UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Disclosure metrics born from statistical evaluations of data utility.

Devyani Biswal (University of Ottawa)

dbisw078@uottawa.ca

Abstract

Statistical Disclosure Control (SDC) categorizes randomization techniques into two distinct groups: non-perturbative and perturbative methods. The statistical limitations of these methods have been examined and concluded to be bounded by the classic utility-risk trade-off. This roadblock creates the motivation for our work; improving data utility of SDC randomization techniques from a primarily statistical perspective.

Motivated by differential privacy, we study different noise distributions and the statistical properties of their outputs as applied to microdata. These insights lead to interesting properties of the data post-randomization, including asymptotic convergence of noise distributions. Experimental methods were used to compare various noise profiles to existing data perturbation methods such as PRAM (Post RAndomization Method) and noise addition in order to evaluate utility. Using statistical goodness of fit tests and risk measures, our findings resulted in a new randomization technique that improves data utility while ensuring a comparable level of disclosure risk.

Disclosure Metrics Born From Statistical Evaluations of Data Utility

Devyani Biswal[*], Luk Arbuckle[**], Rafal Kulik[***]

[*] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada, dbisw078@uottawa.ca

[**] Privacy Analytics, Ottawa, ON, Canada, larbuckle@privacy-analytics.com

[***] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada, rkulik@uottawa.ca

Abstract. Statistical Disclosure Control (SDC) categorizes randomization techniques into two distinct groups: non-perturbative and perturbative methods. The statistical limitations of these methods have been examined and concluded to be bounded by the classic utility-risk trade-off. This roadblock creates the motivation for our work; improving data utility of SDC randomization techniques from a primarily statistical perspective.

Motivated by differential privacy, we study different noise distributions and the statistical properties of their outputs as applied to micro data. These insights lead to interesting properties of the data post-randomization, including asymptotic convergence of noise distributions. Experimental methods were used to compare various noise profiles to existing data perturbation methods such as PRAM (Post RAndomization Method) and noise addition in order to evaluate utility. Using statistical goodness of fit tests and risk measures, our findings resulted in a new randomization technique that improves data utility while ensuring a comparable level of disclosure risk.

1 Introduction

A spectrum of identifiability has been recognized by industry [7], incorporating risk-based framing so that a scalable and proportionate approach to compliance to privacy regulations is provided. Technical models are incorporated into metrics that attempt to quantify measures of what may constitute a disclosure [10], and can capture a range of views that may or may not incorporate identifiability. One key challenge, then, is understanding what is meant by the term identifiable, and how well-established privacy models may support efforts to render data non-identifiable. To be effective at enhancing privacy, however, privacy models need to be used in practice, and that means they also need to be practical and, when used in the context of anonymization, produce useful data [1]. Barriers that discourage or limit the use of anonymization technology will simply drive organizations to use identifiable data, or simply not innovate at all.

Two of the most often cited models to avoid identity disclosure are k-anonymity [8], [9] and differential privacy [4]. The main idea of k-anonymity is to prevent re-identification of an anonymized dataset through record linkage attacks by grouping individuals into sets of at least k individuals with identifiable values on their indirect identifiers (where the size if k determines the level of privacy, since the upper bound of the probability of re-identification should be $1/k$). Differential privacy, on the other hand, provides a probabilistic guarantee that the inclusion of an individual in the dataset does not alter the outcome of a query to the dataset by more than a specified bound (determined largely by the parameter epsilon). These two classes of models, k-anonymity and differential privacy, are largely regarded as two separate, non-comparable models for disclosure control; however, some strides have been made to link key ideas between them [3].

Opinions vary regarding what should be the focus of anonymization techniques with regard to the privacy-utility trade-off [5]. In this paper we explore the basis for these techniques and how we may improve the quality of anonymized data from the perspective of producing statistically useful data. Our goal is to maintain some base level of privacy that we can concretely or objectively understand, while developing an approach that benefits the quality of the data. We begin by considering a combination of the Post-Randomization Method (PRAM) [6] and k-anonymity to form randomization within groups. Inspired by differential privacy, we then model noise addition to improve statistical properties of the output data. Establishing a relationship between k-anonymity and noise addition allows us to compare privacy levels of each technique and focus effort on maximizing data utility. Experimentation shows that considerable improvements can be made to the utility of a dataset by injecting basic noise which is selected to reflect a theoretical grouping size.

2 Theoretical Models

In recent years there have been efforts to connect privacy models with a rigorous probabilistic and statistical theory; see [11], [2]. In this paper we introduce some basic theory on some privacy models that will eventually serve as a foundation for more a more rigorous treatment. We begin with some notation.

Let $\underline{X} = (X_1, \ldots, X_n)$ be a dataset with n records. Let $\underline{Y} = (Y_1, \ldots, Y_n)$ be a transformed dataset. The transformation can be achieved by Post Randomization (PRAM), Grouping (k-anonymity) or noise addition, as in the case of differential privacy (DP).

2.1 PRAM

Post Randomization (PRAM) was formally introduced in [6]. The method applies to categorical data, that is, when the possible realizations of the random variables X_j lie in the set $\{a_i, i = 1, \ldots, M\}$, where a_i are real values. The basic idea is as follows: each of X_j's is transformed into Y_j according to the given transition probabilities:

$$p_{kl} = P(Y_j = a_l \mid X_j = a_k) \ .$$

The disclosure risk in PRAM is measured through *posterior odds*, that is, the relative probability that a rare score in the perturbed dataset $\underline{Y}$ corresponds with a rare score in the original dataset $\underline{X}$. These posterior odds should be small. Data utility is measured through the increase of variance of the estimates due to the measurement error introduced by PRAM. Theoretical formulas for the variances are given.

2.2 *k*-anonymity

With k-anonymity, the dataset is divided into m subgroups according to indirect identifiers. These subgroups are called *equivalence classes*, denoted by EC_i, $i = 1, \ldots, m$. Each individual belongs to one and only one equivalence class. An anonymized dataset $\underline{Y}$ provides k-anonymity, if for each individual Y_j in the given equivalence class, there exist at least $k - 1$ other individuals in the same class with identical values.

The bigger k is, the higher the level of privacy achieved. At the same time, in order to achieve large k one needs either a large population or apply a high level of generalization and suppression. The latter data transformations have a negative impact on data utility.

2.3 k-PRAM (a version of k-anonymity and PRAM)

Similarly to k-anonymity, the dataset $\underline{X}$ is divided into m subgroups in such the way that each subgroup (equivalence class) has at least k entries. If the original dataset is inhomogeneous, with large variability and outliers, this may not be possible to achieve. Rather, if the original data X_i follow a specific probability distribution, then the subgroups are selected in such the way that the expected number of entries in each of them is at least k. To be more specific, assume that X_i's are real-valued. Let $X_{(1)} \leq, \ldots, \leq X_{(n)}$ be the order statistics of $\underline{X}$ and define $\text{Range}(\underline{X}) = X_{(n)} - X_{(1)}$. Let $G_1, \ldots, G_m$ be m consecutive intervals (subgroups) of equal length

$$|G| = \frac{\text{Range}(\underline{X})}{m} \, .$$

That is, $G_1 = [X_{(1)}, X_{(1)} + |G|)$, $G_2 = [X_{(1)} + |G|, X_{(1)} + 2|G|)$ and so on. We require that the expected size of each subgroup is at least k:

$$\mathrm{E}\left[\sum_{j=1}^{n} \mathbb{1}\left\{X_j \in G_i\right\}\right] \geq k \, .$$

After the data are grouped into the intervals $G_1, \ldots, G_m$, we apply randomization using PRAM to each of the individual subgroups G_i separately in such the way that the size of each subgroup remains constant and hence the disclosure risk is at most $1/k$, as with k-anonymity. For simplicity we will call this k-PRAM to contrast with the method introduced in the next section.

Randomization can be applied in this way as a means of misleading would-be attackers or simply to maintain data formats. As we will see in Section 2.4 this is not desirable from the data utility point of view.

2.4 k-noise (a version of k-anonymity and noise addition)

Whereas PRAM and k-anonymity were conceived to limit disclosure risk from microdata, differential privacy was originally proposed to limit disclosure risk from statistical queries. As previously mentioned, differential privacy provides a probabilistic guarantee that the inclusion of an individual in the dataset does not alter the outcome of a query on the dataset by more than a specific bound. It became synonymous with adding noise following a Laplace distribution.

Inspired by this idea, we combine k-anonymity with noise addition, using any suitable probability distribution and without explicitly attempting to meet the definition of differential privacy at this time. If the data are grouped, as with k-anonymity, and an arbitrary noise is added to individual data points, there is no guarantee that the group sizes in the transformed dataset are preserved. However, with carefully prescribed noise addition, the group sizes in the transformed dataset can be controlled. As such, the disclosure risk can be similarly controlled as is the case of k-anonymity, which we will call k-noise.

Once the privacy level is fixed, we can focus efforts on improving data utility. As opposed to the randomization within fixed intervals or groups, as described in Section 2.3, our approach does not introduce bias and hence it has a better data utility.

As in Section 2.3, we divide the dataset into m groups G_i of length $|G| = 2\delta$ with some $\delta > 0$. This implies that in a 2δ-neighbourhood of any record $x \in \underline{X}$, we have at least k individuals X_i:

$$\#\{j : |X_j - x| < 2\delta\} \geq k \ .$$

We note however that we cannot control the number of individuals in a δ-neighbourhood, $\#\{j : |X_j - x| < \delta\}$, as shown in Figure 1.

Let $\underline{Y} = \underline{X}^{(r)} = (X_1^{(r)}, \ldots, X_n^{(r)})$ be a randomized dataset defined by

$$\underline{X}^{(r)} = \underline{X} + \underline{\eta},$$

where $\underline{\eta} = (\eta_1, \ldots, \eta_n)$ is a vector of independent identically distributed random variables.

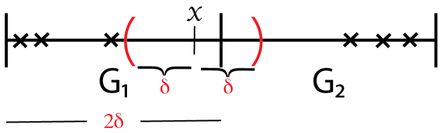


Figure 1: Graphical representation of the δ neighbourhood of an x entry of the dataset.

2.5 Uniformly distributed noise

We will assume in this section that η_j, $j = 1, \ldots, n$, have an uniform distribution with a parameter $a > 0$. If for a particular group G_i, the data X_j are concentrated around its centre, then the choice $a = \delta$ guarantees, with a high probability, that

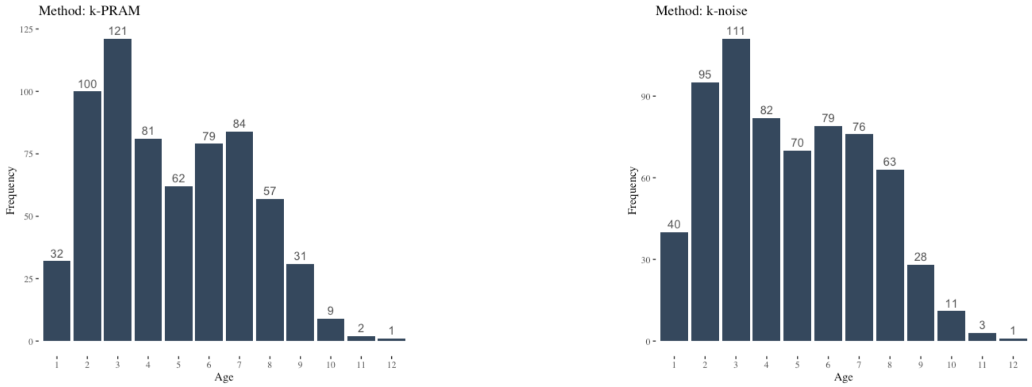
$$\#\{j : |X_j^{(r)} - x| < 2\delta\} \geq k \ .$$

However, if this is not the case the bound cannot be guaranteed. Thus, the theoretical bound must consider the worst-case scenario and be more conservative. The most conservative bound guarantees that by applying a uniform noise with parameter δ there exists at least $\frac{1}{2}k$ other individuals within a 2δ neighbourhood. Furthermore, we show that the underlying distribution of the dataset is not needed in order to guarantee this bound. The theorem and proof are provided in Section 5.

3 Experimental Results

In the first experiment we illustrate that the although the bound obtained in Theorem 1 can be conservative in reality it close to the target value of k. This is shown on Figures 2a and 2b. We show experimental results using a public dataset consisting of 659 records with several categorical and numerical variables. We focus on one numerical variable of interest, Age, and aim to study the effects of data utility when comparing two methods of anonymization. in figure 2, the left-hand image shows the histogram of the original ages and the right-hand image shows the histogram for the noisy data where η_j has the uniform distribution, $\mathrm{Unif}[-\delta, \delta]$. Using the same binning between histograms, we can see that the empirical distributions for both the original and the noisy datasets are nearly identical and hence the data utility (measured by an arbitrary metric) is comparable.

The difference between the k-PRAM and k-noise methods are illustrated on Figure 3 and the proposed noise injection method in Figure 3. With k-noise, the resulting distribution of ages is smoother, which would suggest better utility and has the added benefit of further misleading would-be attackers. The light blue clusters show the inherent bias in the dataset when using the first method k-PRAM, versus the smooth dark blue trend formed when using k-noise. Furthermore, we divide the Age variable into 12 groups, each spanning an interval of 5 years on the interval $[24, 79]$, and we can see from Table 1 that k-noise reduces the bias and error compared to k-PRAM.



(a) k-PRAM

(b) k-Noise

Figure 2: Empirical distributions of randomized dataset.

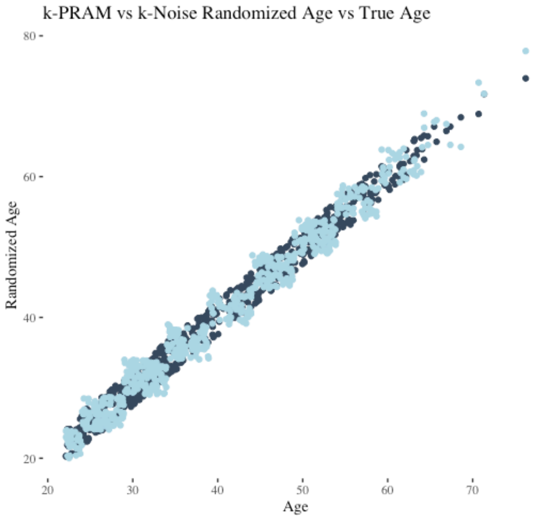
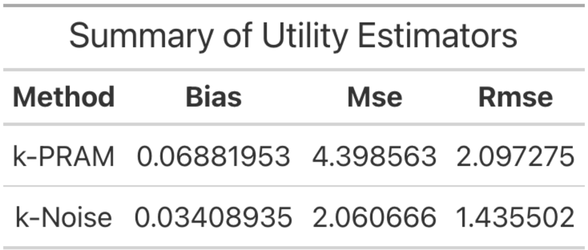


Figure 3: Scatterplot of anonymized ages using two methods



Summary of Utility Estimators

Method	Bias	Mse	Rmse
k-PRAM	0.06881953	4.398563	2.097275
k-Noise	0.03408935	2.060666	1.435502

Figure 4: Different utility measures to compare k-PRAM and k-noise.

To test this further, we employ the use of Monte Carlo simulations to get the expected number of ages, representing individuals, in a neighbourhood of an anonymized entry when using k-noise. k-noise can be thought of as a local measure of k-anonymity, since the group is being compared to the neighbourhood of adjacent points. If this number of ages exceeds or equals the group size of the original entry, then we can determine they are adequately protected within a group. We are treating the underlying dataset as the baseline for comparing to k-noise. Our results far exceed our theoretical bound of $\frac{1}{2}k$ and demonstrates the effectiveness of this approach in practice.

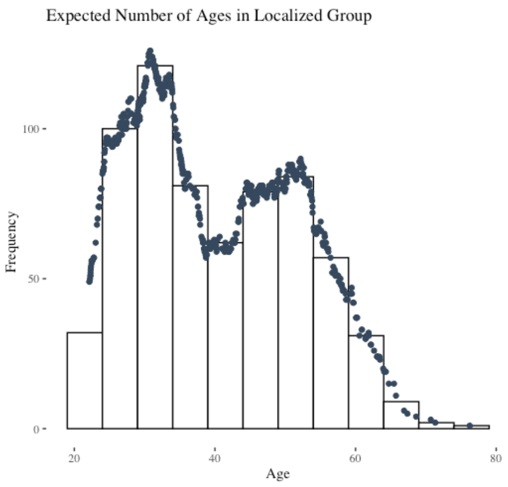


Figure 5: Expected number or records within a neighbourhood of [-2.5,2.5] years of each randomized record in X

4 Conclusion

By adjusting the noise level to achieve an expected minimum threshold k, we are able to improve the distribution of an anonymized variable over the more common approach of randomizing within fixed intervals to satisfy k-anonymity. This approach of noise addition allows us to leverage the well-established concept of k-anonymity, which is easily understood and has well-established precedents for the threshold k. We believe this will allow us to fine-tune noise levels based on other statistical properties and make inroads towards new approaches for privacy models.

5 Appendix: Theorems and Proofs

Theorem 1. *Let $\underline{X} = (X_1, \ldots, X_n)$ be a dataset and $\underline{X}^{(r)} = (X_1^{(r)}, \ldots, X_n^{(r)})$ be a randomized dataset defined by*

$$\underline{X}^{(r)} = \underline{X} + \underline{\eta},$$

where $\underline{\eta} = (\eta_1, \ldots, \eta_n)$ is a vector of independent uniform random variables on $[-a, a]$ for $a > 0$. Let $\delta > 0$ and assume that for each $x \in [X_{(1)}, X_{(n)}]$ we *have*

$$\#\{j : |X_j - x| < 2\delta\} \geq k \ .$$

Assume $a = \delta$. Then

$$\mathbb{E}\left[\#\{j : |X_j^{(r)} - x| \leq 2\delta\} \mid \underline{X}\right] > \frac{1}{2}\#\{j : |X_j - x| < 2\delta\} = \frac{1}{2}k \ .$$

Remark 1. We note that the expectation is calculated conditionally on the database $\underline{X}$, hence the database entries are treated as deterministic and the randomness is due to the noise $\underline{\eta}$. Using the tower property of the conditional distribution we also obtain

$$\mathbb{E}\left[\#\{j : |X_j^{(r)} - x| \leq 2\delta\}\right] > \frac{1}{2}\#\{j : |X_j - x| < 2\delta\} = \frac{1}{2}k \ .$$

Proof of Theorem 1. Let $A_j = -2\delta - X_j + x, B_j = 2\delta - X_j + x$. Then, using the properties of the uniform distribution,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}\left\{-2\delta < X_j^{(r)} < 2\delta\right\} \mid \underline{X}\right] \\
&= \sum_{j=1}^{n} \mathbb{E}\left[\mathbb{1}\left\{-2\delta - X_j + x < \eta_j < 2\delta - X_j + x\right\} \mid \underline{X}\right] \\
&= \frac{2\delta}{a}\sum_{j=1}^{n} \mathbb{1}\left\{-a < A_j, B_j < a\right\} + \sum_{j=1}^{n} \mathbb{1}\left\{A_j < -a, a < B_j\right\} \\
&\quad + \frac{1}{2a}\sum_{j=1}^{n}(a - A_j)\mathbb{1}\left\{-a < A_j, a < B_j\right\} + \frac{1}{2a}\sum_{j=1}^{n}(B_j + a)\mathbb{1}\left\{A_j < -a, B_j < a\right\}.
\end{aligned}$$

For $a = \delta$ the expressions above become

$$\begin{aligned}
&\sum_{j=1}^{n} \mathbb{1}\left\{x - \delta < X_j < x + \delta\right\} \\
&+ \sum_{j=1}^{n} \frac{(3\delta - x + X_j)}{2\delta}\mathbb{1}\left\{x - 3\delta < X_j < x - \delta\right\} \\
&+ \sum_{j=1}^{n} \frac{(3\delta + x - X_j)}{2\delta}\mathbb{1}\left\{x + \delta < X_j < x + 3\delta\right\}.
\end{aligned}$$

We split the last two terms as $J_1 + J_2 + J_3 + J_4$ with

$$\begin{aligned}
J_1 &:= \sum_{j=1}^{n} \frac{(3\delta - x + X_j)}{2\delta} \mathbb{1}\left\{x - 2\delta < X_j < x - \delta\right\} \\
J_2 &:= \sum_{j=1}^{n} \frac{(3\delta - x + X_j)}{2\delta} \mathbb{1}\left\{x - 3\delta < X_j < x - 2\delta\right\} \\
J_3 &:= \sum_{j=1}^{n} \frac{(3\delta + x - X_j)}{2\delta} \mathbb{1}\left\{x + \delta < X_j < x + 2\delta\right\} \\
J_4 &= \sum_{j=1}^{n} \frac{(3\delta + x - X_j)}{2\delta} \mathbb{1}\left\{x + 2\delta < X_j < x + 3\delta\right\} =: I_1 + I_2 + I_3.
\end{aligned}$$

Note that

$$J_1 + J_3 \geq \frac{1}{2} \sum_{j=1}^{n} \mathbb{1}\left\{x - 2\delta < X_j < x - \delta\right\} + \frac{1}{2} \sum_{j=1}^{n} \mathbb{1}\left\{x + \delta < X_j < x + 2\delta\right\}$$

Ignoring J_2 and J_4, the expectation is bounded below by

$$\frac{1}{2} I_1 + J_1 + J_3 \geq \frac{1}{2} \#\{j : |X_j - x| < 2\delta\} \geq \frac{1}{2} k \ .$$

□

References

[1] Luk Arbuckle and Khaled El Emam. *Building an Anonymization Pipeline: Creating Safe Data*. O'Reilly Media, 2020.

[2] T. Tony Cai, Yichen Wang, and Linjun Zhang. The Cost of Privacy: Optimal Rates of Convergence for Parameter Estimation with Differential Privacy. *arXiv:1902.04495 [cs, stat]*, 2020.

[3] Josep Domingo-Ferrer and Jordi Soria-Comas. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, 74:151–158, 2015.

[4] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

[5] Mark Elliot and Josep Domingo-Ferrer. The future of statistical disclosure control. *arXiv preprint arXiv:1812.09204*, 2018.

[6] J M Gouweleeuw, Peter Kooiman, and PP De Wolf. Post randomisation for statistical disclosure control: Theory and implementation. *Journal of official Statistics*, 14(4):463, 1998.

[7] Jules Polonetsky, Omer Tene, and Kelsey Finch. Shades of gray: Seeing the full spectrum of practical data de-intentification. *Santa Clara L. Rev.*, 56:593, 2016.

[8] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. 1998.

[9] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[10] Isabel Wagner and David Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

[11] Larry Wasserman and Shuheng Zhou. A Statistical Framework for Differential Privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.