STM JOURNALS

Recent Trends in
Parallel Computing

ISSN: 2393-8749
Volume 12, Issue 3, 2025
September–December
DOI (Journal): 10.37591/RTPC

https://journals.stmjournals.com/rtpc

Review RTPC

An Effective Privacy Preservation Technique for Enhancing Data Usability

Dhananjay M. Kanade[1,*], Shirish S. Sane[2]

Abstract

The rapid growth and adoption of modern database systems have created immense opportunities for researchers, industries, and organizations to extract meaningful knowledge and make data-driven decisions. While this progress has enabled the discovery of valuable patterns and trends, it has also intensified the challenge of safeguarding individual privacy. Merely removing direct identifiers such as names, social security numbers, or Aadhar card details is no longer sufficient, as adversaries can often exploit quasi-identifiers like gender, date of birth, and postal codes to re-identify individuals with alarming accuracy. To address these vulnerabilities, the field of Privacy-Preserving Data Publishing (PPDP) has emerged, offering techniques that attempt to strike a delicate balance between maintaining data utility and ensuring strong privacy guarantees. This study provides a detailed exploration of prominent PPDP models, including k-anonymity, ℓ-diversity, and t-closeness, while also reviewing newer strategies such as β-likeness and disassociation. Each method's strengths, weaknesses, and real-world applicability are critically assessed. In addition, the study highlights the inherent trade-offs between data protection and usability, underlining the importance of adaptive, efficient, and context-aware solutions for secure data sharing in an era of growing privacy risks.

Keywords: Data anonymization, data utility, privacy-preserving data publishing (PPDP), general data protection regulation (GDPR)

INTRODUCTION

Advancements in database systems have enabled the analysis of large-scale datasets, driving discoveries in fields such as healthcare, finance, and marketing. However, the increasing availability of shared data has raised significant privacy concerns, particularly due to the risk of re-identification in anonymized datasets. Conventional anonymization techniques, which focus on removing direct identifiers, have proven insufficient. Research study revealed that 87% of the US population could be uniquely identified using just three quasi-identifiers: zip code, gender, and date of birth. This highlights the necessity for effective Privacy-Preserving Data Publishing (PPDP) techniques that ensure a balance between data privacy and usability [1–4].

***Author for Correspondence**
Dhananjay M. Kanade
E-mail: dmkanade@kkwagh.edu.in

[1]Research Scholar, Department of Computer Engineering, Karmaveer Kakasaheb Wagh Institute of Engineering Education and Research, Nashik, Savitribai Phule Pune University, Pune, Maharashtra, India
[2]Professor and Research Guide, Principal and Professor, Department of Computer Engineering, Gokhale Education Society's R.H. Sapat College of Engineering, Management Studies and Research, Nashik, Maharashtra, India

Received Date: May 12, 2025
Accepted Date: September 10, 2025
Published Date: October 30, 2025

Citation: Dhananjay M. Kanade, Shirish S. Sane. An Effective Privacy Preservation Technique for Enhancing Data Usability. Recent Trends in Parallel Computing. 2025; 12(3): 35–40p.

Regulatory frameworks like the General Data Protection Regulation (GDPR) emphasize the importance of protecting personal data, holding organizations accountable for implementing robust privacy safeguards. Privacy-Preserving Techniques (PPTs) are designed to anonymize datasets while maintaining their usefulness for data mining, machine learning, and analytical applications. This study provides a comprehensive review of various PPTs, classifying them into non-perturbative, perturbative, de-associative, and synthetic data generation approaches. The focus is on relational

databases and the challenge of minimizing disclosure risks while preserving predictive accuracy and analytical reliability [4–7].

LITERATURE REVIEW

Privacy-preserving data publishing (PPDP) plays a vital role in contemporary data management by transforming raw datasets into secure formats that protect sensitive information while maintaining their analytical utility. PPDP is designed to mitigate privacy risks across various database types, including microdata, tabular data, and query-based systems [8–12]. Dataset attributes are typically classified into four categories:

- *Identifiers:* Directly reveal an individual's identity, such as names, social security numbers, or passport numbers.
- *Quasi-identifiers (QIs):* Attribute combinations like gender, zip code, and date of birth that, when linked with external data, can uniquely identify individuals.
- *Sensitive attributes:* Contain confidential information, including medical histories, financial details, or political preferences.
- *Non-sensitive attributes:* Do not disclose personal or confidential details.

The fundamental objective of PPDP is to prevent both identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is explicitly recognized within a dataset, whereas attribute disclosure happens when sensitive information about an individual is inferred. The key challenge is striking a balance between privacy and data utility ensuring that datasets remain valuable for analysis while safeguarding individual privacy [13–17].

PRIVACY-PRESERVING TECHNIQUES (PPTs)

Figure 1 shows that PPTs are classified into four categories based on their approach to anonymization:

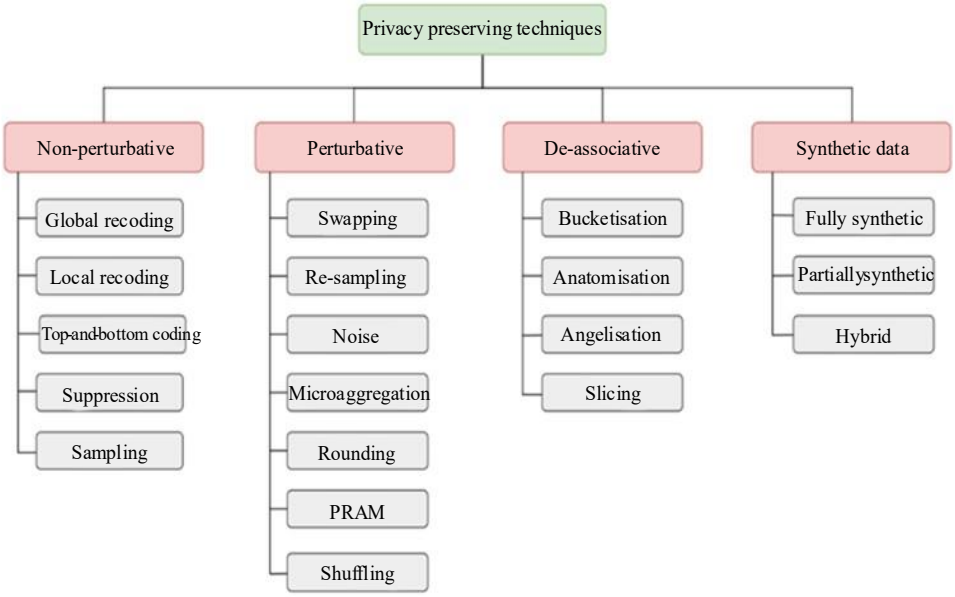

Figure 1. Flowchart of Privacy-Preserving Techniques (PPTs).

Non-Perturbative Techniques

Non-perturbative techniques safeguard privacy by modifying the level of detail in datasets without altering the data's integrity [18–21]. These approaches achieve protection by generalizing or suppressing identifiable attributes while ensuring the data remains accurate and reliable.

- *Global Recoding:* Combines values into generalized categories, such as grouping ages into ranges (e.g., 20–30, 31–40) or combining specific locations into broader regions. This technique ensures uniform treatment of attributes across the dataset.

- *Local Recoding:* Selectively generalizes values in certain regions of the dataset. For example, only specific outliers may be generalized, reducing unnecessary distortion.
- *Top-and-Bottom Coding:* Limits data values to thresholds by capping or flooring extreme values. For example, incomes above Rs 200,000 may be grouped as ">200,000".
- *Suppression:* Eliminates specific values or records entirely to mitigate disclosure risks. For instance, rare combinations of quasi-identifiers may be suppressed to prevent re-identification.
- *Sampling:* Releases a subset of the dataset rather than the entire dataset to reduce re-identification likelihood. However, this may introduce sampling bias if not carefully implemented.

Perturbative Techniques

Perturbative methods introduce uncertainty into datasets by distorting data values. These techniques are particularly effective in protecting privacy while preserving statistical properties.

Noise Addition: Adds random noise to numerical data to obscure individual values while maintaining overall distributions. For example, adding Gaussian noise to salaries can protect privacy without significantly affecting statistical analysis [22–25].

Micro-aggregation: Groups records with similar values and replaces individual values with aggregated statistics such as the mean or median. This technique is commonly used for continuous data.

Swapping: Exchanges attribute values between records to break direct correlations between identifiers and sensitive attributes. For example, swapping addresses between individuals in different regions can obscure residential patterns.

Shuffling: Randomly permutes sensitive attributes within groups, ensuring privacy while retaining the overall structure of the dataset.

De-Associative Techniques

De-associative methods aim to break the relationship between quasi-identifiers and sensitive attributes.

Bucketization: Partitions data into buckets where each bucket contains multiple sensitive values. For example, in a healthcare dataset, bucketization might group patients by age range and shuffle their medical conditions within each group.

Anatomization: Separates quasi-identifiers and sensitive attributes into two linked tables. This ensures sensitive attributes cannot be directly associated with specific individuals.

Slicing: Groups correlated attributes into subsets and permutes values within each subset. This technique preserves attribute correlations while reducing re-identification risks.

Synthetic Data Generation

Synthetic data methods create artificial datasets that replicate the statistical characteristics of the original data while maintaining privacy.

- Fully Synthetic Data: Replaces all original values with simulated data. For example, a generative adversarial network (GAN) might create a synthetic population dataset with realistic demographic and income distributions [26–28].
- Partially Synthetic Data: In partially synthetic data generation only certain attributes or records are synthesized. This approach allows a blend of real and synthetic data to balance utility and privacy.
- Hybrid Method: Combines original and synthetic data to achieve optimal trade-offs between utility and privacy.

PRIVACY RISKS AND MITIGATION

Types of Disclosure Risks

Identity Disclosure: Direct identification of an individual based on unique attribute combinations.

Attribute Disclosure: Inference of sensitive attribute values based on quasi-identifiers.

Inference Disclosure: Statistical inference of confidential information from released data.

Membership Disclosure: Determining whether an individual is part of a dataset.

Risk Assessment and Mitigation

Effective risk assessment involves simulating attacker scenarios to evaluate vulnerabilities. For example, an attacker with access to external datasets may attempt to link records using quasi-identifiers. Techniques like k-anonymity are foundational but have limitations, particularly in datasets with homogeneous sensitive attributes. Enhancements such as ℓ-diversity ensure that equivalence classes contain diverse sensitive values, while t-closeness measures the similarity between class distributions and the overall dataset to mitigate risks [29, 30].

DATA UTILITY AND TRADE-OFFS

The effectiveness of anonymization is closely tied to the usability of the anonymized data. Utility is assessed through various metrics, including:

Information Loss Metrics: measures the extent of distortion caused by anonymization. Techniques such as KL divergence, discernibility, and covariance analysis compare the original dataset with the anonymized version to evaluate differences (Figure 2).

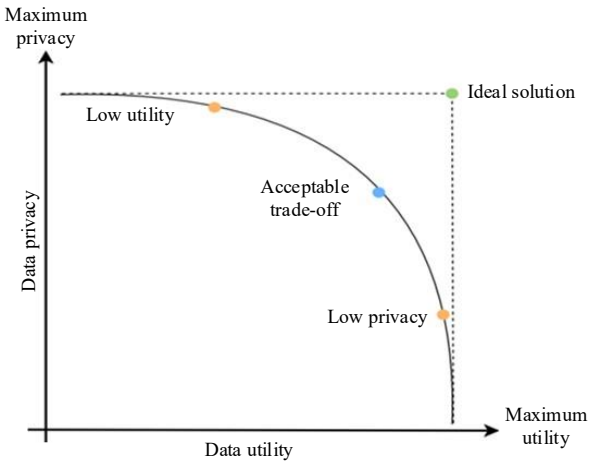

Figure 2. Trade-off between data privacy and utility.

Predictive Performance Metrics: assess how well machine learning models perform when trained on anonymized data. Common evaluation methods include precision, recall, F1-score, and mean squared error for regression-based tasks.

These metrics help balance the trade-off between privacy protection and data utility. For example, excessive generalization in k-anonymity can lead to substantial information loss, reducing the dataset's effectiveness for analytical and machine learning applications. Likewise, introducing too much noise may distort predictive models, resulting in biased or unreliable outcomes.

CASE STUDIES AND EXPERIMENTAL EVALUATIONS

Effectiveness of PPTs

Case studies demonstrate varying impacts of PPTs:

Non-Perturbative Methods: Effective in retaining utility for classification tasks but susceptible to attribute disclosure risks.

Perturbative Methods: Robust against re-identification attacks but may introduce biases affecting statistical analysis.

Synthetic Data: Emerging methods like GANs show promise in balancing privacy and utility but require careful calibration to avoid overfitting.

Implementation Tools

Tools such as ARX, Aircloak, and SDC Micro offer practical solutions for implementing PPTs, allowing users to customize anonymization processes and evaluate data utility.

OPEN ISSUES AND FUTURE DIRECTIONS

Key challenges include:

The approach focuses on several key aspects of data handling and privacy. First, it emphasizes dynamic data handling by developing PowerPoint presentations that can adapt to dataset modifications over time. It also addresses scalability for big data by tackling the curse of dimensionality in large datasets. Finally, the approach ensures privacy in collaborative environments with multiple data custodians by leveraging distributed systems, making it suitable for secure data sharing and processing in decentralized settings.

CONCLUSION

Privacy-preserving data publishing is crucial for secure data sharing and analysis. Although considerable advancements have been made, persistent challenges require the creation of flexible, scalable, and automated solutions. Future research should focus on comprehensive strategies that strike a balance between privacy, utility, and scalability, ensuring that data continues to serve as a valuable tool for innovation.

REFERENCES

1. Vanichayavisalsakul P, Piromsopa K. An evaluation of anonymized models and ensemble classifiers. In Proceedings of the 2018 2nd international conference on big data and internet of things. 2018 Oct 24; 18–22.
2. Sweeney L. k-anonymity: A model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst. 2002 Oct; 10(05): 557–70.
3. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. l-diversity: Privacy beyond k-anonymity. ACM Trans Knowl Discov Data. 2007 Mar 1; 1(1): 3–es.
4. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd international conference on data engineering. 2006 Apr 15; 106–115.
5. Xie L, Lin K, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. arXiv preprint arXiv:1802.06739. 2018 Feb 19.
6. Kanade DM, Sane SS. Evaluating the Effectiveness of Clustering-Based K-Anonymity and KNN Cluster for Privacy Preservation. Int J Intell Syst Appl Eng. 2023 Sep. 6; 11(11s): 85–93. [cited 2025 Sep. 23]. Available from: https://ijisae.org/index.php/IJISAE/article/view/3437
7. Reiter JP. Using CART to generate partially synthetic public use microdata. J Off Stat. 2005 Sep 1; 21(3): 441.
8. Hittmeir M, Ekelhart A, Mayer R. On the utility of synthetic data: An empirical evaluation on machine learning tasks. In Proceedings of the 14th international conference on availability, reliability and security. 2019 Aug 26; 1–6.
9. Majeed A, Lee S. Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data. Appl Intell. 2020 Aug; 50(8): 2555–74.
10. Senosi A, Sibiya G. Classification and evaluation of privacy preserving data mining: a review. 2017 IEEE AFRICON. 2017 Sep 18; 849–55.

11. Buratović I, Miličević M, Žubrinić K. Effects of data anonymization on the data mining results. In 2012 IEEE proceedings of the 35th international convention MIPRO. 2012 May 21; 1619–1623.
12. de Oliveira Silva H, Basso T, de Oliveira Moraes RL. Privacy and data mining: Evaluating the impact of data anonymization on classification algorithms. In 2017 IEEE 13th European Dependable Computing Conference (EDCC). 2017 Sep 4; 111–116.
13. Prasser F, Kohlmayer F, Kuhn KA. A benchmark of globally-optimal anonymization methods for biomedical data. In 2014 IEEE 27th international symposium on computer-based medical systems. 2014 May 27; 66–71.
14. Saranya K, Premalatha K, Rajasekar SS. A survey on privacy preserving data mining. In 2015 IEEE 2nd International Conference on Electronics and Communication Systems (ICECS). 2015 Feb 26; 1740–1744.
15. Bild R, Kuhn KA, Prasser F. Safepub: A truthful data anonymization algorithm with strong privacy guarantees. Proc Priv Enhanc Technol. 2018; 2018(1): 67–87.
16. El Emam K, Dankar FK. Protecting privacy using k-anonymity. J Am Med Inform Assoc. 2008 Sep 1; 15(5): 627–37.
17. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998 Aug 31; 20(8): 832–44.
18. Breiman L. Bagging predictors. Mach Learn. 1996 Aug; 24(2): 123–40.
19. Benbouzid D, Busa-Fekete R, Casagrande N, Collin FD, Kégl B. MultiBoost: a multi-purpose boosting package. J Mach Learn Res. 2012 Mar 1; 13(1): 549–53.
20. Wyner AJ, Olson M, Bleich J, Mease D. Explaining the success of Adaboost and random forests as interpolating classifiers. J Mach Learn Res. 2017; 18(48): 1–33.
21. Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: A new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 2006 Oct 31; 28(10): 1619–30.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011 Nov 1; 12: 2825–30.
23. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007 May 1; 9(03): 90–5.
24. Dankar FK, El Emam K, Neisa A, Roffey T. Estimating the re-identification risk of clinical data sets. BMC Med Inform Decis Mak. 2012 Jul 9; 12(1): 66.
25. Prasser F, Kohlmayer F, Lautenschläger R, Kuhn KA. Arx-a comprehensive tool for anonymizing biomedical data. In AMIA Annual Symposium Proceedings. 2014 Nov 14; 2014: 984.
26. Vanichayavisalsakul P, Piromsopa K. An evaluation of anonymized models and ensemble classifiers. In Proceedings of the 2018 2nd international conference on big data and internet of things. 2018 Oct 24; 18–22.
27. El Emam K. Guide to the de-identification of personal health information. CRC Press; Florida, USA. 2013 May 6.
28. Clifton C, Kantarcioglu M, Vaidya J. Defining privacy for data mining. In National science foundation workshop on next generation data mining. 2002 Nov 1; 1(26): 1.
29. Kanade DM, Patil LA. Internet of Things Security: Challenges and Opportunities. Int J Res Appl Sci Eng Technol. 2018; 6(5): 2188–2194.
30. Qian J, Li XY, Zhang C, Chen L, Jung T, Han J. Social network de-anonymization and privacy inference with knowledge graph model. IEEE Trans Dependable Secure Comput. 2017 Apr 25; 16(4): 679–92.

