Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.37395

A Review of Data Privacy Protection Technologies in the Big Data Environment

Zhenyuan Zhou

Hefei 168 High School, Anhui, China
3853235238@qq.com

Abstract. With the rapid development and widespread application of big data technologies, while personal data generates immense social value, the risk of privacy breaches is also escalating. How to effectively protect personal privacy while ensuring data usability has become a central issue of shared societal concern. Existing reviews often focus on describing individual privacy protection technologies, lacking a multidimensional, cross-comparative analysis of various technical approaches. This paper reviews mainstream privacy protection technologies in the big data environment, classifying them into three major categories: data distortion, data encryption, and data anonymization, while also introducing emerging solutions such as federated learning and trusted execution environments. Through a comparative analysis of various technologies across three dimensions—privacy protection strength, data utility, and computational efficiency—this paper finds that each technology has its own strengths and weaknesses in terms of privacy protection; currently, no single technology can simultaneously meet all requirements, and there is a fundamental core conflict between privacy protection and data utility. In practice, appropriate technologies should be selected based on specific business scenarios, or an implementation approach combining multiple technologies should be adopted.

Keywords: Big Data, Privacy Protection, Differential Privacy, Federated Learning, Homomorphic Encryption, Data Anonymization

1. Introduction

Big data has now permeated every aspect of our daily lives. Whether shopping online, watching short videos, or making mobile payments, data is being collected and analyzed behind the scenes. Data can help businesses make decisions, assist doctors in diagnosing illnesses, support governments in managing cities, and inform everyday decision-making. However, these benefits come at a cost—personal information is being widely transmitted and exploited.

In recent years, privacy breaches have occurred frequently. The Facebook data scandal made headlines worldwide, while platforms in China have also been repeatedly reported to have leaked user information. The public has become increasingly concerned about the safety of their privacy, prompting governments to begin regulating such practices. The European Union introduced the General Data Protection Regulation (GDPR), and China enacted the Personal Information Protection Law (PIPL), restricting enterprises from freely using and disseminating user data.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37395

However, legal regulation alone is insufficient; technological measures must keep pace as well. In recent years, researchers have proposed numerous privacy protection methods, including classic techniques such as differential privacy, homomorphic encryption, and k-anonymity, as well as newer approaches like federated learning and trusted execution environments [1]. Nevertheless, most survey articles tend to focus in depth on a single category of technology, with few providing systematic comparisons of different approaches in terms of their respective advantages and trade-offs. Therefore, this paper aims to comprehensively organize and classify both mainstream and emerging privacy protection technologies, compare their strengths and weaknesses from three perspectives, and examine the scenarios in which they are most applicable.

2. Overview of privacy protection technologies

Before delving into specific technologies, it is necessary to clarify the main types of privacy threats in the big data environment.

2.1. Privacy threats in the big data environment

Privacy threats in the big data environment can be categorized into the following three types [2]:

Privacy threats during data collection. Big data applications typically require the collection of massive amounts of data from multiple sources, including social media, Internet of Things devices, mobile applications, and public databases. At this stage, primary privacy risks include excessive data collection, the unauthorized bulk harvesting by third parties of public or semi-public data without users' awareness, and cross-platform tracking—in which behavioral data from different platforms are aggregated to create comprehensive personal profiles.

Privacy threats during data storage. Big data systems generally store data centrally in distributed databases or cloud platforms. While this improves management efficiency, it also significantly increases the risk of data breaches. In recent years, numerous large-scale personal information leakage incidents have occurred precisely at this stage.

Privacy threats during data analysis and publication. When data is used for mining and analysis or released to the public, privacy risks persist. Typical threats include linkage attacks, inference attacks, and differential attacks [2, 3]. These risks demonstrate that even when data undergoes anonymization before publication, the possibility of privacy leakage cannot be completely eliminated.

2.2. Core design objectives of privacy protection technologies

The design of privacy protection technologies must balance three core objectives. Aguilera Martínez et al., systematically discuss the trade-offs among these objectives from the perspective of the full data lifecycle [4]:

Security. This is the most fundamental requirement. Security means that even if an attacker obtains encrypted data, intermediate computation results, or final outputs, they cannot reverse-engineer specific personal information from them.

Data utility. Data should not become completely unusable after privacy protection is applied, as this would undermine the purpose of the protection. A well-designed solution should preserve as much analytical utility as possible while maintaining adequate privacy protection.

Computational efficiency. Big data processing is already time-consuming and resource-intensive, and privacy protection typically adds additional overhead. Some approaches, such as homomorphic

DOI: 10.54254/2755-2721/2026.37395

encryption, incur extremely high computational costs [5]. Efficiency ultimately determines whether a technology can be practically deployed—excessive time and resource consumption cannot be justified solely on the grounds of security.

3. Mainstream privacy protection technologies

3.1. Data distortion-based privacy protection technologies

Data distortion techniques comprise a class of methods for protecting individual privacy by adding random noise to original data or query results, with the aim of balancing "data usability" and privacy. Its core idea is to prevent attackers from reconstructing specific individuals' real information from published data, while ensuring that macro-level statistical conclusions derived from the data remain consistent with the original data. The typical representative of this category is differential privacy, along with methods such as data randomization.

Differential privacy is widely regarded as one of the most influential mathematical frameworks in the field of privacy protection. It provides a provable mathematical guarantee that does not rely on assumptions about the attacker's background knowledge. Its core idea can be expressed as follows: the result of a query on a dataset should not change significantly due to the addition or removal of any single record. In other words, even if an attacker possesses all information except for one specific record, they cannot infer the value of that record from the query result. Essentially, differential privacy achieves this protection by adding carefully calibrated noise to the data, with a mathematical guarantee that, regardless of the attacker's background information, the attacker cannot reliably determine whether a specific record is included in the dataset [6].

Differential privacy is mathematically defined as follows: for any two adjacent datasets D and D' that differ by only one record, a randomized algorithm M satisfies: Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D') \in S]

where ε represents the privacy budget. A smaller ε indicates stronger privacy protection, meaning larger added noise; conversely, while data usability is higher, the level of privacy protection decreases.

Advantages and Limitations

The main advantage of differential privacy is its mathematical provability, allowing quantitative evaluation of privacy protection levels—a capability that other methods, such as data anonymization and traditional encryption, cannot offer [7].

However, its limitations are equally apparent. Privacy loss under repeated queries is a serious concern: when the same information is queried multiple times, the cumulative privacy loss increases with each query, potentially weakening the overall privacy guarantee. Moreover, as the required privacy protection strength increases, the accumulated noise grows accordingly, causing severe data distortion and giving rise to the "fundamental trade-off between privacy protection and data utility" [6, 8]. Furthermore, differential privacy still relies on certain idealized assumptions and requires further validation in complex real-world scenarios.

3.2. Data encryption technologies

Data encryption technology is another important category of privacy protection methods. It transforms plaintext into ciphertext using cryptographic operations, thereby preventing unauthorized parties from accessing the underlying data, while allowing authorized users to decrypt the ciphertext

DOI: 10.54254/2755-2721/2026.37395

and recover the original information. Typical representatives in this field include homomorphic encryption and secure multi-party computation [5].

(1) Homomorphic Encryption

Homomorphic encryption allows users to perform operations—such as addition and multiplication—directly on ciphertexts, with the decrypted result matching the result obtained by performing the same operation on the corresponding plaintexts. This feature makes it particularly valuable in cloud computing and outsourced computation scenarios: data owners can upload encrypted data to the cloud, and cloud service providers can perform computations without decrypting the data or directly accessing the underlying plaintext. This method effectively prevents data leakage risks during storage and computation.

(2) Secure Multi-Party Computation

Secure multi-party computation (SMPC) is another privacy protection method based on cryptographic techniques. Its origins can be traced back to Academician Yao Qizhi's 1982 formulation of the "Millionaires' Problem." The problem is described as follows: two millionaires wish to determine which of them is wealthier without revealing their exact net worth to each other—a scenario that captures the core tension underlying multi-party data collaboration in the big data era [5]. Multiple parties each hold private data and wish to jointly complete a computation task—such as statistical averaging or collaborative modeling—without exposing their private data to one another. Each party splits their data into multiple "shares" distributed among participants; no single party can reconstruct the complete data, but all parties collaborating can obtain the correct computation result [3], thereby protecting privacy through this mechanism [1].

(3) Advantages and Limitations

The greatest advantage of data encryption technology is its extremely high level of security. Because the data remains in ciphertext form throughout the computation, even if it is stolen or the system is breached, an attacker cannot obtain any information. While it is highly feasible in small-scale, multi-party collaboration scenarios, its implementation in massive big data scenarios is limited.

3.3. Data anonymization techniques

Data anonymization techniques are a class of methods that protect privacy by deleting, generalizing, or obscuring personally identifiable information. The core idea is to process information that can be used to identify individuals—such as names, ID numbers, phone numbers, and exact addresses—before the data is released, thereby preventing attackers from linking records in the data to specific individuals. Representative methods of data anonymization include k-anonymity and its improved models, l-diversity and t-neighbor.

k-Anonymity

The core objective of k-anonymity is to ensure that, in the published data, each record is indistinguishable from at least k-1 other records in the dataset based on "pseudo-identifiers." Pseudo-identifiers refer to attributes that, when considered individually, cannot identify a specific person but, when combined—such as year of birth, gender, or ZIP code—can reveal information pointing to a specific individual. To achieve k-anonymity, pseudo-identifiers—such as age and address—typically require generalization.

l-Diversity and t-Nearest Neighbors

The introduction of k-anonymity was of great significance in the field of privacy protection, but researchers soon discovered significant vulnerabilities in the approach. The most critical issue is the homogeneity attack: in data that satisfies k-anonymity, if all records within a given group share the

DOI: 10.54254/2755-2721/2026.37395

same sensitive attributes, an attacker can determine the sensitive information of any individual in that group—even if they cannot pinpoint a specific person. To address this flaw, Machanavajjhala et al. proposed the l-diversity model, which requires that each equivalence class contain at least l distinct values for the sensitive attribute, thereby increasing the uncertainty of inference. Building on this, Li et al. further proposed the t-nearest-neighbor model, which not only requires diversity in the sensitive attribute but also demands that its distribution closely resemble that of the overall dataset, thereby providing more refined protection of private information.

Advantages and Limitations

Its advantages lie in its intuitive operation and relative simplicity in implementation. Operations such as generalization and de-identification are easy to understand and do not require complex cryptographic computations, making it widely adopted in data publication and sharing scenarios.

However, its limitations are also quite evident. First, background knowledge attacks pose a difficult-to-overcome threat: if an attacker possesses additional external information, they may link the anonymized data to background information they already know, thereby re-identifying individuals. Second, in big data scenarios, data contains an extremely wide variety of information. Meeting the k-anonymity standard requires extensive generalization, which leads to significant loss of fine-grained data details. Furthermore, research on k-anonymity indicates that relying solely on anonymization may not provide quantifiable privacy guarantees. Anonymization techniques lack strictly mathematically provable privacy boundaries, which is the key distinction from differential privacy.

4. Emerging privacy protection technologies

In recent years, the rapid development of artificial intelligence and cloud computing has driven the emergence of a range of new privacy-preserving technologies. Two representative cutting-edge approaches are federated learning and trusted execution environments.

4.1. Federated learning

Federated learning is a machine learning paradigm designed for distributed data scenarios, first proposed by the Google team in 2016. Its core idea is "data stays put, models move" — unlike traditional centralized learning, federated learning does not aggregate data from various locations to a central server; instead,models are distributed to the locations where the data are stored for local training [1].

To further strengthen privacy protection, federated learning is often combined with differential privacy or secure multi-party computation techniques [2]. The advantage of federated learning lies in its particular suitability for domains with extremely high data privacy requirements, such as healthcare and finance. For example, multiple hospitals can jointly train a disease prediction model without sharing patient medical records. However, federated learning also faces several challenges : communication costs among participating parties can be substantial, and when data distributions differ significantly across parties, the performance of the global model may deteriorate [1].

4.2. Trusted execution environments

A trusted execution environment (TEE) is a privacy protection technology based on hardware security mechanisms. Its core idea is to construct an isolated secure area within the CPU. When data

DOI: 10.54254/2755-2721/2026.37395

and code run within this area, even the operating system, hypervisor, or cloud service providers cannot access or tamper with the content.

Compared with purely cryptographic approaches such as homomorphic encryption, the main advantage of trusted execution environments lies in their relatively high computational efficiency, as they typically incur substantially lower computational overhead than homomorphic encryption [1].

However, TEEs also have their limitations. First, their security heavily depends on the hardware itself: if the hardware has design vulnerabilities or suffers side-channel attacks — such as inferring internal states by analyzing power consumption or execution time — the privacy protection may fail [1].

5. Technology comparison and challenges

Differential privacy provides quantifiable privacy guarantees for privacy protection, but its data utility loss is significant in high-dimensional scenarios [6, 7].

Homomorphic encryption excels in security and data losslessness, but its major drawback lies in its excessively low computational efficiency, making it difficult to support large-scale practical applications [5].

k-anonymity is simple to implement and relatively efficient in practice, but its defensive capability is limited when attackers possess extensive background knowledge [9].

Federated learning and trusted execution environments represent more recent privacy protection approaches that can, to some extent, mitigate the trade-off between security and efficiency, but they also face limitations related to data distribution and hardware deployment costs [1, 2].

Therefore, in practice, the choice of privacy protection technology should be tailored to the specific application scenario, either by selecting an appropriate approach or combining multiple technologies [3, 10].

6. Conclusion

In summary, while big data is indeed highly useful in today's era, the risks of privacy leakage are simultaneously growing. This paper has systematically reviewed current mainstream privacy protection technologies—including classic approaches such as differential privacy, homomorphic encryption, and k-anonymity, as well as newer techniques like federated learning and trusted execution environments. The comparison shows that no single technology is universally applicable. For instance, technologies that provide strong protection tend to be computationally slow; those that are fast may not effectively defend against knowledgeable attackers; and attempts to strike a balance often introduce excessive communication overhead. The fundamental trade-off between privacy protection and data utility remains unresolved, and no single approach can fully reconcile these competing objectives. In practice, the choice of technology depends on the specific context. Looking ahead, as artificial intelligence technologies continue to evolve and quantum computing gradually matures, the field of privacy protection will face emerging challenges. Privacy protection is not merely a technical issue but also involves broader legal, ethical, and governance considerations, which require coordination among relevant stakeholders.

References

[1] Li, J., Yin, Y., Li, S., et al. (2024). A review of big data security and privacy computing technologies. *Journal of Cyberspace Security Science, 2*(6). DOI: 10.20172/j.issn.2097-3136.240601.

DOI: 10.54254/2755-2721/2026.37395

[2] Liu, S., Gao, Q., Zhang, F., et al. (2024). A review of privacy-preserving computing technologies for multi-party data fusion analysis. *Journal of Network and Information Security, 10*(6).
[3] Zhang, M., & Li, H. (2025). A review of multi-granularity three-branch decision privacy protection. *Journal of Minnan Normal University (Natural Science Edition), 38*(4), 1–18.
[4] Madhusudhanan, S., & Jose, A. C. (2025). Privacy preservation techniques through data lifecycle: A comprehensive literature survey. *Computers & Security, 155*, 104473. DOI: 10.1016/j.cose.2025.104473.
[5] Huang, L., Tian, M., & Huang, H. (2015). A review of cryptographic techniques for big data privacy protection. *Journal of Software, 26*(4), 945–959.
[6] Aguilera-Martínez, F., & Berzal, F. (2025, June 13). Differential privacy in machine learning: From symbolic AI to LLMs. https://arxiv.org/abs/2506.11687.
[7] Zheng, Z., Yao, L., Hu, H., et al. (2025). Benchmarking relaxed differential privacy in private learning: A comparative survey. *ACM Computing Surveys, 57*(12), 316. https://doi.org/10.1145/3729216
[8] Zhang, L., & Zhang, Y. (2025). Defending against attacks in deep learning with differential privacy: A survey. *Artificial Intelligence Review, 58*, 347.
[9] Anoop, K., Sreejith, S., & Rajesh, R. (2025). Privacy-preserving data mining techniques in big data environments. In *2025 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)* (pp. 696–702). IEEE.
[10] Cuzzocrea, A., & Soufargi, S. (2025). Privacy-preserving multidimensional big data analytics models, methods and techniques: A comprehensive survey. *Expert Systems with Applications, 270*, 126387.