Data Security and Privacy Research Trends: LDA Topic Modeling

Bin Zhao[1], Jie Zhou[2*]

[1] School of Public Administration, Hubei University, Wuhan, China
[2] Department of Professional Competence Education, Ningxia Polytechnic University, Yinchuan, Ningxia, China
[*] Corresponding author: 15779047054@163.com

Abstract: With the rapid advancement of big data technologies, the need for robust data security and privacy measures has intensified. Big data technologies have revolutionized the collection and analysis of a vast volume of research literature, offering unparalleled avenues for scholarly inquiry. Identifying prevalent research topics and discerning developmental trends is paramount, especially when grounded in an expansive literature base. This study examined abstracts and author keywords from 4,311 pertinent articles published between 1980 and 2023, sourced from the Web of Science core collection. The content of abstracts and author keywords underwent LDA theme modeling analysis. Consequently, two predominant research topics emerged: security and privacy measures for mobile applications and ensuring security and information integrity for big data within the Internet of Things framework. The LDA model proficiently pinpoints these salient topics, assisting researchers in comprehending the current state of the domain and guiding potential future research trajectories.
***Keywords*:** data security; privacy; LDA topic modeling; topic trends; word cloud.

1. Introduction

In the modern information age, data has evolved as a pivotal factor in production, drawing significant attention to issues surrounding data security and privacy [1]. These two domains have emerged as paramount research areas within information technology [2], leading to an abundance of literature that addresses data security and privacy from diverse perspectives [3]. Given the proliferation of such research, it becomes imperative to categorize the studies and delineate the evolving trends in this domain.

The primary goal of data security and privacy protection is to shield data from threats such as unauthorized access, tampering, destruction [4], and leakage while also preserving individual privacy against illicit access and misuse [5]. These two areas are intrinsically intertwined; a deficiency in data security can result in privacy breaches, and conversely, inadequate privacy safeguards can compromise data security. Contemporary advancements like the Internet of Things (IoT), smart cities, digital transformation of enterprises, and the burgeoning digital economy have catalyzed exponential data growth. While this proliferation can potentially generate substantial value, it concurrently amplifies risks, including unauthorized intrusions, data breaches, and violations of privacy. Addressing challenges in data security and privacy becomes indispensable in our increasingly digitized society [6].

Building on prior research, this study collated 4,311 papers on data security and privacy published between 1980 and 2023 from the Web of Science core collection. The abstracts of these papers underwent TF-IDF keyword extraction, culminating in a comprehensive keyword canon. Subsequently, a word cloud based on this canon was generated for further analysis. Integrating topic modeling with word cloud analytics, the study dives deep into the textual content, yielding two distinct topics. The overarching goal is to delineate prevailing trends in data security and privacy research, offering insights for future studies. A flowchart of the dataset acquisition and analysis method is shown in Figure 1. The scenario is divided into three sub-processes: data retrieval and preprocessing, word cloud study, and topic analysis.

2. Data and methods

The Web of Science Core Collection, a vast repository of academic literature spanning various disciplines globally, was chosen due to its high-caliber content, ensuring rigorous research outcomes. The data retrieval was conducted on April 11, 2023, focusing on papers and reviews. Drawing on comprehensive knowledge of data security and privacy, an advanced search was executed with the following string:

(TS= (“data security” OR “data safety” OR “date security” OR “data safe” OR “data secure” OR “information security” OR “database security”)) AND TS= (“privacy” OR “privacy-preserving” OR “personal secrets” OR “confidentiality” OR “secret” OR “privacy protect” OR “privacy information” OR “rights of privacy”). Here, TS represents “topic search,” typically signifying the appearance of the search term in the title, abstract, or keyword, guiding the retrieval process. Given that

literature on this subject emerged around 1980, the search spanned from 1980 to 2023, yielding 4,311 papers. Post-retrieval, 313 papers, such as book reviews and editorials, were deemed irrelevant and discarded, leaving 3,998 papers for the study. Figure 2 illustrates an increasing trend in the annual publication volume of pertinent literature.

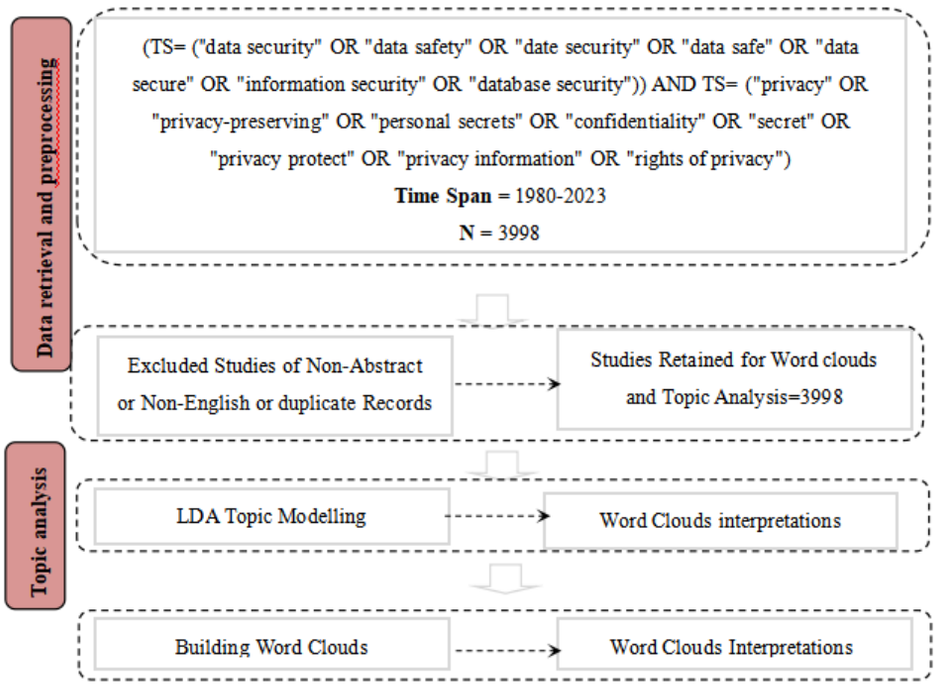

Figure 1. Flowchart of Dataset Acquisition and Analysis Methodology

As depicted in Figure 2, research publications on data security and privacy experienced gradual growth until 2017. Between 1980 and 2010, the annual publication count remained below 50 articles. However, a significant uptick in publications commenced in 2011, with particularly explosive growth observed from 2017 to 2022. This trajectory underscores the escalating significance of data security and privacy in contemporary society. With the expanding reliance on digital technologies, the sheer volume of data transactions has surged, leading to occasional data breaches and privacy infringements. Such incidents have heightened public concern, fueling increased attention from both the academic and industrial sectors to these issues.

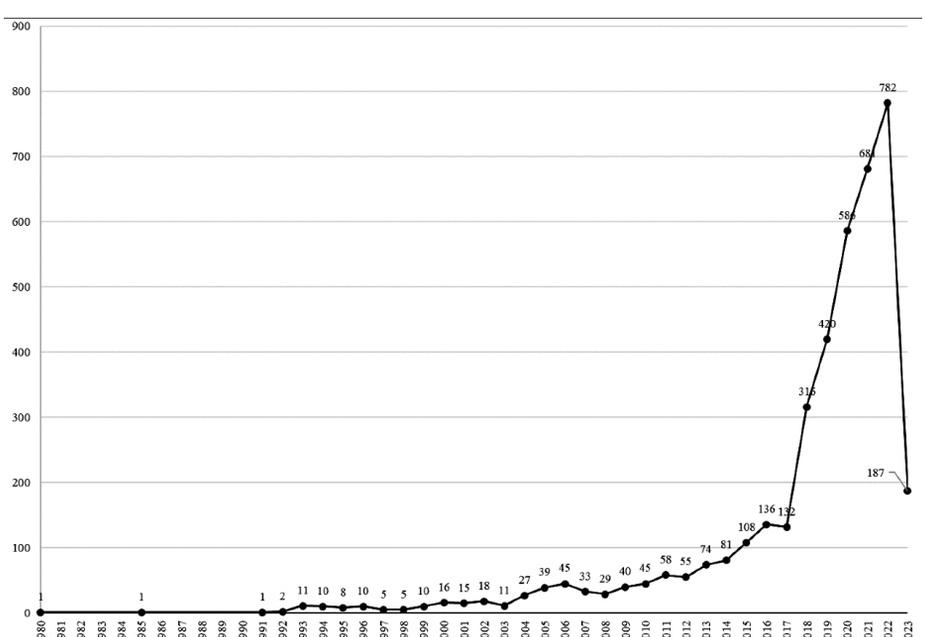

Figure 2. Annual Scientific Production Per Year (1980–2023)

This research utilizes published literature for thematic exploration and analysis. An essential step preceding this analysis involves preprocessing the bibliographic data. This phase encompasses the elimination of common English stop words, such as “we” and “what.” Additionally, synonyms like “data security” and “security” are harmonized, and compound terms such as “cloud computing” and “cloud” are consolidated. Natural language processing (NLP) tools are employed to extract thematic terms from abstracts and author-provided keywords within the bibliographic entries [7]. To derive a more nuanced set of keywords from abstracts, author keywords from the initially curated corpus are extracted and processed to form a tailored lexicon. Ultimately, the preliminary corpus encompassed 5,571 keywords across 3,998 documents.

3. Results

Utilizing Python’s gensim library, the preprocessed abstract text underwent LDA model training. Each term was ranked based on its frequency, starting from the most prevalent to the least prevalent, serving as indicators of the topic’s essence. Topics were discerned by amalgamating these terms. After LDA model training, we discerned the distribution of “topic terms.” Following this, we embarked on topic detection and characterization based on this distribution. The evaluation of topic content hinged on the terms linked to each topic, with emphasis placed on the predominant terms for assessment. Topics were defined based on the detection results, and any that echoed redundant themes or exuded nebulous content were excluded. The results of topic identification are encapsulated in Table 1.

Table 1. Results of Topic Modeling

Topic-1	Topic-2
data	iot
model	big_data
privacy	model
blockchain	information_security
patients	system
privacy-preserving	authentication
framework	image
research	algorithm
cloud	research
data_security	watermarking
healthcare	information
information	challenges
performance	integrity
federated_learning	

4. Hot Topic Analysis

4.1 Information Security and Privacy Protection in Mobile Applications

In the current digital epoch, ensuring information security and privacy protection within mobile applications has emerged as a pressing issue requiring immediate and robust solutions. Figure 3 illustrates the expansive adoption of mobile devices and applications, highlighting the paramount importance of information security, privacy, cloud-facilitated user experiences, and consumer interactions. Simultaneously, the swift progression of mobile applications has ushered in an escalating reliance on these platforms to process sensitive personal data and facilitate various online activities. Consequently, validating the data security and privacy safeguards of mobile applications is imperative. Mobile applications encompass users’ private information, not limited to names, addresses, telephone numbers, and bank account details. The mismanagement or malicious exploitation of this information can lead to severe consequences, such as identity theft and financial fraud, among other abuses. Thus, it is obligatory for mobile application developers to implement stringent security measures to safeguard user data. For instance, Ali Balapour leverages the Communication Privacy Management Theory (CPM) to explore privacy-related perceptions, including privacy risks and the efficacy of privacy policies.

Figure 3. Information Security and Privacy Protection Word Cloud for Mobile Applications

In the current digital epoch, ensuring information security and privacy protection within mobile applications has emerged as a pressing issue requiring immediate and robust solutions. Figure 1 illustrates the expansive adoption of mobile devices and applications, highlighting the paramount importance of information security, privacy, cloud-facilitated user experiences, and consumer interactions. Simultaneously, the swift progression of mobile applications has ushered in an escalating reliance on these platforms to process sensitive personal data and facilitate various online activities. Consequently, validating the data security and privacy safeguards of mobile applications is imperative. Mobile applications encompass users' private information, not limited to names, addresses, telephone numbers, and bank account details. The mismanagement or malicious exploitation of this information can lead to severe consequences, such as identity theft and financial fraud, among other abuses. Thus, it is obligatory for mobile application developers to implement stringent security measures to safeguard user data. For instance, Ali Balapour leverages the Communication Privacy Management Theory (CPM) to explore privacy-related perceptions, including privacy risks and the efficacy of privacy policies.

To enhance users' comprehension of the potential privacy significance of installed applications, Steven Arzt introduced FlowDroid, a novel and exceedingly precise static taint analysis tool for Android applications, which identified a significantly high percentage of data leaks while maintaining low false positive rates[8]. The information security and privacy protection of mobile applications have emerged as pivotal topics. Through the collaborative endeavors of developers, regulators, and users, a secure and reliable mobile application ecosystem can be established to shield users'personal information and privacy. Concurrently, users need to amplify their information security awareness by diligently reviewing application privacy policies and permission requirements, as well as regularly updating and managing their mobile devices and applications.

4.2 Ensuring Big Data Security and Information Integrity in the Internet of Things (IoT) Ecosystem

The Internet of Things (IoT) orchestrates a network, interlinking a myriad of physical devices and sensors through the internet to achieve intelligent automation. The swift expansion and pervasive application of IoT generate and accumulate substantial data, pivotal for enabling applications across smart cities, intelligent transportation, and smart manufacturing, among other domains. Nonetheless, the security and integrity of big data within the IoT are susceptible to threats, posing challenges to the data's trustworthiness and availability [9]. As show in Figure 4, keywords like big data, information security, and algorithm underscore the criticality of big data security within the IoT framework. Given the multitude of devices and sensors encompassed by IoT, their respective security and protective capabilities are heterogeneous. Adversaries may exploit vulnerabilities within IoT devices to orchestrate intrusions, exfiltrate sensitive data, or disrupt device operations. For instance, hackers might access private information by compromising smart home devices or manipulate traffic flow by attacking intelligent transportation systems, rendering the assurance of big data security in IoT paramount.

Figure 4. Big Data Security and Information Integrity Protection Word Cloud in the Internet of Things

Concurrently, in academic research and practical applications, it is vital to focus on nascent technologies and methodologies pertinent to big data security and information integrity protection within the IoT. This includes exploring innovative

cryptographic algorithms and secure transport protocols to counteract emerging security threats and probing into data privacy protection and anonymization technology to safeguard users' privacy rights and interests. Additionally, conducting security performance evaluations and crafting safety standards are imperative to offering quantitative assessments and guidance concerning big data security within the IoT. Fadele Ayotunde Alaba classifies security threats within the IoT domain into applications, architectures, and communications, providing a detailed analysis of IoT security scenarios, potential attacks, and proposing potential enhancements for the IoT security architecture [10]. The assurance of big data security and information integrity within the IoT is a nuanced and complex subject. Adopting thorough security measures and perpetually innovating technical methods are imperative to safeguarding IoT big data, ensuring its safety and reliability, preserving data integrity, and fostering the sustainable development and application of IoT.

5. Conclusions

This investigation analyzes pertinent literature, exploring research topics and publication trends in data security. Analyzing voluminous academic outputs can aid researchers in understanding productivity, categorization, and growth within research domains. Such insights can facilitate researchers in comprehending the structural framework of academic knowledge in a particular field, thereby guiding forthcoming research endeavors.

Acknowledgments

This research was funded by the National Social Science Grant General Project: Government Regulation of Monopolistic Practices in the Platform Economy and Their Healthy Development (Grant No. 22BKS153).

References

[1] Saxena, V. K., & Pushkar, S. (2014, Mar 08-09). Privacy Preserving Model in Cloud Environment. Paper presented at the Conference on IT in Business, Industry and Government (CSIBIG), Sri Aurobindo Inst Technol, Indore, INDIA.

[2] Sun, Z. C., Wang, Y. J., Cai, Z. P., Liu, T. E., Tong, X. R., & Jiang, N. (2021). A two-stage privacy protection mechanism based on blockchain in mobile crowdsourcing. International Journal of Intelligent Systems, 36(5), 2058-2080.

[3] Bertino, E. (2016, Jun 10-14). Data Security and Privacy Concepts, Approaches, and Research Directions. Paper presented at the 40th Annual IEEE Computer Software and Applications Conference Symposium (COMPSAC) / Symposium on Software Engineering Technology and Applications (SETA), Atlanta, GA.

[4] Zhang, J. L., Chen, B., Zhao, Y. C., Cheng, X., & Hu, F. (2018). Data Security and Privacy-Preserving in Edge Computing Paradigm: Survey and Open Issues. Ieee Access, 6, 18209-18237.

[5] Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario, Canada, 5, 12

[6] Yang, P., Xiong, N. X., & Ren, J. L. (2020). Data Security and Privacy Protection for Cloud Storage: A Survey. Ieee Access, 8, 131723-131740.

[7] Huang, L. S., Chen, H. P., Wang, X., & Chen, G. L. (2000). A fast algorithm for mining association rules. Journal of Computer Science and Technology, 15(6), 619-624.

[8] Arzt, S., et al. (2014). FlowDroid: Precise Context, Flow, Field, Object-sensitive and Lifecycle-aware Taint Analysis for Android Apps. Acm Sigplan Notices, 49(6), 259-269.http://dx.doi.org/10.1145/2666356.2594299

[9] Perera, C., Ranjan, R., Wang, L. Z., Khan, S. U., & Zomaya, A. Y. (2015). Big Data Privacy in the Internet of Things Era. It Professional, 17(3), 32-39.

[10] Alaba, F. A., Othman, M., Hashem, I. A. T., & Alotaibi, F. (2017). Internet of Things security: A survey. Journal of Network and Computer Applications, 88, 10-28.