user's perspective and its implications from the context of Law and Social science is given less focus. The issues of data storage and integration of organizational processes particularly between permissioned and permissionless blockchains are relatively understudied as shown by the content of publications under Topic 2. Also, the mechanisms to undo committed computing operations (such as a smart contract) are least investigated in Law and Computer science.

Topic 7 (Society, Privacy, Trust) discuss various aspects of society and theorizes trust and its value by using case studies and cognitive theories. Many publications show unclear focus on the dissemination of key blockchain features in society and whether they would enhance trust and privacy experience of potential users or would it deteriorate over time. Furthermore, many researchers argue that mechanisms enabling on-chain and off-chain trust to be vague, practically difficult to understand, and unevenly associated with blockchain protocol for its enablement. Furthermore, the case study methodology is used in approximately 20 publications and does not clearly comprehend the usefulness of blockchain and its applications. The theoretical instruments used to measure the usefulness of blockchain applications in social sciences are weak and a major overhaul is required.

The ambiguity that whether blockchain applications can easily be adopted by society or not is not strongly supported. The main underlying reasons are the weak focus of research on the user level perspectives of blockchain that further enable understanding of trust, privacy, and security from a variety of Economics and Business perspectives. For example, referring to the Figure 3 of Topic 8, the keywords "user" and "person" appeared in approximately 6 publications and all of them were not fundamentally associated to research disciplines of Social science, Business (Information systems), and Computer science. Therefore, researchers need to seriously examine the deliberate assumptions that people would approve the many "trust-free" characteristics promised by blockchain technology. The weakest word in the topic is "transact" that shows a 0.6% association to the topic and shows fewer publications that examined the aspects of blockchain transactions and implications of both committed transactions and their reversal in case of disputes from the perspectives of Law.

Topic 6 (Business, Blockchain Types), Topic 5 (Mining, Incentives), and Topic 1 (Finance, Novelty, Disruption) are interrelated and show key areas of enquiry for researchers of business aspects, finance, and economics. The classification fitness score for all the three topics was 1.7% and the research focuses on the aspects of design compatibility with different organizational aspects, such as inter-organizational processes and transactions, incentive configurations for internal and external organizational actors, implications

Table 2. LDA topics

LDA Topic No.	LDA Topic Title	LDA Topic Keywords	Score
1.	Finance, Novelty, Disruption	citi, refer, disrupt, sector, revolut, explain, finance, fund, token, definit	0.0172
2.	Protocol, Services, Develop	service, onlin, consensus, protocol, algorithm, account, trust, traceabl, product, general	0.0172
3.	Energy, Market, Trade	energi, electr, grid, search, coordin, market, trade, power, demand, consumpt	0.0172
4.	Measure, Utility, Economics	rat, accuracy, refer, decreas, credit, economi, employ, entity, good, learn	0.0172
5.	Mining, Incentives	mine, miner, game, reward, spend, doubl, adversary, incent, strategi, accuraci	0.0172
6.	Business, Blockchain types	enterpris, messag, argu, avoid, permiss, provid, include, give, play, publish	0.0172
7.	Social, Privacy, Trust	thing, social, internet, devic, collabor, preserv, privacy, service, trust, mitig	0.0172
8.	Computing, Security, Data	data, bitcoin, person, secur, decentr, trust, transact, user, comput, privaci	0.8796

of on-chaining and off-chaining, and environmental consequences of mining and scalability of blockchain systems. In particular, we found that approximately 30% of research in the outlets of Business and Economics discussed incentivization schemas and latent mechanisms particularly under what consensus mechanisms different incentivization system can be developed and how different blockchains can be merged?

Topic 4 (Measure, Utility, Economics) has a classification fitness score of 1.7% and shows a high relevance to Economics research. The number of publications is lower and shows two schools of thought – 1) proponents who argue that the monetary value in a blockchain is fixed such that the blockchain can provide a check on unusual inflation, unlike less trustworthy state banks; and 2) maximalists who propose that forks in a blockchain are equivalent to inflation that lead to orphaned blocks and persistent deviation between chains. In particular, extant research has focused developing closed-form formulas of the fees and latency of Bitcoin processing and other properties; comparing blockchain payment systems to that of traditional payment systems; and suggesting modifications in the protocol design to improve efficiency. However, less attention is given to understand why cryptocurrencies are a favourable alternative to existing payment systems. Therefore, future research needs to focus on why using cryptocurrencies is a viable substitute, and when it may not be a good alternative. Given unclear association of publications of Topic 4 to fundamental research disciplines such as Business, Economics, and Computer science, collaborative research between Business (particularly Information systems) and Economics, and between Economics and Computer science, therefore, can help to curb technological pitfalls and provide a better understanding of emergent phenomena (e.g., Tokenomics) from user's perspective.

A small number of publications discussing aspects of Initial coin offerings (ICO) in allied disciplines of Computer science and Business (such as Information systems) may provide a fruitful avenue for empirical investigation particularly related to crypto-assets using asset pricing theory. Therefore, we believe that empirical investigations related to ICO's will help to normalize the prevalent anomalies in regularizing crypto-assets as an investment vehicle. Because regulatory approvals are often given when there is a level playing field in the crypto asset markets; otherwise, regulators will not favour them.

4.2. Cross-disciplinary analysis, trends, and notes on research methodologies

In this section, we adopt the framework of [17] that uses three-dimensional criteria to classify all research publications across seven research disciplines and eight research methodologies as shown in Table 3. The relevance of research methodology for each publication is manually examined and sorted across – 1) concepts and design, 2) theoretical or empirical, 3) and unclear publications. We identified 26 publications that use unclear research methods whereas; Healthcare, Law, and Social science researchers have lowest publications that use conceptual and simulation methods.

In the dimension of concepts and design, the Law and Economics have the least number of conceptual publications whereas; Business, Computer science, and Engineering dominate in this dimension. The research has focused on the fundamental concepts of blockchain and its principal mechanisms that could harness its adoption in business organizations. For example, technology providers such as Microsoft, Oracle, SAP, and IBM are interested in the relative importance of privacy, security, usability, and latency to determine the plausibility of end-user adoption. However, given the end-users as key actors of the blockchain network, perspectives of security, privacy, and generic assumptions about trust-free characteristics of blockchain are examined by social science research in approximately 70 publications.

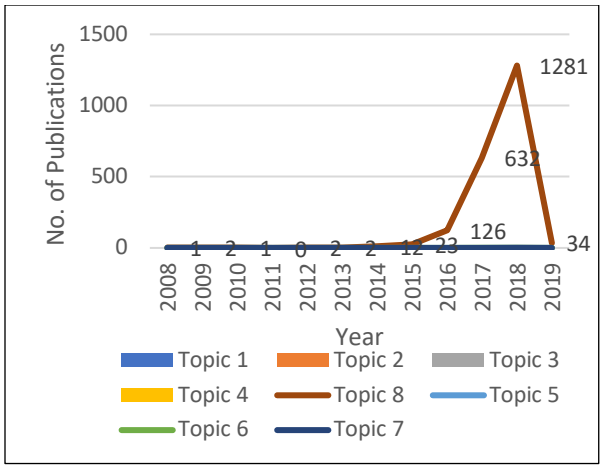

Figure 2. The prominence of topics from 2008-2019

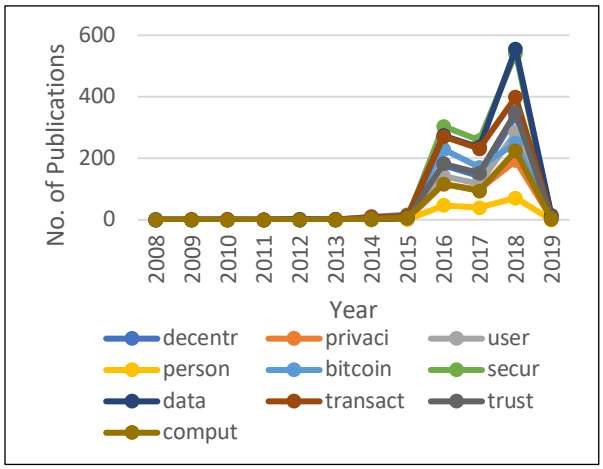

Figure 3. The trend of Topic 8 from 2008-2019

Computer science research has shown tremendous growth of publications since 2008 especially for Topic 8 as shown in Figure 3. Topic 8 encompasses several dimensions of a blockchain protocol, privacy and security issues, and user-level aspects. Although, Table 3 shows that Computer science researchers have used all key research methodologies, however, applied Computer science researchers such as those in Information systems have bridged Computer science and Business research by approximately 21 publications that use survey methods, case studies, and experiments.

In Business discipline, approximately 200 publications mimic integration between established systems and blockchain-based solutions by using design science and prototyping approach. However, performance evaluation of such integrations is rarely considered and is an excellent area for future researchers of Information systems and Business disciplines.

Engineering researchers have approximately 238 publications that use design science or prototyping methodology. The main research areas are scalability issues of the apparatuses and device-level validations. In particular, 29 publications discuss the interoperations of a blockchain network with IoT devices. Whereas, given the issues of scalability, consensus mechanisms on integrated blockchain networks of IoT and transaction risks and their implications are interesting future research areas.

Table 3. Research methods of publications and distribution across seven research disciplines

Criteria	Research Methodology	No. of Relevant Publications	Distribution Across Disciplines						
			Biz.	C.Sc.	Econ.	Engg.	HC.	Law	S.Sc
Concepts and Design	Conceptual	766	244	244	34	140	-	34	70
	Patents/Des. Sc./Prototype	1265	200	764	25	238	13	25	-
	Literature review	8	2	3	1	1	1	-	-
Theoretical/Empirical	Case study	17	7	6	1	1	1	1	1
	Simulation	10	1	7		2	-	-	-
	Survey	29	5	1	2	3	1	1	2
	Experiment	4	-	1	-	2	1	-	-
Unclear	Other	26	-	-	-	-	-	-	-

We found 9 publications that are synergistically aligned with Social science and Economics based on identification tags assigned to each publication as shown in the example of Table 1. The researchers used conceptual and prototyping methodologies to assume initial consequences of blockchain and its design implications, however, the value configurations for different intermediate actors and economic impacts have been least investigated. In particular, one school of thought assumes blockchain to have unclear and temporary value as compare to established processes. Therefore, more direct research efforts are required to investigate the value of blockchain and its integration cost into existing systems. The tradeoffs between anonymity and transparency discussed in approximately 10 publications ultimately question the fitness of scales that can measure value incentives and points to an interesting area for future investigations.

There are approximately 13 publications in Healthcare that uses conceptual or design science approaches to understand the delineation of electronic health records from established systems into a

blockchain network that enables automatic validation of new records and upholds privacy and security of data.

Similarly, publications in the outlets of Law focuses on the governance implications of a public and private blockchain network and arbitration in case of transactional disputes on a highly decentralized network. In particular, it is unclear for which intermediaries public or private blockchain systems constitute a threat or opportunity.

5. Conclusion

This study analyzed research trends and methodologies applied within the blockchain research community. In particular, it examined a corpus of 2,125 articles from 2008 to early 2019 that used the keyword "blockchain". We applied LDA on the corpus to find topics that could interpret the entire blockchain research ecosystem at a high level. The results reveal that (naturally) Computer science dominates the research followed by Economics and Business. Precisely, Computer science research focuses on the protocol development aspects of blockchain whereas, Business and Economics studies business suitability and integration in existing processes by using a high level of conceptual and design methodologies. Meanwhile, Information systems research bridges the gap between Computer science, Social science, and Business by supplementing the literature with qualitative surveys, experiments, and case studies. Blockchain research is different than traditional research because objectives for usual research within a research discipline is clear and easy to follow up. However, for blockchain, it is highly interdisciplinary phenomena and strategic alignments of research are the utmost necessary for it to flourish. Therefore, this paper fills this gap by reporting research trends and identifies potential areas for collaboration.

6. References

[1] F. Glaser, "Pervasive Decentralisation of Digital Infrastructures: A Framework for Blockchain enabled System and Use Case Analysis," in *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, 2017.

[2] M. Avital *et al.*, "Jumping on the Blockchain Bandwagon : Lessons of the Past and Outlook to the Future," 2018.

[3] M. Avital, "Peer review: Toward a blockchain-enabled market-based ecosystem," *Commun. Assoc. Inf. Syst.*, vol. 42, no. 1, pp. 646–653, 2018.

[4] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, "Where Is Current Research on Blockchain Technology?-A Systematic Review," *PLoS One*, vol. 11, no. 10, p. e0163477, 2016.

[5] M. Risius and K. Spohrer, "A Blockchain Research Framework," *Bus. Inf. Syst. Eng.*, vol. 59, no. 6, pp. 385–409, 2017.

[6] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," *Telematics and Informatics*, vol. 36. Elsevier Ltd, pp. 55–81, 01-Mar-2019.

[7] S. Seebacher and R. Schüritz, "Blockchain technology as an enabler of service systems: A structured literature review," *Lect. Notes Bus. Inf. Process.*, vol. 279, pp. 12–23, 2017.

[8] S. P. Nerur, A. A. Rasheed, and V. Natarajan, "The intellectual structure of the strategic management field: An author co-citation analysis," *Strateg. Manag. J.*, 2008.

[9] M. H. MacRoberts and B. R. MacRoberts, "Problems of citation analysis: A critical review," *J. Am. Soc. Inf. Sci.*, 1989.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[11] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, 2010.

[12] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, "Text Mining in Organizational Research," *Organ. Res. Methods*, 2018.

[13] Y. Bao and A. Datta, "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures," *Manage. Sci.*, 2014.

[14] wikipedia, "Anaconda (Python distribution)," *wikipedia*. 2019.

[15] P. J. Jupyter Community, "Project Jupyter," *Project Jupyter*, 2018. .

[16] I. Abu El-Khair, "TF*IDF," in *Encyclopedia of Database Systems*, 2017.

[17] S. Aral, C. Dellarocas, and D. Godes, "Social media and business transformation: A Framework for research," *Inf. Syst. Res.*, vol. 24, no. 1, pp. 3–13, 2013.