Estimating Re-identification Risk with Greater Accuracy: A Sample–Population Uniqueness Approach

P.L.M.K. Bandara
The Open University of Sri Lanka
Nawala, Nugegoda
Sri Lanka

ABSTRACT

The increasing availability of microdata for research and policy analysis raises critical concerns about the risk of re-identification, particularly when datasets contain quasi-identifying attributes. This paper proposes a novel model for estimating re-identification risk through the joint analysis of sample and population uniqueness, and evaluates its performance against the conventional log-linear approach. The methodology combines repeated sampling with aggregation of uniqueness measures to estimate population-level risk, with precision, recall, and F1-score employed to validate accuracy. Empirical evaluation was conducted on three real-world datasets: student performance, insurance claims, and car purchasing inquiries. The results demonstrate that re-identification risk is strongly dataset-dependent. The insurance dataset exhibited near-total uniqueness at both sample and population levels, highlighting an elevated probability of re-identification and the urgent need for robust disclosure controls. In contrast, the student performance and car purchasing datasets showed lower, though still considerable, proportions of unique records. Across all datasets, the proposed model closely aligned with true population counts and consistently outperformed the log-linear model in terms of accuracy.

The findings underscore the inadequacy of traditional risk estimation methods for modern, high-dimensional datasets. The proposed model provides a more accurate and reliable framework for disclosure risk assessment, offering valuable guidance for data custodians and policymakers in balancing data utility with privacy protection.

Keywords

Population Uniqueness, Sample Uniqueness, Reidentification Risk Estimation, Reidentification Example, Reidentification Attack

1. INTRODUCTION

Personal data serves as a critical resource across domains that rely on human behavior, preferences, and activities. As a result, demand for personal data has surged, with open and publicly available datasets playing a pivotal role in advancing research and societal innovation. However, the increasing accessibility of such data has introduced serious concerns regarding individual privacy and the risk of reidentification, even when datasets have been de-identified. Globally, over 80% of countries have enacted personal data protection laws [1], often backed by substantial financial penalties for noncompliance [2]. Regulatory authorities continue to strengthen these legal frameworks based on evolving threats and lessons from prior breaches [3]. Despite these efforts, the risk of reidentifying individuals from ostensibly anonymized data remains significant [4]. Classic examples, such as the well-known reidentification of Massachusetts Governor William Weld's medical records [5], have demonstrated the inadequacy of early anonymization methods and inspired foundational models such as k-anonymity. Since then, both reidentification techniques and privacy-preserving strategies have become more sophisticated, with generative AI technologies further exacerbating the risk landscape. In this study, the effectiveness of disclosure controls was critically assessed by examining uniqueness-based reidentification risks in a publicly available de-identified dataset. Specifically, two dimensions of uniqueness risk: sample uniqueness and population uniqueness, were investigated, which capture how easily individual records can be distinguished in both observed and broader population contexts. By sampling with replacement and simulating realistic adversarial scenarios, the likelihood of reidentification with varying degrees of background knowledge was estimated. The analysis shows that if an adversary possesses partial knowledge about an individual record, there is a greater possibility of correct reidentification. The estimation framework demonstrates accuracy above 75%, outperforming existing models used for similar assessments. These findings indicate substantial vulnerabilities in the current disclosure controls applied to the dataset and signal the need for more robust risk estimation and mitigation techniques. The key contributions of this study are:

(1) A systematic evaluation of reidentification risks—both sample and population uniqueness—in a real-world de-identified dataset.

(2) Identification of limitations in existing disclosure controls and 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 become a critical concern in privacy research. As discussed earlier, one of the earliest and most widely cited examples of such risk is the reidentification of Massachusetts Governor William Weld's medical records by Latanya Sweeney in the late 1990s [5]. This attack exploited linkages between an anonymized dataset released by the Massachusetts Group Insurance Commission (GIC) and publicly available voter registration data, revealing that 87% of individuals could be uniquely identified using a combination of ZIP code, birth date, and gender.

Since then, numerous high-profile reidentification studies have demonstrated the fragility of anonymization techniques. For example, in 2008, Narayanan and Shmatikov reidentified Netflix users by correlating anonymized movie rating patterns with IMDb reviews [6]. In another landmark study, de Montjoye et al. showed that 90% of individuals in a credit card transaction dataset could be reidentified using only four purchases [7]. Similarly, they demonstrated that 95% of individuals' mobile phone trajectories could be uniquely identified using four spatio-temporal points [8]. More recently, Rocher et al. estimated that 99.98% of Americans are uniquely identifiable using just 15 demographic attributes, raising further concerns about the efficacy of current privacy-preserving techniques [9].

Table 1 summarizes notable reidentification case studies that have shaped the discourse on data privacy and anonymity.

Table 2 shows the reference of each row of the table 1.

To mitigate such risks, personal data protection regulations across jurisdictions have introduced legal and technical disclosure controls. Most major frameworks—including the European Union's General Data Protection Regulation (GDPR)1[22], the California Consumer Privacy Act (CCPA) [23], and India's Digital Personal Data Protection Act (DPDP) [24]—recommend or mandate techniques such as *anonymization*, *de-identification*, and *pseudonymization* as safeguards against misuse.

Commonly endorsed principles include data minimization, purpose limitation, explicit consent, and restrictions on cross-border data transfers. However, the specific implementations of technical controls vary by jurisdiction. Table 3 summarizes the use of anonymization and pseudonymization in selected data protection acts.

While widely adopted, these technical and legal measures often fall short in preventing reidentification—particularly when datasets contain unique combinations of quasi-identifiers or when adversaries possess auxiliary information. Anonymization focuses on masking quasi-identifiers [39], de-identification removes direct identifiers [40], and pseudonymization replaces identifiers while retaining relational data structures [41]. Yet, none of these approaches fully address risks arising from uniqueness in sensitive attributes, especially in the presence of modern analytical and linkage techniques.

This study highlights these shortcomings by empirically evaluating reidentification risks in a de-identified dataset. Despite the application of standard disclosure controls, how adversarial models can leverage uniqueness to reidentify individuals with high confidence is demonstrated in this paper. The findings underscore the need for revisiting existing regulatory frameworks and enhancing technical safeguards against evolving threats.

3. METHODOLOGY

This section describes the methodology adopted for estimating sample and population uniqueness risks. The log–linear model is then introduced for comparative evaluation, followed by the approach used to assess the accuracy of the estimations.

3.1 Estimating Sample Uniqueness Risk

Sample uniqueness refers to the extent to which a record can be distinguished from others in the dataset by a combination of quasi-identifying attributes. This follows the conventional approach in disclosure risk literature.

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

$$U = \text{count of rows where } (a_1 = \alpha_1, a_2 = \alpha_2, \ldots, a_n = \alpha_n). \tag{1}$$

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

$$r = \frac{1}{U}. \tag{2}$$

The maximum risk occurs when $U = 1$, indicating that the record is unique within the dataset. Such records are most vulnerable to re-identification. Consistent with prior research, we therefore focus on records with $U = 1$ to highlight the worst-case risk scenario in the absence of protective measures.

3.2 Estimating Population Uniqueness

Population uniqueness measures the likelihood that a record is unique in the broader population rather than in a specific sample. This approach is consistent with strategies proposed in prior influential studies.

An empirical approximation of population uniqueness is obtained by aggregating uniqueness counts across multiple random samples drawn from a larger dataset. Let $x = [a_1 = \alpha_1, a_2 = \alpha_2, \ldots, a_n = \alpha_n]$ denote a record. We define:

$$f(X = x) = \text{round}\left(\frac{1}{m}\sum_{i=1}^{m} U_i\right), \tag{3}$$

where U_i is the uniqueness of record x in the i-th sample, and m is the number of samples considered. As in the case of sample uniqueness, our analysis focuses on records with $f(X) = 1$, which represent those at greatest risk of population-level re-identification.

3.3 Comparison Model: Log–linear Framework

The log–linear model, a widely adopted framework for estimating population uniqueness, was selected for comparison with the proposed approach. This subsection outlines the estimation process.

For population estimation, an offset for the sampling fraction π can be incorporated as:

$$\log(\mathbb{E}[f_c]) = \log(\pi) + \log(\mu_c), \tag{4}$$

where μ_c denotes the expected population frequency for cell c. The fitted values $\hat{\mu}_c$ are then used to estimate disclosure risk measures such as individual risk $1/\hat{\mu}_c$ for sample uniques.

Table 1.
Notable Reidentification Examples in the Literature

No	Dataset	Re-identified Information
1	Massachusetts GIC medical data	**Governor William Weld's medical records**
2	Netflix Prize movie ratings	Several Netflix users, some identified by name
3	Mobile phone location data (from a European telco)	95% of individuals re-identified using 4 spatio-temporal points
4	Credit card transaction dataset	90% of people re-identified with 4 purchases
5	U.S. Census + demographic data	99.98% of Americans unique with 15 demographic attributes
6	AOL search query dataset (2006)	**Thelma Arnold, a 62-year-old widow, was publicly identified**
7	1000 Genomes Project + genealogy databases	Re-identified individuals in the anonymized DNA dataset
8	Facebook profiles + public data	Partial SSNs of individuals inferred
9	U.S. Census microdata	Estimated >60% of U.S. population re-identifiable
10	Washington State hospital discharge data	Re-identified patients including diagnosis and treatment info
11	AOL search queries (2006 release)	Identified several users including Thelma Arnold again
12	NYC Taxi Trip Data	Identified individuals' nightlife and affair patterns
13	Browser history (via CSS/JS sniffing)	Inferred social network membership, visited sites
14	Mobility traces from wireless access logs	Linked people to Twitter/Facebook profiles
15	Genomic + clinical data	Linked genetic data to hospital records
16	Cell tower logs (mobile carrier data)	Estimated user home/work location and identity
17	Deep learning models trained on private data	Determined whether an individual's data was used in training
18	Mobility traces from wireless sensor networks	Re-identified movement paths of university students
19	Smart meter data	Inferred user habits and potentially identity

Table 2.
References for the Notable Reidentification Examples in the Literature

No	Reference
1	[5]
2	[6]
3	[8]
4	[7]
5	[9]
6	[10]
7	[11]
8	[12]
9	[13]
10	[14]
11	[10]
12	'[15]
13	[16]
14	[17]
15	[18]
16	[18]
17	[19]
18	[20]
19	[21]

3.3.1 *Estimating Record-Level Individual Risk.*

For a given cell c in the contingency table, let $\hat{N}_c$ denote the estimated population frequency obtained from the fitted log–linear model. The *individual risk* for records in that cell is defined as:

$$r_c = \frac{1}{\hat{N}_c}. \tag{5}$$

This expression provides a population-level disclosure risk estimate, capturing the probability that a sample-unique record corresponds to a population-unique. Reporting this measure alongside individual risk offers a comprehensive evaluation of re-identification vulnerability.

3.4 Evaluation Metrics: Precision, Recall, and F1-score

To evaluate the effectiveness of the uniqueness estimation, we frame the task as a classification problem and compute precision, recall, and the F1-score using the confusion matrix:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

where:

—TP = True Positives (records correctly predicted as population-unique),
—FP = False Positives (records incorrectly predicted as unique),
—FN = False Negatives (population-unique records missed by the model).

Precision measures the reliability of uniqueness predictions, recall assesses the model's ability to capture all truly unique records, and the F1-score provides a harmonic mean of both, offering a balanced evaluation metric.

Table 3.
Disclosure Controls in Data Protection Regulations

Act	Country/Region	Anonymization	Pseudonymization
GDPR [25]	European Union	Recommended	Recommended
UK DPA 2018 [26]	United Kingdom	Recommended	Required for sensitive data
DPDP Act 2023 [27]	India	Recommended	Not defined
CCPA / CPRA [28]	California, USA	De-identification recommended	Not defined (implied)
LGPD [29]	Brazil	Recommended	Encouraged
PIPEDA [30]	Canada	De-identification recommended	Not defined
Privacy Act 1988	Australia	De-identification recommended	Not defined (implied)
POPIA [31]	South Africa	De-identification recommended	Not defined
PDPA [32]	Singapore	De-identification recommended	Partially defined
NZ Privacy Act 2020	New Zealand	Encouraged	Not defined
KVKK [33]	Turkey	Partially defined	Encouraged
PDPA 2010 [34]	Malaysia	Partially defined	Partially defined
APPI (2022) [35]	Japan	Recommended	Recommended
PIPL & Data Security Law 2021	China	Recommended	Recommended
Data Protection Act 2021 [36]	Kenya	Recommended	Not defined
Data Protection Act [37]	Nigeria	Implied	Not defined
Federal Law on Personal Data [38]	Russia	Recommended	Recommended

4. RESULTS

This section presents the outcomes of the proposed model for re-identification risk estimation and discusses its implications when applied to real-world datasets. The evaluation encompasses both sample- and population-level uniqueness risks, followed by comparative accuracy analysis.

4.1 Description of the Datasets

Re-identification risk was evaluated on more than ten datasets, of which three with the highest risk levels are presented in this paper: the *Students' performance dataset*, the *Insurance claim dataset*, and the *Car purchasing dataset*.

The students' performance dataset contains records of 2,392 high school students, including demographic factors, study habits, parental involvement, extracurricular activities, and academic performance. For this analysis, the following attributes were selected: `SES_Quartile`, `ParentalEducation`, `SchoolType`, `Locale`, `InternetAccess`, `Extracurricular`, `PartTimeJob`, and `GoOut`.

The insurance claim dataset comprises 1,339 records, including attributes such as age, sex, number of children, smoking status, region, and claim amount. The attributes considered for re-identifiability assessment were `age`, `sex`, `number of children`, `region`, and `smoking status`, with `age`, `sex`, and `region` treated as quasi-identifiers.

The car purchasing dataset consists of 1,000 records of individuals' car purchasing inquiries, including sex, age, salary level, and purchase decision. In this case, all attributes were considered due to their re-identifiability, with `age` and `sex` serving as quasi-identifiers.

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

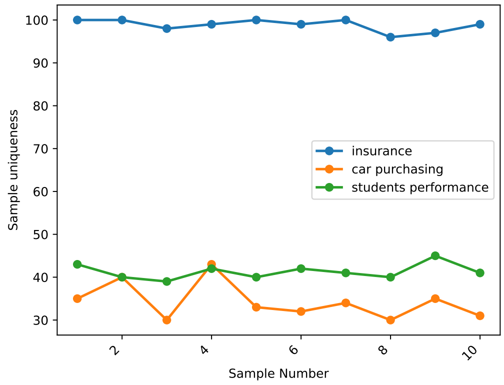

Fig. 1
Percentage of high-risk records based on sample uniqueness in the three datasets.

The figure shows the distribution of sample-unique records across dataset samples. The x-axis represents the sequence of samples, while the y-axis denotes the percentage of records unique within each sample. A higher proportion of unique records corresponds to greater disclosure risk, as these records are more vulnerable to re-identification. The results indicate that the insurance dataset presents a markedly higher disclosure risk, with nearly all records being unique across samples. This finding underscores the need for stringent disclosure control measures prior to release. By contrast, the car purchasing and student performance datasets exhibit lower levels of sample uniqueness, with approximately 35–40% of records classified as high-risk. Moreover, the car purchasing dataset demonstrates a marginally lower risk profile than the student performance dataset, reflecting variability in susceptibility across domains.

4.3 Population Uniqueness Risk

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

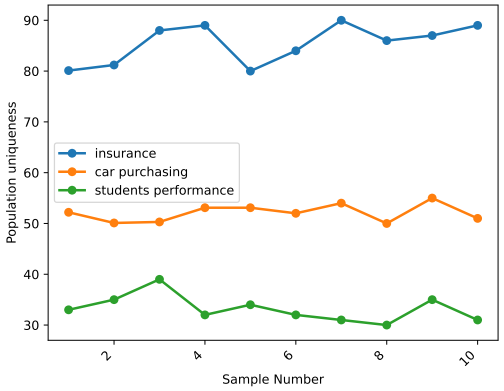

Fig. 2
Percentage of high-risk records based on population uniqueness.

The results illustrate the proportion of records that remain unique at the population level. The x-axis denotes the sample number, while the y-axis shows the percentage of population-unique records. The insurance dataset is particularly vulnerable, with almost all records retaining uniqueness, indicating an exceptionally high probability of re-identification. In contrast, the car purchasing and student performance datasets exhibit reduced levels of population uniqueness, with approximately 35–40% of records remaining unique. The car purchasing dataset demonstrates slightly lower risk than the student performance dataset. These findings highlight that population-level uniqueness amplifies disclosure risks beyond sample-level estimates, underscoring the necessity for robust anonymisation strategies, particularly for highly sensitive datasets such as insurance claims.

4.4 Comparison with the Log-linear Model

The performance of the proposed model was compared with the log–linear model across all three datasets, following the methodology outlined in Section 3. Only high-risk records were included in the comparison, as the study focuses on records most vulnerable to disclosure.
As shown in Figure 3, the estimates produced by the proposed model closely match the true population count of unique records across all datasets. By contrast, the log–linear model systematically underestimates uniqueness, indicating that the proposed model offers more accurate and reliable estimation of population uniqueness.

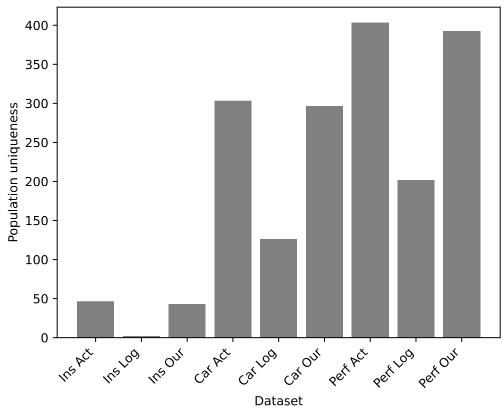

Fig. 3
Comparison of the proposed model with the log-linear model.

4.5 Accuracy of Population Uniqueness Estimation

The accuracy of the proposed model was assessed using precision, recall, and F1-score, as described in Section 3. Equivalent metrics were also computed for the log-linear model for comparison.

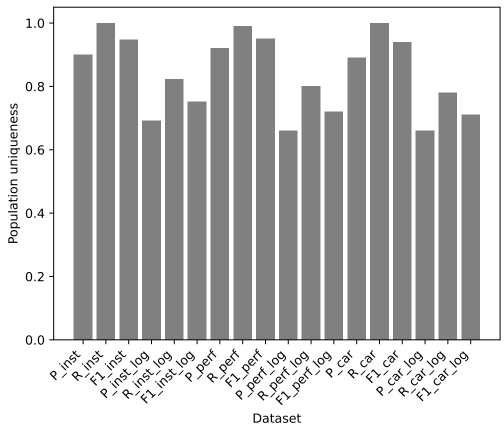

Fig. 4
Comparison of precision, recall, and F1-score between the proposed model and the log-linear model.

Figure 4 demonstrates that the proposed model consistently outperforms the model across all three metrics. Specifically, it achieves higher precision, recall, and F1-scores, reflecting greater reliability in correctly identifying high-risk records. These results confirm that the proposed approach provides a more accurate and robust estimation of population uniqueness, thereby offering stronger support for disclosure risk assessment than the conventional log-linear model.

5. DISCUSSION

The findings of this study provide important insights into the estimation of re-identification risk, both at the sample and population levels. By comparing the proposed model with the conventional log-linear approach, several key observations emerge.

First, the analysis of *sample uniqueness* highlights that datasets vary considerably in their susceptibility to re-identification. The insurance dataset demonstrated an exceptionally high level of disclosure risk, with nearly all records being unique across samples. This result underscores the challenges associated with health- or claim-related datasets, where demographic and behavioural attributes combine to form distinctive patterns. Conversely, the student performance and car purchasing datasets showed relatively lower, though still substantial, proportions of unique records (approximately 35–40%). These findings reinforce the view that disclosure risk is highly domain-dependent, and even datasets with moderate uniqueness can still pose significant privacy threats without adequate safeguards.

Second, the *population uniqueness* analysis revealed that re-identification risks become amplified when considering the broader population. The insurance dataset, in particular, exhibited near-total population uniqueness, indicating that the risk of re-identification extends far beyond the sampled data. This highlights a critical limitation of sample-only assessments: they may systematically underestimate the true magnitude of disclosure risk. By contrast, the car purchasing and student performance datasets displayed comparatively lower population uniqueness, yet their non-trivial proportions of high-risk records further emphasize the necessity for protective measures.

Third, the comparison with the log-linear model demonstrated that the proposed model provides a more accurate and reliable estimation of population uniqueness. While the log-linear framework has long been regarded as a benchmark in disclosure risk analysis, our results show that it consistently underestimates the number of high-risk records. The proposed model's closer alignment with true population counts suggests its greater suitability for contemporary datasets, particularly those characterized by complex attribute interactions.

Finally, the evaluation of *accuracy metrics* (precision, recall, and F1-score) further validates the robustness of the proposed model. Across all metrics, the proposed model outperformed the log-linear benchmark, thereby offering not only improved accuracy but also more dependable identification of high-risk records. This is especially important in practice, where misclassification of unique records may result in underestimated risk and insufficient anonymisation.

Taken together, these findings contribute to the ongoing discourse on privacy-preserving data release. They demonstrate that traditional models may no longer be sufficient for capturing the nuanced risks present in modern, high-dimensional datasets. The proposed model offers a promising alternative that balances computational tractability with accuracy, thereby equipping policymakers, data custodians, and researchers with a more effective tool for assessing and mitigating re-identification risk. Future work may extend these findings by exploring additional domains, incorporating adversarial knowledge into risk estimation, and evaluating how disclosure control techniques such as k-anonymity, l-diversity, and differential privacy interact with the proposed framework.

6. CONCLUSION

This study presented a novel approach for estimating re-identification risk by analysing both sample and population uniqueness, and benchmarking the results against the widely used log-linear model. The findings demonstrate that the proposed model provides a closer approximation to the true number of unique records, thereby addressing a key limitation of traditional methods, which tend to underestimate disclosure risk.

The empirical results across three diverse datasets highlight several important conclusions. First, re-identification risk is highly domain-dependent: the insurance dataset was found to be particularly vulnerable, while student performance and car purchasing datasets exhibited lower but still substantial levels of uniqueness. Second, the proposed model consistently outperformed the log-linear framework in terms of accuracy, as validated through precision, recall, and F1-score metrics. These results confirm that the proposed model offers a more reliable basis for disclosure risk assessment, especially in high-dimensional, real-world datasets.

From a practical perspective, the study underscores the need for robust anonymisation strategies prior to data release, particularly for domains such as health and insurance where near-total uniqueness is prevalent. By providing a more accurate and comprehensive estimation of risk, the proposed model equips data custodians and policymakers with stronger evidence for implementing effective disclosure control measures.

Future research could extend this work by integrating additional data modalities, considering adversarial knowledge scenarios, and evaluating the interaction between the proposed model and widely adopted privacy-preserving techniques such as k-anonymity, l-diversity, and differential privacy. Such investigations would further advance the development of rigorous, evidence-based frameworks for safeguarding personal data in an era of growing data availability and analytical power.

7. 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] e. 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] e. Rocher, “Estimating the success of re-identifications in incomplete datasets using generative models,” *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[10] M. Barbaro, T. Zeller, and S. Hansell, “A face is exposed for aol searcher no. 4417749,” *New York Times*, vol. 9, no. 2008, p. 8, 2006.

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

[12] A. Acquisti and R. Gross, "Predicting social security numbers from public data," *Proceedings of the National academy of sciences*, vol. 106, no. 27, pp. 10 975–10 980, 2009.

[13] e. Golle, "Secure conjunctive keyword search over encrypted data," in *International conference on applied cryptography and network security*. Springer, 2004, pp. 31–45.

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

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

[16] e. Wondracek, "A practical attack to de-anonymize social network users," in *2010 ieee symposium on security and privacy*. IEEE, 2010, pp. 223–238.

[17] 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.

[18] 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.

[19] e. Shokri, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

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

[21] e. Xu, "N-doped nanoporous co3o4 nanosheets with oxygen vacancies as oxygen evolving electrocatalysts," *Nanotechnology*, vol. 28, no. 16, p. 165402, 2017.

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

[23] E. Illman and P. Temple, "California consumer privacy act," *The Business Lawyer*, vol. 75, no. 1, pp. 1637–1646, 2019.

[24] 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.

[25] "General Data Protection Regulation (GDPR)," 2018. [Online]. Available: https://gdpr-info.eu/

[26] D. P. Act, "Data protection act 2018," *[online] GOV. UK.*, 2018.

[27] C. 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] e. Canedo, "Proposal of an implementation process for the brazilian general data protection law (lgpd)." in *ICEIS (1)*, 2021, pp. 19–30.

[30] D. 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] e. 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] W. 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] İ. Sevinç and N. Karabulut, "A review on the personal data protection authority of turkey," *Akademik Hassasiyetler*, vol. 7, no. 13, pp. 449–472, 2020.

[34] Z. M. Yusoff, "The malaysian personal data protection act 2010: A legislation note," *NZJPIL*, vol. 9, p. 119, 2011.

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

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

[37] E. 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] e. Lison, "Anonymisation 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] S. Garfinkel *et al.*, *De-identification of Personal Information:*. US Department of Commerce, National Institute of Standards and Technology, 2015.

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