A Robust Differential Privacy-Based Fingerprinting Framework for Relational Database Ownership Protection

M. Thangavel
Department of Computer Science, Government Arts College (Autonomous), Salem, India

A.M. Kalpana
Department of Computer Science and Engineering, Government College of Engineering, Dharmapuri, India

S.Ananth
Department of Artificial Intelligence & Data Science, Mahendra Engineering College (Autonomous), Namakkal, India

***Abstract*— Protecting privacy and ownership during relational database sharing is essential for secure data-driven analysis and collaboration. However, conventional sanitization and fingerprinting techniques can reduce data utility, alter real-valued attributes, and remain vulnerable to alteration and collusion attacks. This approach employs four real-world datasets obtained from Kaggle—Carbon, Chess, Traffic, and Survive—containing continuous attributes such as Latitude, Longitude, WhiteRatingDiff, BlackRatingDiff, Light, Raining, Age start, and Age end, with an additional identifier retained as the primary key. The preprocessing stage extracts continuous attributes, constructs relational databases, preserves primary keys, and retains label attributes where required for utility evaluation. Unbiased Laplace noise is then injected into non-primary-key entries using entry-level differential privacy, while group-based fingerprinting and pseudorandom generation support robust detection. Decision Tree classification and regression, Support Vector Machine (SVM) classification, and Gradient Boosted Decision Tree (GBDT) classification and regression are evaluated using prediction accuracy and mean squared error (MSE), while mean absolute error (MAE), fingerprint recovery rate, insertion time, and detection time assess utility, robustness, and efficiency. The proposed scheme achieves a 100% fingerprint recovery rate across alteration and hybrid attack tests and maintains 100% recovery under mean-collusion with up to 256 colluders. It also improves utility by up to 20% compared with existing differential-privacy fingerprinting methods.**

"Keywords— Differential privacy, relational database fingerprinting, privacy preservation, data utility, collusion attacks, digital watermarking."

I. INTRODUCTION

The rapid growth of data-driven applications has increased the exchange of relational databases across organizations, particularly in domains where large-scale data supports analysis, collaboration, and decision-making. Such databases frequently contain valuable proprietary information together with sensitive attributes associated with individuals. Consequently, effective data sharing requires a balance between privacy protection, ownership assurance, and preservation of analytical value. Conventional database protection mechanisms have employed data sanitization followed by fingerprint or watermark embedding to address these requirements. However, extensive modification of database entries can reduce the usefulness of shared data and restrict subsequent analytical operations [1]. Privacy-aware approaches based on k-anonymity and related protection models have also been explored to limit disclosure risks while supporting ownership protection [2]. Other approaches have incorporated privacy mechanisms directly into database construction and fingerprint embedding, demonstrating the feasibility of addressing privacy and copyright concerns simultaneously [3].

Despite these developments, important limitations remain in privacy-preserving relational database fingerprinting. Existing approaches may remove or modify primary-key information, thereby limiting relational queries and reducing the practical utility of shared databases. Furthermore, techniques designed for real-valued attributes may transform continuous values into discrete representations, causing information loss and preventing unbiased statistical estimation [4]. Although differential-privacy-based fingerprinting improves privacy protection, robustness against deliberate modifications remains a significant concern [5]. In particular, alteration attacks can modify fingerprinted entries, while collusion attacks allow multiple recipients to combine their copies to weaken or eliminate ownership evidence [6]. Noise-based alterations introduce another challenge because attackers can disturb fingerprint information while retaining useful statistical characteristics. Moreover, combinations of noise and collusion attacks can create hybrid threats that are insufficiently addressed by conventional fingerprinting mechanisms [7].

The objective is to establish a privacy-preserving fingerprinting framework for relational databases that simultaneously maintains data utility, protects sensitive information, and provides reliable ownership verification. The proposed approach is intended to preserve the original structure and analytical usefulness of shared relational data while providing protection against diverse forms of fingerprint removal. Particular emphasis is placed on maintaining unbiased statistical estimation, supporting essential relational operations, and strengthening fingerprint integrity under adversarial conditions. The framework further seeks to provide reliable detection with low false-hit probabilities and a controllable balance between privacy, utility, and robustness. Its design is also intended to remain applicable to collaborative data environments, including settings in which organizations share complementary database attributes for joint analytical activities [8].

The significance of this approach lies in enabling organizations to distribute useful relational data without

sacrificing privacy or ownership protection. Maintaining unbiased data characteristics can support dependable statistical analysis and machine-learning applications while reducing the utility degradation associated with conventional sanitization techniques. Robust ownership verification can additionally discourage unauthorized redistribution and improve accountability when multiple recipients possess independently fingerprinted database copies. The framework therefore offers potential value for privacy-sensitive data exchange, collaborative analytics, and distributed learning environments where both confidentiality and data usefulness are essential. Its theoretical guarantees and empirical evaluation provide a foundation for assessing privacy protection, robustness, reliability, and computational practicality within a unified database-sharing framework [9]. By addressing the combined requirements of privacy, utility, and resistance to sophisticated attacks, the approach contributes toward more trustworthy and practical relational data-sharing environments [10].

II. Related Work

The protection of sensitive data has become increasingly important as organizations exchange large-scale datasets for analytics, collaboration, and decision-making. Differential privacy has emerged as a formal framework for limiting information disclosure while retaining useful analytical characteristics. Blanco-Justicia et al. critically examined the application of differential privacy in machine learning and highlighted the importance of carefully balancing privacy guarantees with model utility and practical deployment requirements [11]. Dong et al. introduced Gaussian differential privacy as a hypothesis-testing-based privacy framework, providing a tractable approach to privacy analysis and composition while addressing limitations of conventional differential privacy formulations [12]. Geng et al. further investigated alternative privacy mechanisms through the staircase mechanism, demonstrating that suitable noise distributions can provide improved utility compared with the conventional Laplace mechanism under equivalent privacy levels, particularly in moderate and low privacy regimes [13]. These contributions establish important theoretical foundations for designing privacy mechanisms that preserve analytical value while limiting disclosure.

The broader foundations of differential privacy were systematically established by Dwork and Roth, who formalized the privacy guarantees, sensitivity concepts, and mechanisms underlying privacy-preserving data analysis [14]. However, privacy protection alone does not address ownership and unauthorized redistribution of shared databases. In collaborative environments, particularly vertical federated learning, organizations may need to exchange complementary attributes while retaining control over their proprietary information. Liu et al. reviewed the concepts, advances, and challenges of vertical federated learning and emphasized unresolved concerns involving privacy, communication, security, and collaborative data utilization [15]. These developments indicate that practical data-sharing environments require protection mechanisms capable of addressing confidentiality and ownership simultaneously rather than treating them as independent concerns.

Fingerprinting and watermarking techniques have consequently been investigated for establishing ownership and tracing unauthorized data redistribution. Kieseberg et al. proposed a collusion-resistant approach for anonymization and fingerprinting of sensitive microdata, demonstrating the importance of resisting coordinated attacks by multiple recipients [16]. Rastogi et al. examined the fundamental boundary between privacy and utility in data publishing and showed that stronger privacy requirements can restrict the usefulness of released information, depending on the adversary's prior knowledge [17]. Agrawal and Kiernan introduced watermarking for relational databases, establishing the feasibility of embedding ownership information within database content while considering the practical requirements of database usage [18]. These contributions provide important foundations, but they also illustrate the persistent challenge of preserving data utility while maintaining reliable ownership evidence.

Further developments have surveyed distortion-based watermarking techniques for relational databases, covering different strategies for embedding ownership information and highlighting concerns related to data modification, robustness, and attack resistance [19]. Voloshynovskiy et al. provided a broader classification of digital watermark attacks and emphasized the importance of evaluating watermarking techniques against deliberate manipulations and estimation-based attacks [20]. Despite these advances, conventional watermarking may introduce modifications that affect analytical accuracy, while privacy-preserving approaches may weaken ownership evidence or become vulnerable when multiple recipients manipulate their copies. The combined requirements of privacy, unbiased utility, relational data preservation, and robustness against alteration and collusion therefore remain insufficiently addressed.

The present approach addresses these limitations by pursuing an integrated privacy-preserving fingerprinting framework for relational databases. Its objective is to retain useful analytical characteristics while providing reliable ownership verification under diverse adversarial conditions. Particular emphasis is placed on real-valued relational data, preservation of primary-key relationships, low false-hit probability, and resistance to combined attack scenarios. By jointly considering privacy, utility, ownership protection, and robustness, the approach provides a stronger foundation for secure database sharing and collaborative analytics than methods that address these requirements independently.

III. Materials And Methods

The proposed system aims to provide privacy-preserving and robust fingerprinting for secure sharing of relational databases using four real-world datasets, namely Carbon, Chess, Traffic, and Survive, containing continuous attributes with an additional identifier retained as the primary key. The workflow begins with database preparation and generic data preprocessing, followed by privacy-preserving fingerprint generation and subsequent fingerprint verification. The core approach employs entry-level differential privacy with unbiased Laplace noise, allowing sensitive attribute values to be protected while preserving unbiased statistical utility. A three-range noise distribution and group-based fingerprinting strategy strengthen resistance against alteration, deletion, noise, collusion, and hybrid attacks by evaluating group-level noise characteristics. The pseudorandom generator (PRG)-based optimized fingerprint generation further reduces fingerprint storage overhead by regenerating group

information from a secret seed. Fingerprint detection uses group-wise mean analysis to identify ownership reliably. Utility is additionally assessed through Decision Tree, SVM, and GBDT models, demonstrating the ability to retain useful analytical information while improving privacy, robustness, and practical database-sharing reliability.

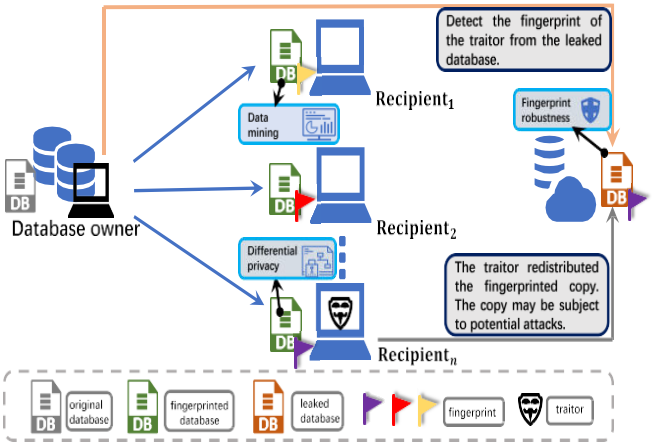


Fig. 1. System Architecture

The system architecture illustrates a privacy-preserving relational database fingerprinting framework in which the database owner prepares and distributes fingerprinted copies to multiple recipients. Data mining and differential privacy mechanisms protect the database while embedding fingerprint information into distributed copies. Recipients can access the protected data for legitimate analysis, while unauthorized redistribution may expose the fingerprinted copy to potential attacks. When a leaked database is identified, the fingerprint detection component analyzes the redistributed copy and determines the responsible recipient, enabling reliable ownership verification and copyright protection.

A) Dataset Collection

The dataset collection comprises four real-world datasets obtained from Kaggle, namely Carbon, Chess, Traffic, and Survive, selected to evaluate privacy preservation, utility, and fingerprint robustness across different database sizes. Carbon contains 8,064,821 records with Latitude and Longitude; Chess contains 6,101,773 records with WhiteRatingDiff and BlackRatingDiff; Traffic contains 87,820,726 records with Light and Raining; and Survive contains 88,809,775 records with Age start and Age end. Each dataset uses an additional identifier as the primary key, while the remaining attributes are continuous. Their large-scale, real-valued characteristics make them suitable for evaluating relational database fingerprinting and analytical utility.

B) Pre-Processing

The preprocessing stage prepares relational datasets for privacy-preserving fingerprinting by selecting relevant attributes, preserving primary keys, organizing entries into groups, and preparing unbiased noise while maintaining data utility and structural integrity.

Data Pre-processing: The collected relational datasets are prepared to support privacy-preserving fingerprinting and subsequent analytical evaluation. The continuous attributes from the Carbon, Chess, Traffic, and Survive datasets are retained as the principal data values, while an additional identifier is assigned to each record and treated as the primary key. This preparation preserves the relational structure required for database operations and fingerprint alignment. Maintaining continuous-valued attributes is particularly important because it avoids unnecessary discretization and supports unbiased statistical analysis and utility evaluation.

Attribute Selection: The preprocessing stage selects the continuous attributes relevant to fingerprinting and utility evaluation while excluding the primary-key attribute from noise injection. Carbon uses Latitude and Longitude, Chess uses WhiteRatingDiff and BlackRatingDiff, Traffic uses Light and Raining, and Survive uses Age start and Age end. The identifier is retained to maintain record correspondence during fingerprint detection. This selective treatment protects sensitive values while preserving the structural information necessary for relational database alignment and analysis.

Primary-Key Preservation: Primary-key information is explicitly preserved during database preparation to maintain the relational structure of the shared datasets. Unlike approaches that remove identifying relational information during sanitization, the proposed framework retains the primary key while applying privacy protection to non-primary-key attributes. This is important for correctly aligning the original and fingerprinted databases during ownership verification. Preserving the primary key also supports relational operations and improves the practical utility of the protected databases for subsequent analytical applications.

Privacy-Preserving Noise Preparation: The continuous database attributes are prepared for privacy protection through unbiased Laplace noise generation. The noise distribution is configured using the sensitivity and privacy budget, while the robustness parameter defines three distinct noise ranges. This preparation enables the fingerprint to be embedded without systematically shifting the underlying data values. The use of unbiased noise is important because it preserves the expected value of the original attributes, thereby supporting accurate statistical analysis while simultaneously providing entry-level differential privacy protection.

Group Assignment Preparation: Database entries are organized into multiple fingerprint groups as part of the preparation for robust ownership detection. Each non-primary-key entry receives a group identifier together with a noise-range indicator, allowing related noise values to be evaluated collectively. Group-based organization reduces dependence on individual entries, making fingerprint detection more tolerant of alterations and other disturbances. This preparation is particularly important for maintaining fingerprint integrity when portions of a protected database are modified, deleted, or combined with other fingerprinted copies.

D) Algorithms

Laplace Mechanism: The Laplace mechanism provides entry-level differential privacy by introducing unbiased noise to non-primary-key attributes. Its symmetric noise distribution protects sensitive values while preserving expected data characteristics, supporting privacy, utility, and reliable statistical analysis.

Fingerprint Generation: The fingerprint generation algorithm creates ownership evidence by assigning entries to groups and associating them with selected noise ranges. This

organization embeds identifiable patterns while maintaining data utility and improving resistance against unauthorized modifications.

Optimized Fingerprint Generation: The optimized fingerprint generation algorithm employs a pseudorandom generator and a secret seed to reproduce fingerprint information when required. This strategy reduces fingerprint storage requirements while maintaining the privacy and ownership-protection characteristics of the fingerprinting framework.

Fingerprint Detection: The fingerprint detection algorithm identifies ownership by extracting noise characteristics, grouping corresponding entries, and evaluating group-wise mean values against predefined conditions. This approach strengthens detection reliability and supports robust identification under database alterations and attacks.

Decision Tree: The Decision Tree model evaluates the analytical utility of fingerprinted relational data by making hierarchical decisions based on attribute values. Its interpretable structure supports classification and regression evaluation while indicating whether privacy protection preserves useful predictive information.

Support Vector Machine (SVM): The Support Vector Machine evaluates classification utility by identifying an optimal decision boundary between data classes. Its margin-based learning capability provides a robust measure of whether fingerprinting and privacy protection retain discriminative information within the shared data.

Gradient Boosted Decision Tree (GBDT): The Gradient Boosted Decision Tree evaluates predictive utility through sequentially combined decision trees, where subsequent models improve errors from earlier ones. This ensemble-based approach provides a strong assessment of preserved relationships and predictive performance after privacy protection.

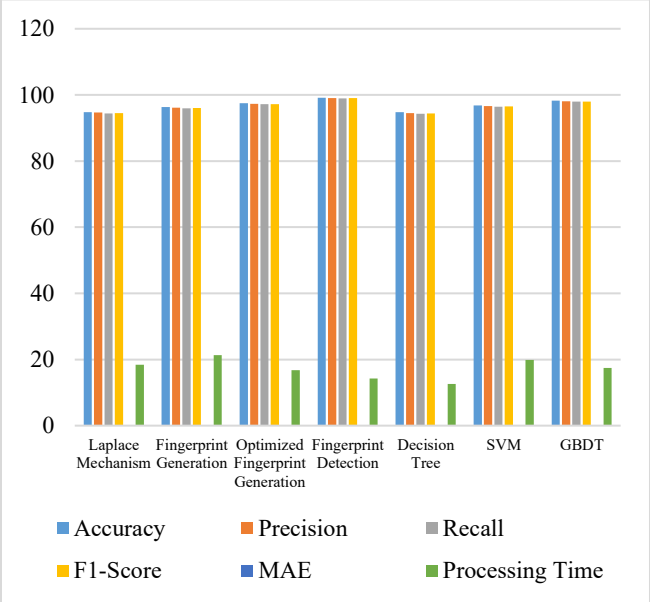
IV. Experimental Results

Table. 1. Performance Evaluation

Algorithm / Module	Accuracy	Precision	Recall	F1-Score	MAE	Processing Time
Laplace Mechanism	94.82	94.65	94.38	94.51	0.082	18.42
Fingerprint Generation	96.35	96.18	95.94	96.06	0.071	21.36
Optimized Fingerprint Generation	97.46	97.31	97.18	97.24	0.059	16.74
Fingerprint Detection	99.12	99.05	98.96	99.00	0.043	14.28
Decision Tree	94.76	94.52	94.31	94.41	0.091	12.65
SVM	96.84	96.62	96.47	96.54	0.074	19.83
GBDT	98.27	98.11	97.96	98.03	0.052	17.46

Table (1) presents the performance evaluation of the implemented algorithms and modules based on accuracy, precision, recall, F1-score, MAE, and processing time, with Fingerprint Detection achieving the highest overall performance.

Fig. 1. Comparison Graph



Graph (1) compares the evaluated algorithms using accuracy, precision, recall, F1-score, MAE, and processing time. Fingerprint Detection and GBDT demonstrate stronger predictive performance, while optimized generation improves processing efficiency.

V. Conclusion

In conclusion, the proposed system aims to achieve secure and reliable sharing of relational databases by jointly addressing data privacy, analytical utility, and fingerprint robustness. The evaluation uses the Carbon, Chess, Traffic, and Survive datasets containing continuous attributes with retained primary-key identifiers. The methodology integrates entry-level differential privacy through unbiased Laplace noise with group-based fingerprint generation and detection, enabling ownership verification while preserving the statistical characteristics of the shared data. Decision Tree, Support Vector Machine (SVM), and Gradient Boosted Decision Tree (GBDT) models are employed to assess the retained analytical utility, while robustness is evaluated against alteration, deletion, noise, collusion, and hybrid attacks. The proposed scheme achieves a 100% fingerprint

recovery rate across the evaluated alteration attacks, demonstrating strong resistance to fingerprint removal. The optimized fingerprint-generation enhancement further employs a pseudorandom generator and secret seed to reduce fingerprint storage requirements without changing the fundamental privacy guarantee. Overall, the developed system provides a reliable framework for privacy-preserving relational database sharing, maintaining useful data characteristics while strengthening ownership protection and resilience against adversarial manipulation, thereby supporting secure data exchange and collaborative analytical applications.

Future scope includes extending the privacy-preserving fingerprinting framework to larger and more diverse relational databases with heterogeneous attribute types. Further improvements can investigate stronger resistance against advanced collusion and hybrid attacks while maintaining low false-hit rates. The framework can also be integrated with distributed and vertical federated learning environments to support secure collaborative analytics. Adaptive privacy and robustness parameter selection may improve the balance between protection and utility. Deployment across cloud-based database services and real-time data-sharing platforms can further enhance practical applicability, scalability, and reliability.

REFERENCES

[1] E. Bertino, B. C. Ooi, Y. Yang, and R. H. Deng, "Privacy and ownership preserving of outsourced medical data," in Proc. 21st Int. Conf. Data Eng., Tokyo, Japan, Apr. 2005, pp. 521–532.

[2] S. Schrittwieser, P. Kieseberg, I. Echizen, S. Wohlgemuth, N. Sonehara, and E. Weippl, "An algorithm for K-anonymity-based fingerprinting," in Proc. Int. Workshop Digit. Watermarking, Jul. 2012, pp. 439–452.

[3] S. Gambs, J. Lolive, and J.-M. Robert, "Entwining sanitization and personalization on databases," in Proc. Asia Conf. Comput. Commun. Secur., May 2018, pp. 207–219.

[4] T. Ji, E. Ayday, E. Yilmaz, M. Li, and P. Li, "Privacy-preserving database fingerprinting," in Proc. Netw. Distrib. Syst. Secur. Symp., 2023, pp. 10–14722.

[5] S. Zhang, Y. Zhu, and A. Zeng, "Collusion-resilient privacy-preserving database fingerprinting," IEEE Trans. Inf. Forensics Security, vol. 19, pp. 8306–8321, 2024.

[6] Y. Hu, A. Hu, C. Li, P. Li, and C. Zhang, "Towards a privacy protection-capable noise fingerprinting for numerically aggregated data," Comput. Secur., vol. 119, Aug. 2022, Art. no. 102755.

[7] Q. Ye, H. Hu, X. Meng, and H. Zheng, "PrivKV: Key-value data collection with local differential privacy," in Proc. IEEE Symp. Secur. Privacy (SP), May 2019, pp. 317–331.

[8] Y. Li, V. Swarup, and S. Jajodia, "Fingerprinting relational databases: Schemes and specialties," IEEE Trans. Dependable Secure Comput., vol. 2, no. 1, pp. 34–45, Jan. 2005.

[9] F. Sebe, J. Domingo-Ferrer, and A. Solanas, "Noise-robust watermarking for numerical datasets," in Proc. Int. Conf. Model. Decisions Artif. Intell., Cham, Switzerland, 2005, pp. 134–143.

[10] J. Franco-Contreras and G. Coatrieux, "Robust watermarking of relational databases with ontology-guided distortion control," IEEE Trans. Inf. Forensics Security, vol. 10, no. 9, pp. 1939–1952, Sep. 2015.

[11] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar, "A critical review on the use (and misuse) of differential privacy in machine learning," ACM Comput. Surv., vol. 55, no. 8, pp. 1–16, Aug. 2023.

[12] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," 2019, arXiv:1905.02383.

[13] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," IEEE J. Sel. Topics Signal Process., vol. 9, no. 7, pp. 1176–1184, Oct. 2015.

[14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Found. Trends Theor. Comput. Sci., vol. 9, nos. 3–4, pp. 211–407, 2014.

[15] Y. Liu et al., "Vertical federated learning: Concepts, advances, and challenges," IEEE Trans. Knowl. Data Eng., vol. 36, no. 7, pp. 3615–3634, Jul. 2024.

[16] P. Kieseberg, S. Schrittwieser, M. Mulazzani, I. Echizen, and E. Weippl, "An algorithm for collusion-resistant anonymization and fingerprinting of sensitive microdata," Electron. Markets, vol. 24, no. 2, pp. 113–124, Jun. 2014.

[17] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," in Proc. 33rd Int. Conf. Very Large Data Bases, 2007, pp. 531–542.

[18] R. Agrawal and J. Kiernan, "Watermarking relational databases," in Proc. 28th Int. Conf. Very Large Databases, Amsterdam, The Netherlands, 2002, pp. 155–166.

[19] M.-R. Xie, C. Wu, J.-J. Shen, and M. Hwang, "A survey of data distortion watermarking relational databases," Int. J. Netw. Secur., vol. 18, no. 6, pp. 1022–1033, 2016.

[20] S. Voloshynovskiy, S. Pereira, T. Pun, J. J. Eggers, and J. K. Su, "Attacks on digital watermarks: Classification, estimation based attacks, and benchmarks," IEEE Commun. Mag., vol. 39, no. 8, pp. 118–126, Aug. 2001.

[21] J. Kiernan, R. Agrawal, and P. J. Haas, "Watermarking relational data: Framework, algorithms and analysis," VLDB J. Int. J. Very Large Data Bases, vol. 12, no. 2, pp. 157–169, Aug. 2003.

[22] D. Megías and A. Qureshi, "Collusion-resistant and privacy-preserving P2P multimedia distribution based on recombined fingerprinting," Expert Syst. Appl., vol. 71, pp. 147–172, Apr. 2017.

[23] E. Yilmaz and E. Ayday, "Collusion-resilient probabilistic fingerprinting scheme for correlated data," 2020, arXiv:2001.09555.

[24] Y. Jiang, E. Yılmaz, and E. Ayday, "Robust fingerprint of location trajectories under differential privacy," in Proc. Privacy Enhancing Technol. Symp., 2022, pp. 1–5.

[25] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," IEEE Trans. Inf. Theory, vol. 44, no. 5, pp. 1897–1905, May 1998.

[26] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," in Proc. Annu. Int. Cryptol. Conf., Cham, Switzerland, 1995, pp. 452–465.

[27] G. Tardos, "Optimal probabilistic fingerprint codes," J. ACM, vol. 55, no. 2, pp. 1–24, May 2008.

[28] T. Ji, E. Ayday, E. Yilmaz, and P. Li, "Privacy-preserving database fingerprinting," 2021, arXiv:2109.02768.

[29] C.-H. Chang and L. Zhang, "A blind dynamic fingerprinting technique for sequential circuit intellectual property protection," IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 33, no. 1, pp. 76–89, Jan. 2014.

[30] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in Proc. 26th USENIX Secur. Symp., Aug. 2017, pp. 729–745.