The Risk-Utility Tradeoff for Data Privacy Models

M. Moein Almasi, Taha R. Siddiqui, Noman Mohammed, Hadi Hemmati
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, CA R3T2K8
Email: {almasi, trsid, noman, hemmati}@cs.umanitoba.ca

***Abstract*—Nowadays with growth of information technologies, organizations are constantly collecting information about individuals. Public availability of these datasets can considerably benefit the society. To ensure data privacy of a released dataset, various privacy models have been introduced. While many privacy models and techniques have been proposed for data sanitization, the area of sanitized data evaluation has received less attention. This paper investigates the four most well-known data privacy models: k-anonymity, l-diversity, t-closeness, and ϵ-differential privacy. We evaluate the data utility (usefulness of sanitized data) and the disclosure risk (re-identification risk of an individual) of the sanitized data for each model. We use a combination of several data utility and risk metrics to measure the impact of a privacy parameter (e.g., k, ϵ) on a particular privacy model. This enables us to compare the risk-utility tradeoff of semantic privacy models such as ϵ-differential privacy to the early syntactic models such as k-anonymity on the same scale. We used the Adult dataset from the UCI machine learning repository to conduct our experiments. Experimental results show that ϵ-differential privacy outperforms other privacy models in terms of both data utility and disclosure risk.**

***Keywords*—Privacy models, data utility, data disclosure risk, data sharing, data sanitization**

I. Introduction

Gradually, various entities such as research organizations and businesses are making their data, which are collected from individuals, publicly available. The availability of such data can significantly benefit society and in particular medical research, such as identifying causes of certain diseases and effectiveness of the treatment. However, most of these databases contain confidential information of individuals which cannot be published due to the risk of privacy breach. Each record in the database consists of some sensitive attributes (i.e., disease and income) and some quasi-identifiers (i.e., zip code, age and sex) that maybe combined together in order re-identify an individual. Therefore, these datasets must be sanitized before public release.

There have been many proposals for privacy models [1], [2], [3], [4] that attempt to formalize the notion of data privacy. k-anonymity [5], [6] is the first notable proposed model in the literature for relational databases and it uses the concept of generalization to provide privacy. The core idea is to ensure that there are at least k records with respect to quasi-identifiers. l-diversity [2] is then introduced to address the limitations of k-anonymity. In particular, l-diversity ensures that each equivalent class of the sanitized dataset is well-diverged so that an adversary is unable to launch homogeneity or background knowledge attack. The privacy model t-closeness [3] further refines the concept of diversity and requires that the distribution of the sensitive values of each equivalent class be as close as to the overall distribution of the dataset. Dwork *et al.* [4] has introduced ϵ-differential privacy which ensures that adding or removing a single individual record does not significantly impact the outcome of any an analysis There are a number of other privacy models [7] with little variations.

Challenges. Data sanitization techniques, prior to data release, modify an original dataset into a distorted version in order to decrease the disclosure risk while keeping the data utility as high as possible. A fundamental problem in privacy preserving data publishing is identifying the right tradeoff between disclosure risk and utility. While many privacy models and techniques have been proposed for data sanitization, the area of sanitized data evaluation has received less attention. The problem of evaluating risk-utility tradeoff of the sanitized data has several challenges. Achieving the right risk-utility tradeoff requires answering the following questions:

1) *What parameter should be selected for a particular privacy model?*
2) *What is the risk-utility tradeoff for various privacy models with different parameters?*

This paper proposes an analytical framework that overcomes the above challenges. In particular, we evaluate four most studied privacy models (as mentioned before) in terms disclosure risk and data utility. Our framework can also be used to evaluate other privacy models.

Current Technique. Cormode *et al.* [8] is the only known work that addressed these challenges by introducing a general notion of empirical privacy and utility, and compared traditional privacy models with differential privacy by varying the respective parameters of each model. They concluded that the difference between these models is quite less dramatic than it has been assumed.

Our Contributions. Unlike [8], we have defined the notion of risk in terms of uniqueness which is the most commonly used risk metric employed for real-life applications [9], [10]. In addition, we have utilized multiple utility metrics: KL-divergence, average equivalence class size, and height (see Section III for more discussion) for evaluating risk-utility tradeoff of various privacy models.

The proposed framework studies the tradeoff between disclosure risk and data utility of various privacy models by comparing them on a single scale. This allows a data owner to determine the effectiveness of various privacy models and facilitates the choosing of a privacy parameter for a particular privacy model. Finally, we conducted extensive experiments

978-1-5090-2914-3/16/$31.00 ©2016 IEEE

using real-life dataset from the UCI machine learning repository. Experimental results show that ϵ-differential privacy outperforms other privacy models in terms of both data utility and disclosure risk.

Organization. The rest of the paper is organized as follows. Section II reviews related works. Descriptions of privacy and utility metrics and the proposed risk-utility tradeoff framework are described in Section III. Section IV experimentally evaluates various privacy models using our framework in a single scale. Section V concludes the paper and points out some directions for future work.

II. Related Work

Cormode *et al.* [8] proposed a unified framework to compare various privacy models by formalizing the notion of empirical privacy and utility. They concluded that no particular privacy model has an advantage over other models in terms of risk-utility tradeoff. Their work [8] is the closest risk-utility tradeoff evaluation to our proposal with the difference that they have used single risk (privacy breach increase) and utility (relative query error) metrics while in this work we have examined the risk-utility tradeoff with respect to multiple metrics (details are presented in section III). Moreover, we measure risk using uniqueness which is the most accepted notion for evaluating disclosure risk [9].

Li *et al.* [11] utilized modern portfolio theory for financial investment and proposed a privacy-utility tradeoff model for data publishing. They have concluded that it is inappropriate to directly compare privacy with utility. Feinberg [12] used log-linear model to measure the risk-utility tradeoff for releasing contingency table with differential privacy guarantee. Risk was measured in terms of the protection for small counts in a contingency table. Utility was measured with respect to noise added to provide differential privacy. Khokhar *et al.* [13] measured risk-utility tradeoff by quantifying privacy in terms of monetary value. They propose a cost model that takes into consideration the monetary cost due to potential privacy breaches.

Moreover, Duncan *et al.* [14] introduced R-U Confidentiality Map as an illustration to trace the impact of several de-identification methods and their parameters. Teng *et al.* [15] compared k-anonymity and randomization technique by adapting R-U confidentiality map. Loukides *et al.* [16] also used R-U Confidentiality Map [17] to measure the risk and utility for transaction data sharing. Domingo-Ferrer *et al.* [18] measured the risk and utility of sanitized data using information theory. Dasgupta *et al.* [19] introduced a visualization approach to measure the risk-utility tradeoff various privacy models. Though closely related, all these works (except [8]) don't compare all privacy models in a unified framework (Figure 5 in Section IV).

III. Methodology

In this section, we present our method for analyzing risk-utility tradeoff of sanitized data. We have defined the risk-utility tradeoff by considering privacy as an individual concept (privacy of each individual must be protected) and utility as an aggregate concept (utility is gained if knowledge about a large group of individuals is learnt).

For each sanitized dataset, we measure its disclosure risk and data utility. First, we analyze the impact of a privacy parameter of a model (i.e., k of k-anonymity model) on disclosure risk and data utility, separately. Second, we obtain a sets of (risk (R), utility (U)) points for each sanitized dataset. We then plot the (R, U) points on a 2-dimensional graph like R-U Confidentiality Map [14], where x-axis represents a privacy model's disclosure risk (average re-identification risk) and y-axis depicts its data utility (information loss).

In order to measure disclosure risk and data utility, we need to fix metrics. In this paper, we have utilized multiple well-known and industry wide accepted data utility and disclosure risk metrics. In the next two subsections, we present these metrics.

A. Data Utility Metrics

In this section, we present several data utility metrics: KL-divergence, normalized average equivalence class size, and height.

1) Kullback–Leibler Divergence [20]: Kullback–Leibler or KL-divergence is a metric for finding the distance between two frequency distributions. In the context of utility, it is used to find the distance between distribution of a sensitive attribute in an equivalence class and the distribution of the same attribute in the whole dataset. In other words, it denotes the amount of information loss when the distribution of a sensitive attribute in the whole dataset is used to approximate the distribution of the same attribute in an equivalence class. In our experiment we have utilized KL-divergence to measure the distance between distributions of sensitive values of original (Q) and sanitized (P) data sets.

$$D(KL) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

KL-divergence measures the logarithmic differences between discrete probability distributions of P and Q for all i (absolute continuity).

2) Normalized Average Equivalence Class Size [6]: Average equivalence class size measures information loss based on the size of the equivalence classes resulting from data sanitization. The intuition behind this metric is that the bigger the size of an equivalence class, the smaller the utility of the sanitized data. Thus, it measures the utility loss resulting from the generalization or suppression of values.

$$C(AVG) = \frac{\frac{T}{H}}{K} \tag{2}$$

Here, T is total number of records, H is total equivalent classes and K is the privacy model constraint.

3) Height [21]: This metric takes into consideration the height of the hierarchy tree of all quasi-identifier attributes. Figure 1 shows sample hierarchy trees for attributes *Job* and *Age*. It measures information loss based on sum of the applied generalization level. Utility depends on generalization height for different attributes in the QD.

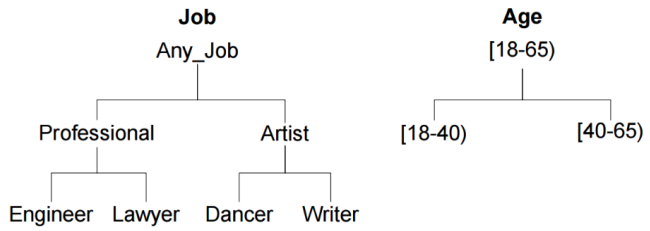

Fig. 1. Sample Hierarchy Tree for Generalization

$$H = \sum_i \frac{L_i}{H_i - 1} \tag{3}$$

Here, H_i is the height of the hierarchy tree for attribute i and L_i is the applied level of generalization for the sanitized data. The information loss increases as we apply higher level of generalization. In particular, for each higher level of generalization, the information lose factor is $1/H - 1$. Thus, the information loss is 0 for the leaf nodes and 1 for the root node.

B. Disclosure Risk Metrics

1) Uniqueness: Sanitization of dataset prior to publication is a common approach to protect individuals data. One of the most common attack is identity linkage attack, where a sanitized dataset is linked with public records. Uniqueness, in this context, is a metric for measuring re-identification risk of an individual.

Population uniqueness is a commonly utilized measure of re-identification risk [22] [23]. In the simplest definition, if there is an equivalence class of only one record then that record is considered unique. Unique records are more likely to be re-identified than non-unique records [24]. Uniqueness [10] can be measured in several ways. Mainly, it is the probability of record being unique in the original dataset [25].

Another approach would be measuring proportion of records in the public dataset that are unique. The following equation measures the proportion of records in a given dataset (i.e., voter registration list) that are unique [10]:

$$\lambda = \frac{\sum_i I(F_i = 1)}{N} \tag{4}$$

I is an indicator function. $I(F_i = 1)$ is 1 if the record is unique in the corresponding equivalence class, and 0 otherwise. Having said that, it is not possible to measure uniqueness precisely, therefore it should be estimated.

2) Estimating Uniqueness: Various models have been proposed in the literature to estimate the uniqueness of a sample dataset. Dankar et al. [10] conducted an experiment and applied Pitman [26], Zayatz [27] and SNB [28] uniqueness estimators on a clinical dataset and concluded that there was no single estimator that performed well across all conditions. Pitman estimator was the most accurate of all but only for low sampling fractions, while the SNB and Zayatz performed equally accurate for the higher sampling fraction. Hence, a decision rule has been adopted which chooses appropriate models based on sampling fraction. We use this to estimate the uniqueness in our experiments.

IV. Evaluation

We have used Adult dataset from UCI Machine Learning Repository (which is multivariate and it has been widely adapted in previous studies) in our evaluations. The data contains 45,222 records with 30,162 trained tuples and 15,060 test tuples.

A. Experiment Setup

We have considered income level, work hours per week, sex, education and work-class as quasi-identifiers, and occupation as sensitive attribute. We have utilized *ARX* [29] data anonymization tool in order to evaluate *k*-anonymity, *l*-diversity, *t*-closeness, and ϵ-differential privacy. We have executed and analyzed the impact of each data utility and disclosure risk metric per privacy model with various parameter values. For measuring re-identification risk, we have executed each experiment 10 times and reported the average re-identification risk in Figures 2 and 5.

ARX is a comprehensive data anonymization tool that provides both utility and re-identification risk analysis module. We have used default configurations for de-identification process except for privacy crietria and utility measure. In addition, we have specified the generalization hierarchy for each quasi-identifier attribute in Adult dataset. All the experiments were conducted on a 2.29GHz CPU with 8GB dedicated RAM.

B. Impact Analysis

We have examined the re-identification risk (uniqueness) for each privacy model. Figure 2 shows the impact of changing *k*, *l*, *t*, and ϵ for respective privacy models. Experiment results show that ϵ-differential privacy outperforms other models in terms of average re-identification risk ($< 2\%$).

Figure 3 demonstrates the impact of parameter values on utility loss measured by KL-divergence metric. *t*-closeness (*EMD= 0.45*) provides about 80% information (i.e., 20% information loss) while maintaining low re-identification risk ($< 10\%$). These experiments can be used by a data owner to set the appropriate parameter value for respective privacy models. For example in order to maintain low re-identification risk ($< 10\%$) using k-anonymity, the k should be set 4 or above. Respectively, in order to have better utility (at most 40% information loss) while using t-closeness privacy model, t value should be set 0.45 and above.

Observation 1: *Both ϵ-differential privacy and t-closeness provide high utility while maintaining low re-identification risk.*

Figure 4 represents the impact of parameter change on information loss measured by average equivalence class size. We are not reporting the detail of experimental results for utility metrics height as the general trend for these metrics are similar to KL-divergence for all privacy models.

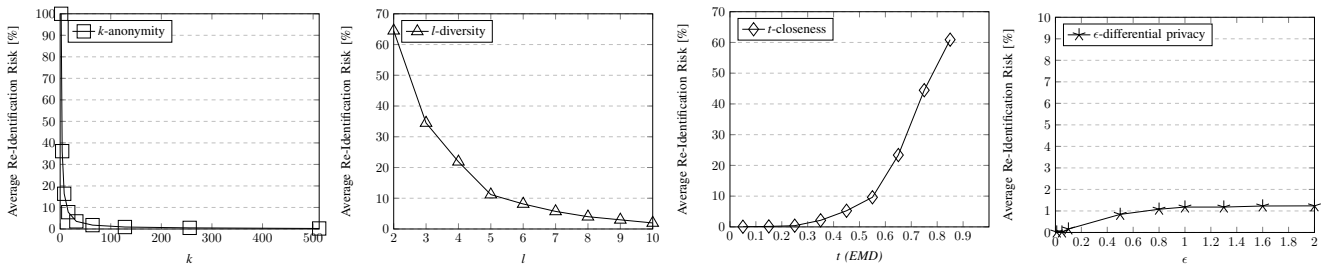

Fig. 2. Impact of Parameter Change on Average Re-Identification Risk (Uniqueness)

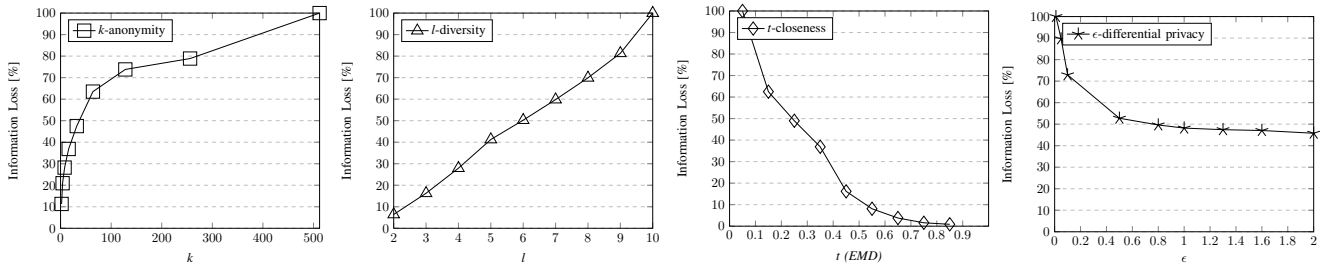

Fig. 3. Impact of Parameter Change on Information Loss (KL-divergence)

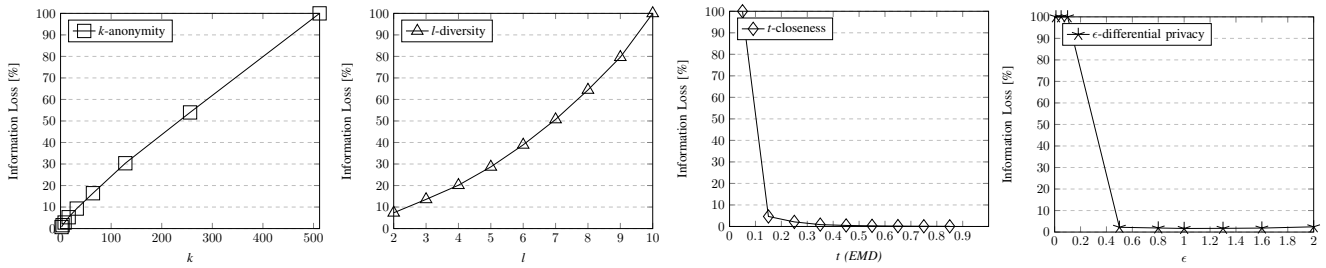

Fig. 4. Impact of Parameter Change on Information Loss (Avg. Equivalence Class Size)

Observation 2: *Impact of parameter change on information loss has the same trend for all privacy models considering various utility metrics (i.e. KL-divergence, height and average equivalence class size).*

C. Comparison of Privacy Models

One of the main objectives of this paper is to compare different models into a single framework (i.e., using same notion of privacy and utility metric). Figure 2 and Figure 3 show the performance of each model, for different parameter values, with respect to utility and privacy. Using the $< Parameter, Risk >$ and $< Parameter, InformationLoss >$ pairs from both chart of a privacy model, we obtain values for $< Risk, InformationLoss >$. We are then able to compare different privacy models on the single plot as shown in Figure 5: the x-axis represents privacy loss in terms of uniqueness (average re-identification risk) and y-axis represents utility in terms of information loss measured by KL-divergence utility metric. For instance the data points on l-diversity curve correspond to varying l from 1 to 10. The bottom-left corner of Figure 5 depicts the ideal sanitization cases which provide simultaneously high utility and privacy. This makes differential privacy the best candidate when making a choice between privacy models.

V. Conclusion

In this paper, we studied the risk-utility tradeoff among various privacy models. We conclude that ϵ-differential privacy outperforms other privacy models in terms of both utility and re-identification risk. However, the difference between t-closeness and ϵ-differential privacy is not significant. Our finding differs from previous study [8] (which concludes that there is no major differences among privacy models) due to the use of uniqueness to measure the re-identification risk. We argue that uniqueness should be used as it is the most widely used metric for measuring re-identification risk.

Moreover, the results of our experiment can be used by data owners to not only select the appropriate privacy model but also facilitates the choosing of a privacy parameter for a particular privacy model. In the future, we would like to release a risk analysis tool for sanitized dataset. It will enable to investigate the risk of released data as a digital forensic tool. It will also be of interest to measure the risk-utility tradeoff based on personalized privacy and utility metrics.

Acknowledgments

We sincerely thank the reviewers for their insightful comments. The research is supported in part by the NSERC

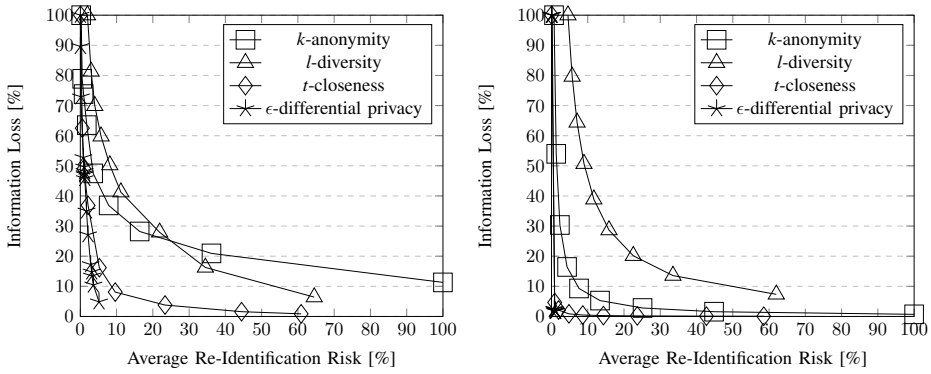

Fig. 5. Disclosure Risk vs. Data Utility (Left: KL-divergence, Right: Avg. Equivalence Class Size)

Discovery Grants (RGPIN-2015-04147), University of Manitoba Startup Grant, and Research Incentive Fund from Zayed University.

REFERENCES

[1] 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.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.

[4] C. Dwork, "Differential privacy," in *Encyclopedia of Cryptography and Security*. Springer, 2011, pp. 338–340.

[5] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 49–60.

[6] ——, "Mondrian multidimensional k-anonymity," in *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006, pp. 25–25.

[7] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to privacy-preserving data publishing: concepts and techniques*. CRC Press, 2010.

[8] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu, "Empirical privacy and empirical utility of anonymized data," in *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 77–82.

[9] P. A. Inc., "Privacy Analytics Risk Monitor," 2016, http://www.privacy-analytics.com/. [Online]. Available: http://www.privacy-analytics.com/

[10] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC medical informatics and decision making*, vol. 12, no. 1, p. 66, 2012.

[11] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 517–526.

[12] S. E. Fienberg, A. Rinaldo, and X. Yang, "Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables," in *Privacy in Statistical Databases*. Springer, 2010, pp. 187–199.

[13] R. H. Khokhar, R. Chen, B. C. Fung, and S. M. Lui, "Quantifying the costs and benefits of privacy-preserving health data publishing," *Journal of biomedical informatics*, vol. 50, pp. 107–121, 2014.

[14] G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes, "Disclosure risk vs. data utility: The ru confidentiality map," in *Chance*. Citeseer, 2001.

[15] Z. Teng and W. Du, "Comparisons of k-anonymization and randomization schemes under linking attacks," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 1091–1096.

[16] G. Loukides, A. Gkoulalas-Divanis, and J. Shao, "On balancing disclosure risk and data utility in transaction data sharing using ru confidentiality map," *Joint UNECE/Eurostat work session on statistical data confidentiality*, vol. 19, 2011.

[17] Q. Zhao, V. Hautamaki, and P. Fränti, "Knee point detection in bic for detecting the number of clusters," in *Advanced Concepts for Intelligent Vision Systems*. Springer, 2008, pp. 664–673.

[18] J. Domingo-Ferrer and D. Rebollo-Monedero, "Measuring risk and utility of anonymized data using information theory," in *Proceedings of the 2009 EDBT/ICDT Workshops*. ACM, 2009, pp. 126–130.

[19] A. Dasgupta, M. Chen, and R. Kosara, "Measuring privacy and utility in privacy-preserving visualization," in *Computer Graphics Forum*, vol. 32, no. 8. Wiley Online Library, 2013, pp. 35–47.

[20] S. Kullback, *Information theory and statistics*. Courier Corporation, 1968.

[21] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical report, SRI International, Tech. Rep., 1998.

[22] K. El Emam, A. Brown, and P. AbdelMalik, "Evaluating predictors of geographic area population size cut-offs to manage re-identification risk," *Journal of the American Medical Informatics Association*, vol. 16, no. 2, pp. 256–266, 2009.

[23] M. R. Koot, G. Noordende, C. Laat *et al.*, "A study on the re-identifiability of dutch citizens," 2010.

[24] J. G. Bethlehem, W. J. Keller, and J. Pannekoek, "Disclosure control of microdata," *Journal of the American Statistical Association*, vol. 85, no. 409, pp. 38–45, 1990.

[25] C. J. Skinner and M. Elliot, "A measure of disclosure risk for microdata," *Journal of the Royal Statistical Society: series B (statistical methodology)*, vol. 64, no. 4, pp. 855–867, 2002.

[26] J. Pitman, "Random discrete distributions invariant under size-biased permutation," *Advances in Applied Probability*, pp. 525–539, 1996.

[27] L. V. Zayatz, "Estimation of the percent of unique population elements on a microdata file using the sample," in *Statistical Research Division Report Number: Census/SRD/RR-91/08*. Citeseer, 1991.

[28] G. Chen and S. Keller-McNulty, "Estimation of identification disclosure risk in microdata," *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, vol. 14, pp. 79–95, 1998.

[29] F. Prasser, F. Kohlmayer, R. Lautenschläger, and K. A. Kuhn, "Arx-a comprehensive tool for anonymizing biomedical data," in *AMIA Annual Symposium Proceedings*, vol. 2014. American Medical Informatics Association, 2014, p. 984.