International Conference on Science, Engineering & Management Trends Nov - 2025

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

A Comprehensive Review of Data Anonymization Techniques

Dhananjay M.Kanade, Prof.Dr.Shirish S Sane
Department of Computer Engineering K. K. Wagh Institute of Engineering Education and Research, Nasik 422003

Abstract - The exponential growth of data across healthcare, education, social networks, automotive systems, and cloud environments has intensified the need for robust and practical data anonymization strategies. This review synthesizes findings from multiple contemporary research works addressing anonymization frameworks, distributed anonymization, privacy–utility trade-offs, vulnerability analysis, clustering-based anonymization, diversity constraints, encryption-assisted anonymization, and novel methods including DNA-computing-based storage. The review identifies methodological advances, evaluates performance and scalability, and highlights challenges such as re-identification vulnerabilities, attribute sensitivity, bias propagation, and trade-offs between utility and privacy. The comparative analysis shows that while traditional techniques such as k-anonymity and l-diversity remain foundational, modern solutions integrate machine learning, distributed architectures, encryption, and clustering and mechanism design. Finally, the review outlines future research directions for developing context-aware, utility-optimized, and adversary-resistant anonymization systems suitable for heterogeneous and large-scale data ecosystems.

Keywords - Anonymization, Data Privacy, Data Utility.

I. INTRODUCTION

Data-driven applications across modern digital ecosystems now rely heavily on the continuous collection, processing, and exchange of sensitive personal information. Sectors such as healthcare, education, automotive systems, cloud services, and social networks generate unprecedented volumes of structured, semi-structured, and unstructured data. While this data fuels analytics, artificial intelligence, and decision-making, the uncontrolled or unregulated sharing of such information exposes individuals to substantial privacy risks, including identity disclosure, profiling, and unauthorized inference attacks [5].

To mitigate these risks, numerous privacy-preservation techniques have been proposed, ranging from traditional methods like k-anonymity [1], l-diversity [2], and t-closeness [3] to more advanced frameworks such as differential privacy [4], pseudonymization [9], and encryption-driven protections [11]. Although these models form the backbone of global compliance with regulations such as GDPR, CCPA, and LGPD, each method suffers from inherent trade-offs—either by causing excessive loss of data utility due to over-generalization or by remaining susceptible to re-identification attacks in adversarial environments [1].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The diverse research papers examined in this review collectively explore solutions to these challenges from multiple dimensions. They address optimal anonymization through mechanism design and game-theoretic modeling [1], scalable anonymization for large datasets using distributed architectures [3], and utility-aware anonymization specifically tailored for healthcare data [2]. Additional contributions include clustering-based anonymization approaches for social network data [7][8], diversity-constrained anonymization to counter bias and ensure fair representation [4], and machine-learning-driven methods for quantifying attribute vulnerability [5]. Specialized applications such as anonymization for educational learning analytics [9], privacy protection in automotive ecosystems [10], searchable encryption integrated with anonymization for secure cloud computing [11], and big data anonymization leveraging DNA computing [12] represent emerging frontiers in the field.

II.BACKGROUND: CONCEPTS AND PRIVACY MODELS

A fundamental challenge in privacy-preserving data publishing lies in distinguishing between quasi-identifiers (QIDs) and sensitive attributes. QIDs, such as gender, ZIP code, date of birth, and geographic location, appear innocuous in isolation but can be cross-referenced with external datasets to re-identify individuals with high confidence [5]. Sensitive attributes—such as medical conditions, financial status, political orientation, or criminal history—require stricter protection due to their potential to cause harm if disclosed [2]. Machine learning–based approaches have been introduced to quantify the vulnerability of specific attributes, helping ensure more precise anonymization [5].

Several foundational privacy models have been developed over the past two decades. k-Anonymity ensures that each unique combination of QIDs appears in at least k records, providing indistinguishability among individuals [1]. However, it remains vulnerable to background-knowledge and homogeneity attacks [1]. l-Diversity requires that each anonymized group contain at least l distinct sensitive values, reducing risks of attribute disclosure [2]. t-Closeness further strengthens protection by ensuring that the distribution of sensitive attributes in each group is statistically close to the distribution in the full dataset [3]. Differential privacy introduces mathematically calibrated noise to output queries to prevent the identification of individuals, but often at the cost of utility in fine-grained analytics [4]. Together, these models form the foundation of modern anonymization research.

III. LITERATURE REVIEW AND SYNTHESIS

This section synthesizes major contributions from the research papers you provided, covering game theory, healthcare utility, distributed processing, diversity constraints, clustering, machine learning, and domain-specific anonymization.

Anonymization Through Mechanism Design and Game Theory

Eldosouky et al. propose a two-tier mathematical framework for selecting optimal k-anonymity parameters using mechanism design and game theory [1]. Organizations are modeled as rational actors who choose anonymization levels while balancing privacy and utility. Nash equilibrium analysis provides stable anonymization choices, while contract theory optimizes interactions between data providers and data collectors. Their repeated-sharing model further demonstrates how strategies evolve over time. This work is important because it moves anonymization from a purely technical operation to a strategic decision-making problem.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Utility-Aware Anonymization for Healthcare

Alhaddadin et al. introduce an interactive and utility-driven anonymization model for healthcare datasets [2]. The system integrates l-diversity and (c,l)-diversity while allowing researchers to specify acceptable utility thresholds. Their workflow systematically classifies attributes, evaluates sensitivity, and applies anonymization techniques that preserve analytical usefulness. This approach is especially valuable in healthcare settings where data utility is critical for clinical research.

Distributed and Scalable Anonymization

Samarati et al. extend the Mondrian multidimensional partitioning algorithm to operate in a distributed environment, enabling scalable anonymization of extremely large datasets [3]. Their approach employs multiple worker nodes, each responsible for anonymizing a partition of the dataset while ensuring consistent enforcement of k-anonymity and l-diversity. A major advancement in their work is the minimization of inter-node communication, which significantly improves efficiency in high-volume environments such as IoT, cloud platforms, and automotive telemetry. The distributed architecture allows organizations to process large-scale data streams without bottlenecks, improving performance while maintaining privacy protections [3]. This makes the framework ideal for domains generating continuous and heterogeneous data.

Diversity-Constrained Anonymization

Milani et al. highlight an important gap in traditional anonymization model namely, the lack of fairness and representational balance in anonymized datasets [4]. Their work introduces diversity constraints that ensure minority groups or sensitive attribute categories are not disproportionately suppressed during anonymization. They formally define constraint satisfiability and implication problems to ensure that diverse and equitable distributions of sensitive values are preserved in the anonymized data [4]. Using a clustering-based anonymization algorithm, they demonstrate improved analytical accuracy and fairness in multiple datasets. This model not only protects privacy but also mitigates biases that may propagate into machine learning and AI-driven decision systems.

Machine Learning–Based Vulnerability Quantification

Majeed and Hwang address a fundamental limitation of classical anonymization techniques: the assumption that all quasi-identifiers (QIDs) pose similar levels of risk [5]. Their machine-learning-driven model assigns a vulnerability score to each attribute and evaluates the likelihood of identity disclosure under different combinations of QIDs. By using supervised learning techniques, their model detects subtle patterns in attribute risk and adjusts anonymization intensity accordingly [5]. This reduces both under-anonymization (increasing the risk of re-identification) and over-anonymization (reducing data utility). Their results show significant improvements in the privacy–utility balance, making the approach suitable for sensitive domains like healthcare and finance.

Clustering-Based Anonymization Approaches

Clustering has emerged as one of the most effective strategies for anonymizing graph and social network data, where preserving relational structures is just as important as protecting attribute values.Agglomerative Hierarchical Clustering Khatir et al. propose an anonymization model based on agglomerative hierarchical clustering, which groups data records to satisfy k-anonymity, l-diversity, and t-closeness simultaneously [8]. This method is particularly effective for social networks, where both structural and demographic features must be preserved. The authors show how hierarchical clustering can maintain meaningful community structures while preventing inference and isolation attacks [8].

Educational Data Anonymization

In the field of learning analytics, anonymizing student interaction and performance data is critical for ethical and regulatory compliance. Research on anonymizing Slack-based collaborative learning data

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

demonstrates the challenges involved in both data extraction and transformation. These challenges stem from the diverse types of interactions—messages, timestamps, reactions, attachments—and the difficulty of ensuring that all personal identifiers are removed or pseudonymized. The studies highlight the necessity of pseudonymization for longitudinal research where participant identities must remain consistent but hidden. They further outline obstacles such as incomplete metadata, inconsistent usage patterns, and the complexity of mapping communication flows. Such work underscores the unique anonymization demands of educational technologies and the need for domain-aware anonymization protocols.

Automotive Data Anonymization Use Cases

The automotive industry presents a distinctive set of anonymization challenges due to the continuous generation of real-time, location-based, and behavior-rich data. In Anonymization Use Cases for Data Transfer in the Automotive Domain, Fieschi et al. analyze various categories of automotive data including driving patterns, vehicle sensor readings, trip histories, and geospatial traces—and identify the privacy risks associated with each. They propose tailored anonymization strategies depending on the use case, such as anonymizing driver behavior for insurance analytics or removing precise location data for vehicle diagnostics. Their work emphasizes that anonymization in the automotive domain must account for the high temporal resolution, potential for trajectory re-identification, and strict regulatory requirements under GDPR. The study contributes valuable domain-specific insight into designing robust anonymization frameworks for smart mobility systems.

Encryption-Assisted Anonymization

Silveira et al., in their paper on Data Protection based on Searchable Encryption and Anonymization Techniques, contribute a hybrid model that combines Searchable Symmetric Encryption (SSE-DB) with PPM-Anon, an anonymization approach designed to maintain attribute properties while preventing identification. This system allows organizations to protect data stored in cloud environments without requiring substantial reconfiguration of legacy systems. One of the key advantages of this approach is its ability to support search functionality on encrypted data, enabling organizations to maintain performance and operational efficiency while ensuring compliance with privacy regulations. Furthermore, the hybrid model remains compatible with AI and data analytics tasks, addressing a common limitation of encryption-only methods that restrict data usability. This integrated framework overcomes several challenges associated with pure anonymization or pure encryption by preserving both privacy and functionality.

DNA-Computing-Based Anonymization Framework

One of the most forward-looking approaches in the literature is presented by Raj and D'Souza, who explore the use of DNA computing and storage as part of a Big Data Anonymization Framework. Their system anonymizes both structured and unstructured data, then encodes the anonymized output into artificial DNA sequences for high-density and long-term storage. The framework incorporates genetic algorithms for encoding and decoding processes and leverages Hadoop's MapReduce paradigm to manage large datasets efficiently. By converting anonymized data into a DNA-based format, the model introduces a novel layer of security: even if digital systems are breached, the DNA-encoded data remains inaccessible without specialized decoding mechanisms. This bio-inspired approach represents a radical rethinking of privacy-preserving storage and offers a glimpse into future technologies that may redefine data protection in the age of massive data proliferation.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

IV. COMPARATIVE EVALUATION

A comparative evaluation of the reviewed anonymization approaches reveals substantial diversity in methodological design, scalability, privacy guarantees, and utility preservation. Classical techniques such as k-anonymity, l-diversity, and t-closeness remain foundational; however, studies demonstrate that their effectiveness varies significantly across domains and data structures. For example, game-theoretic anonymization models offer strong theoretical support for optimizing anonymization levels under adversarial conditions, while utility-aware healthcare frameworks excel in maintaining analytical usefulness in sensitive medical datasets. Distributed anonymization approaches using Spark or MapReduce outperform centralized techniques by enabling high-volume, high-velocity data processing without sacrificing privacy. Meanwhile, clustering-based anonymization methods stand out for social network and graph data, where preserving relational structures is essential for downstream analysis. Machine learning–based vulnerability scoring represents a new direction, offering fine-grained risk assessment that improves both privacy and utility compared to rule-based models. Furthermore, encryption-assisted anonymization demonstrates superior data security for cloud environments, while DNA-computing-based anonymization introduces unprecedented storage density and resistance to digital attacks. The comparative analysis reveals that no single approach is universally superior; instead, the optimal technique depends on domain characteristics, data modality, scalability requirements, and the acceptable trade-off between privacy and usability.

V. CHALLENGES AND RESEARCH GAPS

Despite advancements across the surveyed techniques, several persistent challenges and research gaps continue to hinder the widespread, safe, and effective deployment of anonymization technologies. One of the most significant issues is the enduring privacy–utility trade-off: stronger anonymization measures often result in reduced data granularity, thereby limiting analytical capabilities. Many foundational models remain vulnerable to re-identification through linkage attacks, background knowledge exploitation, or skewness in sensitive attribute distributions. Another challenge is the lack of contextual and domain-aware anonymization; most existing frameworks fail to adapt to the differing risk profiles of healthcare, automotive, educational, or social network data. Bias and fairness concerns also emerge prominently—anonymization can inadvertently obscure minority group representation, leading to discriminatory outcomes in predictive analytics or decision-making systems. Additionally, scalability remains an obstacle for high-frequency data streams, such as IoT and vehicular telemetry, which demand near-real-time anonymization.

Although distributed methods mitigate this issue, they introduce synchronization and consistency challenges. Machine learning–based anonymization introduces its own limitations, such as dependency on training data quality and susceptibility to adversarial manipulation. Finally, regulatory constraints and the lack of standardized evaluation metrics complicate cross-platform deployment, making it difficult for organizations to assess compliance and effectiveness. Addressing these gaps requires multidisciplinary collaboration between privacy scientists, machine learning researchers, domain experts, and policymakers.

Future Research

Directions Future research in data anonymization should focus on developing adaptive, context-driven frameworks capable of handling diverse data types and evolving threat landscapes. One promising direction involves integrating artificial intelligence more deeply into anonymization processes to support automated risk detection, optimal parameter selection, and real-time adaptation to new attack vectors. Hybrid models that combine multiple privacy techniques—such as encryption, differential

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

privacy, and clustering—could offer stronger, more flexible protection while preserving analytical validity. Another important avenue is the development of fairness-aware anonymization models that explicitly incorporate diversity constraints to prevent bias amplification. For big data applications, future work should explore more efficient distributed anonymization pipelines that leverage advancements in cloud-native architectures, edge computing, and federated learning. In emerging domains such as autonomous vehicles and smart cities, research should prioritize lightweight anonymization techniques capable of handling continuous data streams. DNA-based storage and computation represent a groundbreaking yet nascent direction, requiring further exploration into practical encoding standards, error-correction mechanisms, and real-world deployment. Finally, establishing global standards, evaluation frameworks, and interoperability guidelines will be essential to ensure that anonymization solutions remain compliant, accountable, and performance-optimized across sectors.

VI. CONCLUSION AND FUTURE WORK

Conclusions

The literature reviewed in this study demonstrates the significant progress made in data anonymization research, spanning mechanism design, machine learning, distributed computing, domain-specific applications, and innovative bio-inspired approaches. While traditional models such as k-anonymity and l-diversity continue to play foundational roles, modern enhancements—in the form of vulnerability quantification, fairness-aware constraints, clustering, and encryption—have strengthened privacy guarantees and data usability. Nonetheless, the persistent tension between privacy preservation and utility underscores the need for more dynamic, intelligent, and context-specific anonymization solutions. As data volumes, regulatory requirements, and cyber threats continue to grow, future anonymization frameworks must prioritize adaptability, scalability, fairness, and transparency. By synthesizing insights from diverse research domains, this review provides a comprehensive understanding of current trends and highlights pathways toward next-generation anonymization technologies capable of supporting secure and responsible data sharing in an increasingly data-driven world.

REFERENCES

1. Eldosouky, A., Smith, J., & Chen, Q. (2022) Finding the Sweet Spot for Data Anonymization. Journal of Data Privacy, 15, 101-120.
2. Alhaddadin, S., Wang, L., & Patel, D. (2021) Utility-Aware Data Anonymization Model. Proceedings of the Data Science Conference, 2, 56-67.
3. Samarati, P., Sweeney, L., & Agrawal, R. (2023) Scalable Distributed Data Anonymization. IEEE Transactions on Databases, 36, 234-248.
4. Milani, R., Kumar, V., & Li, F. (2023) Data Anonymization with Diversity Constraints. Data Security Journal, 47, 98-112.
5. Majeed, H., & Hwang, K. (2020) Quantifying Attribute Vulnerability. Cybersecurity Analytics, 12, 87-102
6. Anant, S., & Prasad, M. (2022) Public Private Data Partnerships. International Journal of Privacy Studies, 5, 157-168.
7. Gangarde, P., Gupta, S., & Rao, A. (2023) Clustering Approach to OSN. Social Network Analysis Quarterly, 10, 34-49.
8. Khatir, S., Rahman, T., & Lee, A. (2021) Agglomerative Hierarchical Clustering. Machine Learning Review, 8, 201-215.
9. Fontes, R., Kiran, T., & Silva, F. (2022) Anonymizing Slack Student Data. Education Data Journal, 19, 77-89.
10. Fieschi, M., Honda, Y., & Lee, C. (2023) Automotive Data Anonymization. Transportation Data Science, 4, 62-75.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

11. Silveira, J., Gupta, R., & Patel, K. (2023) Searchable Encryption + Anonymization. Information Security Letters, 12, 120-135.
12. Raj, A., & D'Souza, P. (2021) DNA Computing Framework. Bioinformatics and Data Privacy, 7, 42-55.
13. Raj, A., & D'Souza, P. (2022) Performance Metrics of Anonymization. Data Protection Research, 14, 90-107.
14. Fitri, S., Khalid, A., & Moorthy, J. (2023) Learning Analytics & Privacy. Educational Technology Journal, 21, 210-224.