World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Is your personal data safer to disclose? An exploratory analysis of reidentification risk

Kelani Bandara *

Information Technology Division, The Open University of Sri Lanka, Nugegoda, Sri Lanka.

World Journal of Advanced Research and Reviews, 2025, 28(01), 1004-1013

Publication history: Received on 16 August 2025; revised on 25 September 2025; accepted on 29 September 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.1.3301

Abstract

The ubiquity of personal data generated through human-centric devices—such as smartphones and wearable technologies—has intensified concerns over individual privacy and reidentification risk. Despite the implementation of data protection regulations that mandate strict disclosure controls, numerous studies have demonstrated the persistent vulnerability of de-identified datasets. In this study, I conduct a comprehensive risk assessment on a publicly available de-identified dataset, focusing on two dimensions of uniqueness-based risk: sample uniqueness and population uniqueness. The analysis reveals that, under an adversarial knowledge scenario, the probability of correctly reidentifying an individual record reaches 0.35. Furthermore, over 45% of records are susceptible to reidentification when seven quasi-identifiers are known, while even four attributes suffice to reidentify more than 9% of records. The proposed estimation framework achieves accuracy exceeding 75%, outperforming several baseline models. These findings highlight the limitations of existing anonymization techniques and underscore the need for more robust disclosure control mechanisms, particularly for datasets that involve sensitive personal attributes.

Keywords: Population Uniqueness; Sample Uniqueness; Reidentification Risk Estimation; Reidentification Example; Reidentification Attack

1. Introduction

Personal data has become a critical resource across domains that rely on insights into human behavior, preferences, and activities. Consequently, the demand for personal data has surged, with open and publicly available datasets playing a pivotal role in advancing research and driving societal innovation. However, the widespread accessibility of such data raises serious concerns regarding individual privacy and the risk of reidentification, even when datasets have undergone de-identification procedures.

Globally, more than 80% of countries have enacted personal data protection laws [1], often supported by substantial financial penalties for noncompliance [2]. Regulatory authorities continue to strengthen these legal frameworks in response to evolving threats and lessons learned from prior breaches [3]. Despite these efforts, the risk of reidentifying individuals from ostensibly anonymized data remains substantial [4]. Classic cases, such as the well-known reidentification of Massachusetts Governor William Weld's medical records [5], revealed the inadequacy of early anonymization methods and inspired foundational models, including k-anonymity. Since then, both reidentification techniques and privacy-preserving strategies have grown increasingly sophisticated, with recent advances in generative AI further amplifying the risk landscape.

In this study, we critically assess the effectiveness of disclosure controls by examining uniqueness-based reidentification risks in a publicly available de-identified dataset. Specifically, we focus on two dimensions of uniqueness risk: sample uniqueness and population uniqueness, which together capture how easily individual records can be distinguished within both the observed dataset and the broader population. By employing sampling with

* Corresponding author: Kelani Bandara

replacement to simulate realistic adversarial scenarios, we estimate the likelihood of reidentification under varying levels of background knowledge.

Our findings reveal that when an adversary possesses partial knowledge of an individual record, the probability of correct reidentification reaches 35%. Moreover, more than 45% of records can be reidentified when seven quasi-identifiers are known, while over 9% remain vulnerable even when only four are available. The proposed estimation framework achieves accuracy above 75%, outperforming existing models used in similar risk assessments. These results underscore significant vulnerabilities in the disclosure controls currently applied to the dataset, highlighting the need for more robust risk estimation and mitigation strategies.

The key contributions of this study are as follows

- A systematic evaluation of reidentification risks in students' performance in a real-world de-identified dataset.
- Exploration of imperfections accompanied by a demonstration of how risk estimation accuracy can be significantly improved.

The rest of this paper is organized as follows. Section 2 reviews prior work on reidentification risk estimation and relevant background concepts. Section 3 outlines the methodology used for assessing twofold uniqueness risk and evaluating estimation accuracy. In Section 4, I present empirical results, including how risk varies across attributes, sample sizes, and totality assumptions. Section 5 discusses the broader implications of my findings for privacy-preserving data publishing. Finally, Section 6 concludes the paper with a summary and directions for future work.

2. Background

Assessing reidentification risk in ostensibly anonymized datasets has emerged as a central concern in privacy research. One of the earliest and most widely cited examples is Latanya Sweeney's reidentification of Massachusetts Governor William Weld's medical records in the late 1990s [5]. By linking an anonymized dataset released by the Massachusetts Group Insurance Commission (GIC) with publicly available voter registration records, Sweeney demonstrated that 87% of individuals could be uniquely identified using only ZIP code, birth date, and gender.

Since then, numerous high-profile studies have underscored the fragility of anonymization techniques. In 2008, Narayanan and Shmatikov reidentified Netflix users by correlating anonymized movie ratings with IMDb reviews [6]. De Montoya et al. further revealed that 90% of individuals in a credit card transaction dataset could be reidentified using only four purchases [7], and that 95% of mobile phone users could be uniquely identified with just four spatio-temporal points [8]. More recently, Rocher et al. estimated that 99.98% of Americans are uniquely identifiable using only 15 demographic attributes, raising critical concerns about the robustness of current privacy-preserving strategies [9].

Table 1 Notable reidentification examples in the literature

No	Dataset	Re-identified Information
1	Massachusetts GIC medical data [5]	Governor William Weld's medical records
2	Netflix Prize movie ratings [6]	Several Netflix users, some identified by name
3	Mobile phone location data (from a European telco) [8]	95% of individuals re-identified using 4 patio-temporal points
4	Credit card transaction dataset [7]	90% of people re-identified with 4 purchases
5	U.S. Census + demographic data [9]	99.98% of Americans are unique with 15 demographic attributes
6	AOL search query dataset (2006) [10]	Thelma Arnold, a 62-year-old widow, was publicly identified
7	1000 Genomes Project + genealogy databases [11]	Re-identified individuals in the anonymized DNA dataset
8	Facebook profiles + public data [12]	Partial SSNs of individuals inferred

9	U.S. Census microdata [13]	An estimated>60% of the U.S. population is re-identifiable
10	Washington State hospital discharge data [10]	Re-identified patients, including diagnosis and treatment info
11	AOL search queries (2006 release) [14]	Identified several users, including Thelma Arnold again
12	NYC Taxi Trip Data [15]	Identified individuals' nightlife and affair patterns
13	Browser history (via CSS/JS sniffing) [16]	Inferred social network membership, visited sites
14	Mobility traces from wireless access logs [17]	Linked people to Twitter/Facebook profiles
15	Genomic + clinical data [17]	Linked genetic data to hospital records
16	Cell tower logs (mobile carrier data) [18]	Estimated user home/work location and identity
17	Deep learning models trained on private data [19]	Determined whether an individual's data was used in training
18	Mobility traces from wireless sensor networks [20]	Re-identified movement paths of university students
19	Smart meter data [21]	Inferred user habits and potentially identity

Table 1 summarizes landmark reidentification case studies that have shaped the discourse on data privacy and anonymity, along with their corresponding references.

In response to such risks, personal data protection regulations across jurisdictions have introduced legal and technical disclosure controls. Major frameworks—including the European Union's General Data Protection Regulation (GDPR) [21], the California Consumer Privacy Act (CCPA) [22], and India's Digital Personal Data Protection Act (DPDP) [23]—recommend or mandate techniques such as anonymization, de-identification, and pseudonymization as safeguards against misuse. These frameworks also emphasize principles such as data minimization, purpose limitation, explicit consent, and restrictions on cross-border transfers. However, the specific implementation of technical measures varies across jurisdictions. Table 2 compares the treatment of anonymization and pseudonymization in selected data protection acts.

Table 2 Disclosure controls in data protection regulations

Act	**Country/Region**	**Anonymization**	**Pseudonymization**
GDPR [24]	European Union	Recommended	Recommended
UK DPA 2018 [25]	United Kingdom	Recommended	Required for sensitive data
DPDP Act 2023 [26]	India	Recommended	Not defined
CCPA / CPRA [27]	California, USA	De-identification recommended	Not defined (implied)
LGPD [28]	Brazil	Recommended	Encouraged
PIPEDA [29]	Canada	De-identification recommended	Not defined
Privacy Act 1988	Australia	De-identification recommended	Not defined (implied)
POPIA [30]	South Africa	De-identification recommended	Not defined
PDPA [31]	Singapore	De-identification recommended	Partially defined
NZ Privacy Act 2020	New Zealand	Encouraged	Not defined

KVKK [32]	Turkey	Partially defined	Encouraged
PDPA 2010 [33]	Malaysia	Partially defined	Partially defined
APPI (2022) [34]	Japan	Recommended	Recommended
PIPL & Data Security Law 2021	China	Recommended	Recommended
Data Protection Act 2021 [35]	Kenya	Recommended	Not defined
Data Protection Act [36]	Nigeria	Implied	Not defined
Federal Law on Personal Data [37]	Russia	Recommended	Recommended

Despite their widespread adoption, these technical and legal measures often fail to prevent reidentification—particularly when datasets contain unique combinations of quasi-identifiers or when adversaries possess auxiliary information. Anonymization seeks to mask quasi-identifiers [38], de-identification removes direct identifiers [39], and pseudonymization replaces identifiers while maintaining data relationships [40]. Yet none of these methods adequately address risks arising from uniqueness in sensitive attributes, especially when modern linkage and analytical techniques are applied.

This study addresses these shortcomings by empirically evaluating reidentification risks in a de-identified dataset. Despite the application of standard disclosure controls, our analysis demonstrates that adversarial models can exploit uniqueness to reidentify individuals with high confidence. These findings underscore the limitations of current regulatory frameworks and highlight the urgent need for stronger technical safeguards to counter the evolving threats of reidentification.

3. Methodology

This section presents our two-fold methodology for estimating re-identification risk: (1) sample uniqueness risk, and (2) population uniqueness risk. Our approach aligns with established practices in disclosure risk estimation, as seen in leading journals such as the Journal of Privacy and Confidentiality and the Journal of Official Statistics.

3.1. Estimating Sample Uniqueness Risk

Sample uniqueness refers to the degree to which an individual record can be distinguished based on a combination of quasi-identifying attributes. This approach follows the standard framework in the disclosure risk literature.

Let the dataset consist of n quasi-identifiers $a_1, a_2, \dots, a_n$ and let a record have attribute values $\alpha_1, \alpha_2, \dots, \alpha_n$. The uniqueness U of this record is defined as:

U = count of rows where $(a_1 = \alpha_1, a_2 = \alpha_2, , , , , a_n = \alpha_n)$ ……. (1)

The sample uniqueness risk r, i.e., the probability that this record is unique and therefore identifiable, is computed as:

$$r = \frac{1}{U} \text{……..} \quad (2)$$

The risk is maximal when U = 1, indicating that the record is unique in the dataset. As emphasized in risk analysis literature, such records are the most vulnerable to re-identification. Therefore, we focus on records with U = 1 to highlight the worst-case scenarios in the absence of appropriate safeguards.

3.2. Estimating Population Uniqueness

While sample uniqueness assesses risk within a given dataset, population uniqueness estimates the likelihood that a record is also unique in the broader population. This approach reflects techniques adopted in high-impact publications.

We construct an empirical approximation of the population uniqueness distribution by aggregating uniqueness counts over multiple samples drawn from a larger dataset. Let x = [a1 = α1, a2 = α2, an = αn] be a record. We define:

$$f(X = x) = \left(\frac{1}{m}\sum_{i=1}^{m} U_i\right) \quad \ldots\ldots\ldots\ldots \quad (3)$$

where Ui is the uniqueness of record x in the i-th sample, and m is the total number of samples considered. As with sample uniqueness, we highlight records for which f(X) = 1, indicating high population-level risk.

3.3. Accuracy of Population Uniqueness Estimation

To validate our population uniqueness estimation, we adopt a simulation-based approach. A known dataset is treated as the population, and we repeatedly draw random sub-samples from it. We then compare the estimated uniqueness from each subsample to the actual uniqueness within the full dataset. Figure 3 summarizes the estimation accuracy.

Accuracy A is computed as

$$A = \frac{(count(\hat{U}=U)}{t} X100 \quad \ldots\ldots\ldots\ldots \quad (4)$$

were

- Uˆ is the estimated uniqueness from sub-samples,
- U is the actual uniqueness from the population,
- t is the total number of evaluated records.

This metric quantifies the extent to which the estimation aligns with the ground truth and serves as a benchmark for evaluating re-identification risk estimation methods.

4. Results

This section presents the results of the proposed model for re-identification risk estimation and discusses its implications when applied to a real-world dataset. The evaluation encompasses both sample and population uniqueness risks, as well as accuracy analysis and attribute-level risk evaluation.

4.1. Dataset Description

We used the Students Performance Dataset, which comprises detailed records of 2,392 high school students, including demographic factors, study habits, parental involvement, extracurricular activities, and academic performance. For our analysis, we selected the following attributes: Sequential, Parental Education, School Type, Locale, Internet Access, Extracurricular Activities, Part-time Job, and Gout.

Categorical variables were numerically encoded to facilitate computation. To ensure statistical robustness, we generated 10 random samples of 1,000 records each by shuffling the dataset prior to each draw.

4.2. Sample Uniqueness Risk

Sample uniqueness was computed using Equation (1). Figure 1 illustrates the percentage of high-risk records (i.e., records with uniqueness U = 1) across 10 samples.

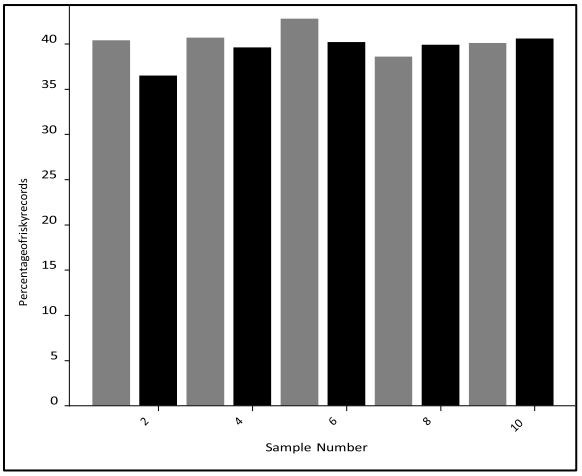

Figure 1 Percentage of high-risk records based on sample uniqueness

As shown in Figure 1, more than 35% of records in each sample are classified as high-risk. The highest risk was observed in Sample 5, while Sample 2 had the lowest.

4.3. Population Uniqueness Risk

Population uniqueness was estimated using Equation (3) by aggregating uniqueness values over multiple samples. Figure 2 presents the percentage of high-risk records based on population uniqueness.

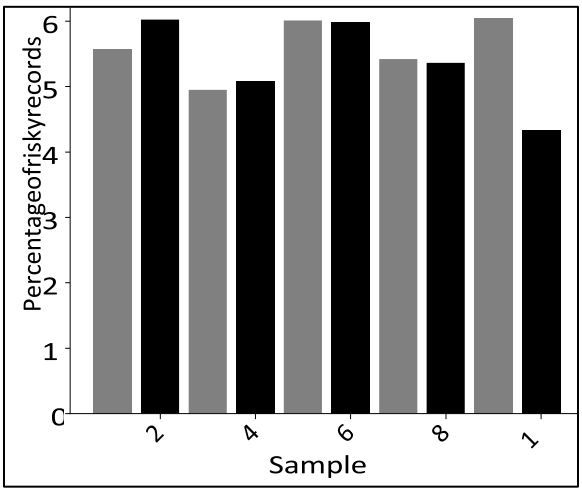

Figure 2 Percentage of high-risk records based on population uniqueness

More than 45% of records were identified as unique at the population level. Sample 2 exhibited the highest risk of population uniqueness, while Sample 10 had the lowest.

4.4. Accuracy of Population Uniqueness Estimation

We validated the accuracy of our population uniqueness estimation by treating the entire dataset as the population and comparing the estimated uniqueness from sub-samples with the actual values. Figure 3 shows the estimation accuracy across 10 runs.

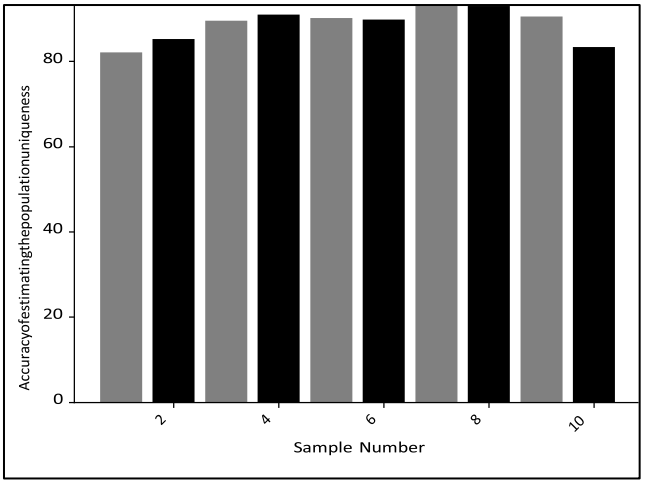

Figure 3 Accuracy of population uniqueness estimation across samples

The results demonstrate that estimation accuracy exceeds 80% in all samples. This level of reliability suggests that the proposed method provides a robust approximation of true population uniqueness and outperforms several existing risk estimation models [9].

4.5. Attribute-Level Risk Analysis

To evaluate how attribute selection affects re-identification risk, we repeated the population uniqueness analysis using a reduced set of four attributes: Sequential, Parental Education, School Type, and Locale. Results are depicted in Figure 4.

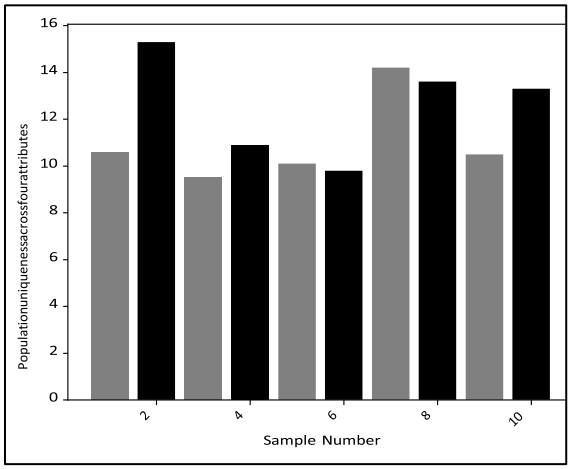

Figure 4 Percentage of high-risk records using four selected quasi-identifiers

Even with just four attributes, more than 9% of records remained unique across all samples. Sample 2 again exhibited the highest re-identification risk, while Sample 3 had the lowest. This demonstrates that even a small number of quasi-identifiers can yield significant disclosure risks.

5. Discussion

The findings of our study reveal a substantial reidentification risk within the dataset, despite the absence of direct identifiers. Based on the sample uniqueness analysis, more than 35% of records in each of the ten random samples were unique. This implies that if an adversary possesses background knowledge about an individual's attributes, the probability of correctly re-identifying their record is approximately 0.35. Such a high probability indicates a serious privacy vulnerability.

The population uniqueness analysis further exacerbates this concern, with more than 45% of records identified as unique when broader population-level distributions are considered. These results align with prior literature, which emphasizes that re-identification risk is often underestimated when relying solely on sample-level statistics.

The robustness of our findings is supported by the high estimation accuracy, which exceeds 80% across all samples. By repeating our analysis of over ten independent runs and validating against a known population baseline, we ensure the methodological reliability of our estimation approach. This reinforces the validity of our conclusions and aligns with best practices recommended in recent studies on disclosure risk estimation.

Although the dataset complies with legal data disclosure requirements—such as the removal of direct identifiers under data protection laws like the GDPR and the Sri Lankan Personal Data Protection Act—the results highlight that such minimal de-identification strategies are insufficient. The dataset includes quasi-identifiers such as Sequential, Parental Education, school Type,

5.1. Locale, Internet Access, Extracurricular

Part-time job, and Gout, all of which are attributes that can be known or inferred by an adversary. Consequently, the risk of acquaintance-based re-identification is considerable.

Even when the analysis was limited to just four attributes—Sequential, Parental Education, school Type, and Locale—more than 9% of records were found to be unique across all samples. This illustrates that even a limited set of quasi-identifiers can pose significant privacy risks, particularly in cases where adversaries possess additional knowledge.

These findings suggest that legal compliance alone does not guarantee effective anonymization. The persistence of high uniqueness rates, even under attribute reduction, underscores the need for stronger, risk-based anonymization techniques. Our results support the call for adopting advanced disclosure control methods such as k-anonymity, l-diversity, t-closeness, or differential privacy, particularly when releasing datasets in the public domain.

In summary, our analysis demonstrates that the current disclosure controls applied to the dataset are inadequate for mitigating re-identification risks. We recommend that data controllers complement legal compliance with quantitative risk assessment and employ context-aware anonymization strategies that account for real-world adversarial knowledge.

6. Conclusion

This study assessed the re-identification risk in a publicly available dataset on high school student performance by evaluating two core dimensions of disclosure risk: sample uniqueness and population uniqueness. Our methodology was designed for statistical reliability, incorporating repeated random sampling, accuracy estimation, and population-level generalization.

The analysis revealed a substantial proportion of high-risk records across both sample and population evaluations. When nine quasi-identifiers were considered, a significant number of records were found to be unique, posing a credible risk of re-identification. Even when the attribute set was reduced to only four quasi-identifiers—Sequential, Parental Education, School Type, and Locale—over 9% of records remained unique across all samples.

Importantly, while the dataset complies with relevant data protection legislation by removing direct identifiers, our findings demonstrate that such compliance does not necessarily equate to effective anonymization. The persistence of high uniqueness rates highlights a critical gap between legal disclosure controls and actual privacy protection, especially in scenarios involving acquaintance-based attacks.

The accuracy of our population uniqueness consistently exceeded 80%, validating the robustness of our method. This provides a strong foundation for recommending our approach as a practical and reliable framework for assessing re-identification risk in microdata releases.

In conclusion, current de-identification practices, though legally sufficient, may not be adequate for safeguarding individual privacy. We recommend that data controllers adopt risk-based anonymization techniques and conduct quantitative reidentification risk assessments prior to publishing data. Future work should explore the integration of advanced methods such as differential privacy, synthetic data generation, and context-aware anonymization to achieve a stronger balance between data utility and privacy.

Compliance with ethical standards

Disclosure of conflict of interest

The author declares that she has no conflict of interest to disclose.

References

[1] G. Greenleaf, "Global data privacy laws 2023: 162 national laws and 20 bills (Feb 10, 2023)," 181 Privacy Laws and Business International Report (PLBIR) 1, 2-4, UNSW Law Research Paper No. 23-48, 2023.

[2] J. Wolff and N. Atallah, "Early GDPR penalties: Analysis of implementation and fines through May 2020," Journal of Information Policy, vol. 11, pp. 63–103, 2021.

[3] A. K. Saraswat and V. Meel, "Protecting data in the 21st century: Challenges, strategies and future prospects," Information technology in industry, vol. 10, no. 2, pp. 26–35, 2022.

[4] M. Finck and F. Pallas, "They who must not be identified—distinguishing personal from non-personal data under the GDPR," International Data Privacy Law, vol. 10, no. 1, pp. 11–36, 2020.

[5] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[6] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," arXiv preprint cs/0610105, 2006.

[7] De Montjoye, "Unique in the shopping mall: On the reidentifiability of credit card metadata," Science, vol. 347, no. 6221, pp. 536–539, 2015.

[8] "Unique in the crowd: The privacy bounds of human mobility," Scientific reports, vol. 3, no. 1, pp. 1–5, 2013.

[9] Rocher, "Estimating the success of re-identifications in incomplete datasets using generative models," Nature Communications, vol. 10, no. 1, pp. 1–9, 2019.

[10] Barbaro, "A face is exposed for aol searcher no. 4417749," New York Times, vol. 9, no. 2008, p. 8, 2006.

[11] Gymrek, "Identifying personal genomes by surname inference," Science, vol. 339, no. 6117, pp. 321–324, 2013.

[12] Golle, "Secure conjunctive keyword search over encrypted data," in International Conference on Applied Cryptography and Network Security. Springer, 2004, pp. 31–45.

[13] L. Sweeney, "Discrimination in online ad delivery," Communications of the ACM, vol. 56, no. 5, pp. 44–54, 2013.

[14] A. Tockar, "Riding with the stars: Passenger privacy in the NYC taxicab dataset," Neustar Research, September, vol. 15, no. 6, 2014.

[15] Wondracek, "A practical attack to de-anonymize social network users," in 2010 IEEE Symposium on Security and Privacy. IEEE, 2010, pp. 223–238.

[16] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," Journal of biomedical informatics, vol. 37, no. 3, pp. 179–192, 2004.

[17] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in Proceedings of the 17th annual international conference on Mobile computing and networking, 2011, pp. 145–156.

[18] Shokri, "Membership inference attacks against machine learning models," in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 3–18.

[19] Haeberlen, "Peer review: Practical accountability for distributed systems," ACM SIGOPS operating systems review, vol. 41, no. 6, pp. 175– 188, 2007.

[20] Xu, "N-doped nanoporous Co3O4 nanosheets with oxygen vacancies as oxygen evolving electrocatalysts," Nanotechnology, vol. 28, no. 16, p. 165402, 2017.

[21] G. D. P. Regulation, "Gdpr. 2019," 2019.

[22] E. Illman and P. Temple, "California Consumer Privacy Act," The Business Lawyer, vol. 75, no. 1, pp. 1637–1646, 2019.

[23] N. Gupta and A. George, "Digital personal data protection act, 2023: Charting the future of india's data regulation," in Data Governance and the Digital Economy in Asia. Routledge, 2025, pp. 34–53.

[24] "General Data Protection Regulation (GDPR)," 2018. [Online].

[25] Available: https://gdpr-info.eu/

[26] D. P. Act, "Data Protection Act 2018," [online] GOV. UK, 2018.

[27] Malhotra and U. Malhotra, "Putting interests of digital nagriks first: Digital personal data protection (DPDP) Act 2023 of India," Indian Journal of Public Administration, vol. 70, no. 3, pp. 516–531, 2024.

[28] "California Consumer Privacy Act (CCPA)," 2024, California Privacy Protection Agency. [Online]. Available: https://cppa.ca.gov/faq.html

[29] Canedo, "Proposal of an implementation process for the Brazilian general data protection law (LGPD)." in ICEIS (1), 2021, pp. 19–30.

[30] Jaar and P. E. Zeller, "Canadian privacy law: The Personal Information Protection and Electronic Documents Act (PIPEDA)," Int l. In-House Counsel J., vol. 2, p. 1135, 2008.

[31] Staunton, "Protection of personal information act 2013 and data protection for health research in south africa," International Data Privacy Law, vol. 10, no. 2, pp. 160–179, 2020.

[32] B. Chik, "The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform," Computer Law & Security Review, vol. 29, no. 5, pp. 554–575, 2013.

[33] Sevinc¸ and N. Karabulut, "A review on the personal data protection authority of Turkey," Akademik Hassasiyetler, vol. 7, no. 13, pp. 449– 472, 2020.

[34] M. Yusoff, "The Malaysian Personal Data Protection Act 2010: A Legislation Note," NZJPIL, vol. 9, p. 119, 2011.

[35] Okada, "On the revision of Japanese personal information protection system in 2021," Ph.D. dissertation, Waseda University, 2023.

[36] Kevins and K. Brian, "Defining data protection in Kenya: Challenges, perspectives and opportunities," Perspectives and Opportunities (November 7, 2022), 2022.

[37] Adeoti, "A new era of data protection and privacy; unveiling innovations & identifying gaps in the Nigeria Data Protection Act of 2023," Unveiling Innovations & Identifying Gaps in the Nigeria Data Protection Act of 2023.

[38] A. Gurkov, "Personal data protection in Russia," The Palgrave Handbook of Digital Russia Studies, pp. 95–113, 2021.

[39] Lison, "Anonymization models for text data: State of the art, challenges and future directions," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4188–4203.

[40] Garfinkel et al., De-identification of Personal Information:. US Department of Commerce, National Institute of Standards and Technology, 2015.

[41] Kohlmayer, "Pseudonymization for research data collection: is the juice worth the squeeze?" BMC medical informatics and decision making, vol. 19, pp. 1–7, 2019.