Obtaining (ϵ, δ)-differential privacy guarantees when using the Poisson distribution to synthesize tabular data

James Jackson[0000−0002−4832−6638], Robin Mitra[2][0000−0001−9584−8044], Brian Francis[1][0000−0001−7926−9085], and Iain Dove[3][0000−0002−1145−2999]

[1] Lancaster University, Lancaster, UK
[2] Department of Statistical Science, UCL, London, UK
[3] Office for National Statistics, Titchfield, UK

Abstract. We show that differential privacy type guarantees can be obtained when using a Poisson synthesis mechanism to protect counts in contingency tables. Specifically, we show how to obtain (ϵ, δ)-probabilistic differential privacy guarantees via the Poisson distribution's cumulative distribution function). We demonstrate this Poisson synthesis mechanism empirically with the synthesis of the ESC_{rep} data set, an administrative-type database that resembles the English School Census.

Keywords: differential privacy · synthetic data · tabular data

1 Introduction

Differential privacy (DP) [8] is a property of a perturbation mechanism that formally quantifies how accurately any individual's true values can be established, given all other individuals' true values are known. Originally developed as a way to protect the privacy of summary statistics (queries), it soon expanded as a way to protect entire data sets. Differentially private data synthesis (DIPS) has since become a popular area of research; see, for example, [1], [14], [5], [15], [4], [16], [7].

In [12], [11], we proposed a synthesis approach for categorical data sets, which takes place at the tabular level, and which uses saturated count models. The use of saturated models means we effectively use count distributions, such as the Poisson, to apply noise to the counts in the original data's contingency table. This approach therefore shares traits with DP mechanisms; the Laplace mechanism, for example, applies Laplace noise to original counts.

In this paper, we consider the ability to obtain DP-guarantees when using the Poisson distribution to synthesize counts in tabular data (contingency tables). We show that although ϵ-DP cannot be satisfied, (ϵ, δ)-DP guarantees can be obtained through the use of the Poisson's cumulative distribution function (CDF). With the exception of [16], the use of count distributions has largely been overlooked as a way to satisfy DP. Similarly, the use of CDFs to satisfy (ϵ, δ)-probabilistic DP has been underexplored. In the future, this work can then

be extended from the Poisson to more complex count distributions (such as the Poisson inverse-Gaussian), where additional parameters provide scope for fine-tuning.

The paper is structured as follows. Section 2 introduces some terminology and definitions. Section 3 looks at existing DP mechanisms for contingency tables, such as the (discretised) Laplace and Gaussian mechanisms. Section 4 gives our novel contribution, the ability to obtain (ϵ, δ)-DP guarantees when using a Poisson synthesis mechanism. Section 5 gives an empirical example using an administrative database. Section 6 gives some concluding remarks.

2 Terminology and definitions

Rinott et al. [17] set out how DP extends into a contingency table setting. Following their notation, let $\mathbf{a} = (a_k, \ldots, a_K) \in \mathcal{A}$ and $\mathbf{b} = (b_k, \ldots, b_K) \in \mathcal{B}$ denote vectors of cell counts in the original and synthetic data's contingency tables, respectively, where K denotes the number of cells and $\mathcal{A}$ and $\mathcal{B}$ denote the range of obtainable original and synthetic counts (respectively). For contingency tables, we suppose that $\mathcal{A} = \mathcal{B} = \mathbb{Z}_{\geq 0}^K$, where $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers.

Moreover, we describe $\mathbf{a}$ and $\mathbf{a}'$ as neighbours, denoted by $\mathbf{a} \sim \mathbf{a}'$, whenever all but one of the counts in $\mathbf{a}$ and $\mathbf{a}'$ are identical and the differing count differs by exactly one. Henceforth, without loss of generality, we suppose $\mathbf{a}$ and $\mathbf{a}'$ differ in their kth element only, i.e. $a_k = a'_k - 1$ and $a_i = a'_i$ for $i = 1, \ldots, K$, $i \neq k$. Thus $\mathbf{a}$ represents the data held by the intruder (who knows all but one of the individuals' true values) and $\mathbf{a}'$ represents the completed data where the "unknown individual" has been added to the cell in which they truly belong.

The ϵ-DP definition revolves around the likelihood ratio, or, more accurately, around a series of likelihood ratios.

Definition 1 (ϵ-DP). *A perturbation mechanism $\mathcal{M}$ satisfies ϵ-DP ($\epsilon > 0$) if:*

$$exp(-\epsilon) \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = \mathbf{b})} \leq exp(\epsilon), \quad (1)$$
$$\forall\, \mathbf{a} \sim \mathbf{a}' \in \mathcal{A} \text{ and } \forall\, \mathbf{b} \in \mathcal{B}.$$

Definition 1 is the special case of the standard DP definition, given in [8], for when the range of $\mathcal{A}$ and $\mathcal{B}$ are discrete. For any $\mathbf{a}$, $\mathbf{a}'$ and $\mathbf{b}$, whenever the ratio $\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)/\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)$ is either small or large, relatively too much is gleaned about the unknown individual's true values. It is worth noting, too, that the above definition considers all possible synthetic data sets in $\mathcal{B}$, illustrating that DP is not a risk metric for a particular synthetic data set but rather a property of a synthesis mechanism.

Somewhat confusingly, there are two similar but different relaxations of ϵ-DP. The first is (ϵ, δ)-differential privacy [9]. The second is known as (ϵ, δ)-probabilistic differential privacy [14]. These are given below in Definitions 2 and 3. In the remainder of this paper, we focus on (ϵ, δ)-probabilistic DP. Yet whenever (ϵ, δ)-probabilistic DP is satisfied, (ϵ, δ)-DP is also satisfied [10].

Definition 2 ((ϵ, δ)-DP). *A perturbation mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP ($\epsilon > 0$; $0 \leq \delta \leq 1$) if:*

$$\frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq exp(\epsilon) \quad and \quad \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b) - \delta}{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)} \leq exp(\epsilon) \qquad (2)$$
$$\forall\ \mathbf{a} \sim \mathbf{a}' \in \mathcal{A},\ \mathbf{b} \in \mathcal{B}.$$

Definition 3 ((ϵ, δ)-probabilistic DP). *A perturbation mechanism $\mathcal{M}$ satisfies (ϵ, δ)-probabilistic DP ($\epsilon > 0$; $0 \leq \delta \leq 1$) if:*

$$p\left[\frac{1}{exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq exp(\epsilon)\right] > 1 - \delta \quad \forall\ \mathbf{a} \sim \mathbf{a}' \in \mathcal{A},\ \mathbf{b} \in \mathcal{B}. \qquad (3)$$

Theorem 1 ((ϵ, δ)-probabilistic DP implies (ϵ, δ)-DP). *If a perturbation mechanism $\mathcal{M}$ satisfies (ϵ, δ)-probabilistic DP, then it also satisfies (ϵ, δ)-DP. (Proof: see [10])*

3 Examples of existing DP mechanisms

We now give examples of existing DP mechanisms suitable for synthesizing counts in contingency tables.

Example 1 (The (discretised) Laplace mechanism). A random variable $X \sim \text{Laplace}(\mu, d)$ has probability density function f_L:

$$f_L(x; \mu, d) = \frac{1}{2d} \exp\left(-\frac{|x - \mu|}{d}\right)$$

The Laplace mechanism $\mathcal{M}_L$ uses the Laplace distribution to add random noise to the original counts $\mathbf{a}$. Specifically, $\mathcal{M}_L(\mathbf{a}) = \mathbf{a} + \mathbf{c}$ where $\mathbf{c}$ is a K-dimensional vector of discretised Laplace$(0, 1/\epsilon)$ random variates. To show that this mechanism does indeed satsify DP, consider an arbitrary $\mathbf{b} \in \mathcal{B}$ (i.e. $\mathbf{b} = \mathcal{M}_L(\mathbf{a})$); then

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_L(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_L(\mathbf{a}') = \mathbf{b})} &= \frac{\mathbb{P}(\mathbf{c} = \mathbf{b} - \mathbf{a})}{\mathbb{P}(\mathbf{c} = \mathbf{b} - \mathbf{a}')} \\ &= \frac{f_L(\mathbf{b} - \mathbf{a})}{f_L(\mathbf{b} - \mathbf{a}')} \\ &\leq \exp(\epsilon|\mathbf{a} - \mathbf{a}'|) = \exp(\epsilon),\end{aligned}$$

as, by definition, $|\mathbf{a} - \mathbf{a}'| = 1$. By a similar argument it can be shown that the LHS is also greater than or equal to $\exp(1/\epsilon)$, thus satisfying the ϵ-DP definition given in (1).

Example 2 (The Gaussian mechanism). A random variable $X \sim \text{Normal}(\mu, \sigma^2)$ has probability density function f_G:

$$f_G(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

The application of discretised Normal(0, σ^2) random noise results in a synthesis mechanism, say $\mathcal{M}_G$, that achieves (ϵ,δ)-differential privacy. Recall that $a_i = a'_i$ for $i = 1,\ldots,k-1,k+1,\ldots,K$ and that $a_k + 1 = a'_k - 1$. Then it follows that

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_G(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}_G(\mathbf{a}') = b)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k - a_k}{\sigma}\right)^2\right]}{\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{b_k - a_k + 1}{\sigma}\right)^2\right]} \\ &= \exp\left[-\frac{1}{2\sigma^2}(2a_k - 2b_k - 1)\right].\end{aligned}$$

Recall that (ϵ,δ)-probabilistic DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}(\mathbf{a}) = b)}{\mathbb{P}(\mathcal{M}(\mathbf{a}') = b)} \leq \exp(\epsilon) \quad \text{with probability } 1-\delta,$$

which, in this instance, occurs whenever

$$-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon \quad \text{with probability } 1-\delta.$$

The probability $1-\delta$ can be obtained from Φ the normal distribution's CDF [2], as $b_k \sim \text{Normal}(a_k, \sigma^2)$.

$$\begin{aligned}1-\delta &= \mathbb{P}(-\epsilon \leq -\frac{1}{2\sigma^2}(2a_k - 2b_k - 1) \leq \epsilon) \\ &= \mathbb{P}(a_k - \sigma^2\epsilon - 1/2 \leq b_k \leq a_k + \sigma^2\epsilon - 1/2) \\ &= \Phi\left(\frac{a_k + \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) - \Phi\left(\frac{a_k - \sigma^2\epsilon - 1/2 - a_k}{\sigma}\right) \\ &= \Phi\left(\sigma\epsilon - 1/2\right) - \Phi\left(-\sigma\epsilon - 1/2\right)\end{aligned}$$

Example 3 (Multinomial-Dirichlet synthesizer). A multinomial-Dirichlet synthesis mechanism [1], say $\mathcal{M}_{MD}$, can also yield DP guarantees. The original counts $\mathbf{a}$ can be converted to cell probabilities $\boldsymbol{\pi}$ simply by dividing by n (the number of individuals in the data). A Dirichlet prior with concentration parameters $\boldsymbol{\alpha} = (\alpha_k, \alpha_2, \ldots, \alpha_K)$ is placed on $\boldsymbol{\pi}$ (see [1] for more on this approach). Using the same "without loss of generality" assumptions as previous, it follows that

$$\begin{aligned}\frac{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}') = \mathbf{b})} &= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a'_k + \alpha_k)}{\Gamma(b_k + a'_k + \alpha_k)}. \\ &= \frac{\Gamma(b_k + a_k + \alpha_k)}{\Gamma(a_k + \alpha_k)} \cdot \frac{\Gamma(a_k - 1 + \alpha_k)}{\Gamma(b_k + a_k - 1 + \alpha_k)} \\ &= \frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k}. \end{aligned} \tag{4}$$

Recall again that DP is satisfied whenever

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_{MD}(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon).$$

As the expression in (4) is always greater than or equal to one, and hence always greater than $1/\exp(\epsilon)$, DP is satisfied whenever

$$\frac{b_k + a_k - 1 + \alpha_k}{a_k - 1 + \alpha_k} \leq \exp(\epsilon),$$

and, as $a_k \geq 1$ and $b_k \leq n$, whenever

$$\frac{n + \alpha_k}{\alpha_k} \leq \exp(\epsilon) \quad \Rightarrow \quad \alpha_k \geq \frac{n}{\exp(\epsilon) - 1}$$

Considering all counts $a_1, \ldots, a_K$ gives that DP is satisfied when:

$$\max_i \alpha_i \geq \frac{n}{\exp(\epsilon) - 1}, \quad \text{a result from [14].}$$

4 Satisfying (ϵ, δ)-probabilistic DP with a Poisson synthesis mechanism

When using saturated count models to synthesize categorical data expressed as contingency tables, as set out in [12], we are effectively using a count distribution, e.g. the Poisson, to apply noise to original counts. We assume that a constant pseudocount $\alpha > 0$ is added to every element of $\mathbf{a}$ (i.e. to *all* original counts, not just to zero counts as in [12]), which opens up the possibility that original counts of zero can be synthesized to non-zeros. When using the Poisson we apply the following mechanism, which we denote by $\mathcal{M}_P$, to obtain a set of synthetic counts:

$$b_i \mid a_i, \alpha \sim \text{Poisson}(a_i + \alpha), \quad i = 1, \ldots, K,$$

$$\text{i.e.} \quad \mathbb{P}(\mathcal{M}_P(a_i) = b_i) = \frac{\exp(-a_i - \alpha)(-a_i - \alpha)^{b_i}}{b_i!}, \quad i = 1, \ldots, K.$$

Supposing once again that $\mathbf{a}$ and $\mathbf{a}'$ differ in their kth element only, we have:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} = \exp(-1)\left(\frac{a_k + \alpha}{a_k - 1 + \alpha}\right)^{b_k}. \tag{5}$$

This quantity is bounded below by exp(-1), with this minimum occurring when $b_k = 0$. It is unbounded above, however, as b_k can be any integer up to infinity; i.e. the expression in (5) tends to infinity as b_k tends to infinity. Thus ϵ-DP cannot be satisfied.

Instead, we now consider the (ϵ, δ)-probabilistic DP relaxation, first considering the left-hand inequality of the DP definition (Def. 1):

$$\frac{1}{\exp(\epsilon)} \leq \frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \quad \Rightarrow \quad b_k \geq \frac{1-\epsilon}{\log\left(\frac{a_k+\alpha}{a_k-1+\alpha}\right)}.$$

When $\epsilon \geq 1$, this inequality holds with probability 1. When $0 < \epsilon < 1$, the probability that this inequality holds can be determined through the Poisson's CDF, since b_k is a realization from a Poisson random variable. This probability is given as:

$$1 - F^P_{a_k+\alpha}\left[\frac{1-\epsilon}{\log\left(\frac{a_k+\alpha}{a_k-1+\alpha}\right)}\right], \tag{6}$$

where $F^P_{a_k+\alpha}$ is the CDF of the Poisson distribution with mean $a_k + \alpha$.

We next consider the right-hand inequality of Def. 1:

$$\frac{\mathbb{P}(\mathcal{M}_P(\mathbf{a}) = \mathbf{b})}{\mathbb{P}(\mathcal{M}_P(\mathbf{a}') = \mathbf{b})} \leq \exp(\epsilon) \quad \Rightarrow \quad b_k \leq \frac{1+\epsilon}{\log\left(\frac{a_k+\alpha}{a_k-1+\alpha}\right)}.$$

For all ϵ, this inequality holds with probability

$$F^P_{a_k+\alpha}\left[\frac{1+\epsilon}{\log\left(\frac{a_k+\alpha}{a_k-1+\alpha}\right)}\right]. \tag{7}$$

Recall that in (ϵ, δ)-probabilistic DP, $1-\delta$ is the probability that DP is satisfied, i.e. the probability that both inequalities hold. A non-trivial question when $0 < \epsilon < 1$ is how to combine the probabilities given in (6) and (7) and hence compute δ? This is an area of future research.

When $\epsilon > 1$, however, the left-hand inequality of Def. 1 always holds, thus we need only focus on (7). Although non-trivial – note, a formal proof has been omitted here but extensive empirical simulation results have been undertaken – for any $\epsilon \geq 1$ and $\alpha > 0$, (7) is minimised when $a_k = 1$ (when $a'_k = 0$). Thus:

$$1 - \delta = F^P_{1+\alpha}\left[\frac{1+\epsilon}{\log\left(\frac{1+\alpha}{\alpha}\right)}\right]. \tag{8}$$

This also demonstrates the role of α as a tuning parameter for risk. In general, a larger α value corresponds to a lower δ value. Yet δ is not a decreasing function of α. For a very brief explanation, this is because increasing α increases the value of the expression inside the squared bracket in (8), but it also increases the mean of the Poisson random variable from which a synthetic count is drawn. Figure 1 illustrates the nature of the relationship between α and δ for different values of ϵ. For example, setting $\alpha = 0.1$ satisfies approximately (3,0.3)-probabilistic DP and (1.5,0.6)-probabilistic DP.

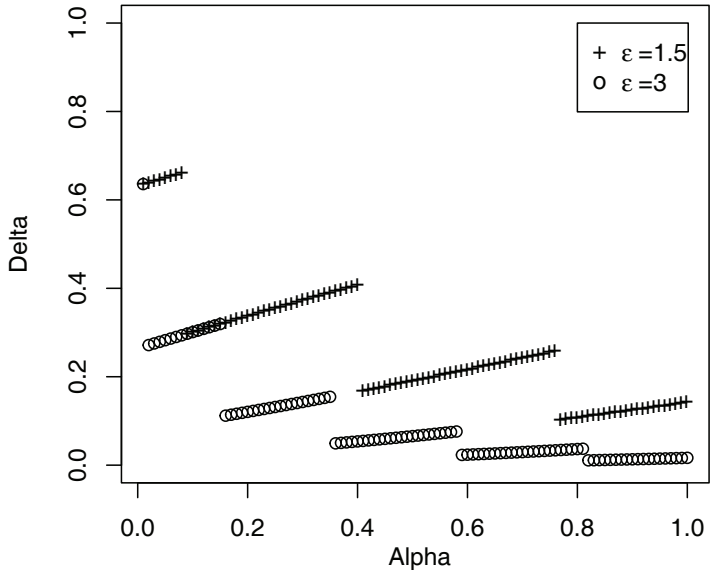

Fig. 1. The relationship between α and δ in the Poisson synthesis mechanism for $\epsilon = 1.5$ and $\epsilon = 3$.

In contingency tables where there are no zero counts, a (ϵ, δ)-DP guarantee can be obtained when $\alpha = 0$. In this instance, δ is determined by the smallest original count, i.e.:

$$1 - \delta = F^P_{a_i+\alpha}\left[\frac{1+\epsilon}{\log\left(\frac{\min_i a_i + 1}{\min_i a_i}\right)}\right]. \tag{9}$$

In a sense, in this example we have violated the traditional (ϵ, δ)-probabilistic DP definition given in (3) because δ is dependent on a particular set of original counts $\mathbf{a}$ – not all original counts.

We can easily replace the Poisson with any other count distribution (e.g. the negative binomial, Poisson inverse-Gaussian, Delaporte, Sichel, etc.), which would lead to a different result for (5).

5 An empirical example

5.1 The English School Census administrative database

The English School Census (ESC) is a large administrative database belonging to the UK's Department for Education (DfE), which holds information about pupils attending state-funded schools in the UK. Owing to the presence of sensitive data, strict privacy guarantees would be required for data from the ESC to

be made available to researchers. There is therefore great appeal to DP-type approaches, where more formal guarantees of privacy can be obtained.

Access to the real ESC data is currently restricted, even for the sake of demonstrating the effectiveness of privacy methods. For this reason, staff at the Office for National Statistics (ONS) created a substitute data set using publicly-available data sources, such as previously published ESC data and 2011 census tables. A key feature of this data set, which we name ESC_{rep}, is that it replicates some of the statistical properties present in the ESC. The version we use here has approximately 8×10^6 individuals (rows) and 5 categorical variables (columns). As all variables are categorical, the data set can be expressed as a contingency table. More information about the data set – as well as the data set itself – is available at [3].

5.2 Applying the Poisson synthesis mechanism

We now apply the Poisson synthesis mechanism to the ESC_{rep} data, considering different values of α, and considering $\epsilon > 1$ values.

The ESC_{rep} data has a high proportion of zero counts (roughly 90%), hence the expression in (8) for δ applies. Figure 2 gives combinations of (ϵ, δ) values that can be achieved for the ESC_{rep} data when using α values of 0.1, 0.2, 0.5 and 1. For example, when $\epsilon = 2$, an α value of 1 is required to obtain a δ value of 0.05; when $\alpha = 0.1$, a δ value of 0.05, is obtained only for ϵ values greater than 6.

DP methods, in general, are known to have a detrimental effect on utility. To assess the effect of δ on specific utility [18], we compare analysis results obtained from the original and synthetic data. We fit a log-linear model involving the ethnicity and age variables and compute the parameters' confidence intervals from both the original and synthetic data. We then use the confidence interval overlap metric [13] to compare the original and synthetic data confidence intervals.

The boxplots in Figure 3 show that even small values of α have an adverse effect on overlap, i.e. utility. It suggests that in order to obtain meaningful results from the synthetic data, the synthesizer would need to keep α as low as possible. This, of course, is intrinsically linked to the risk-utility trade-off, as to achieve this, the synthesizer would have to choose a higher value of ϵ or δ (or both). An advantage of DP synthesis approaches is that they provide a parameter characterising risk. Although risk and utility are of course intertwined, there are generally fewer risk metrics available to the synthesizer than there are utility metrics. Choosing suitable values for ϵ and δ values is far from trivial (see [6]) and in general the synthesizer would consider a range of such values.

6 Discussion

To summarise, this paper is an investigation into obtaining DP-type guarantees when using a Poisson synthesis mechanism to protect the privacy of counts in contingency tables. Going forward, we believe other count distributions, such

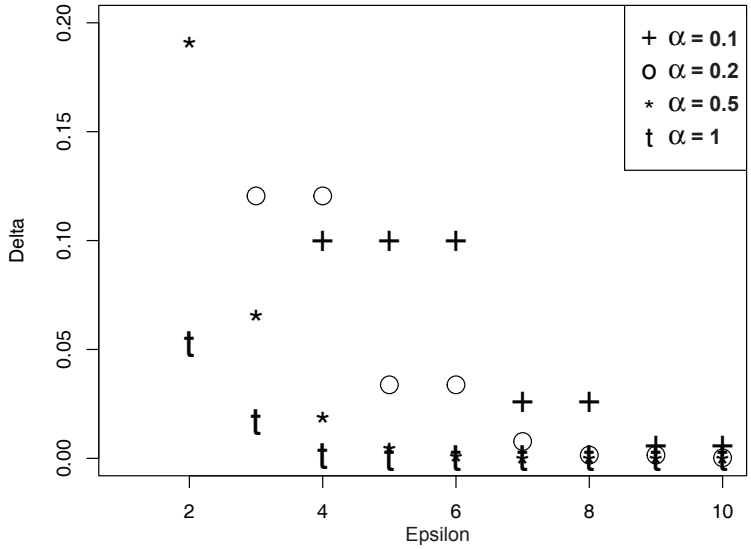

Fig. 2. Combinations of δ such that (ϵ, δ)-probabilistic DP is achieved when the Poisson is used, for various $\max_i a_i$ and ϵ equal to 1.5, 2, 2.5 and 3.

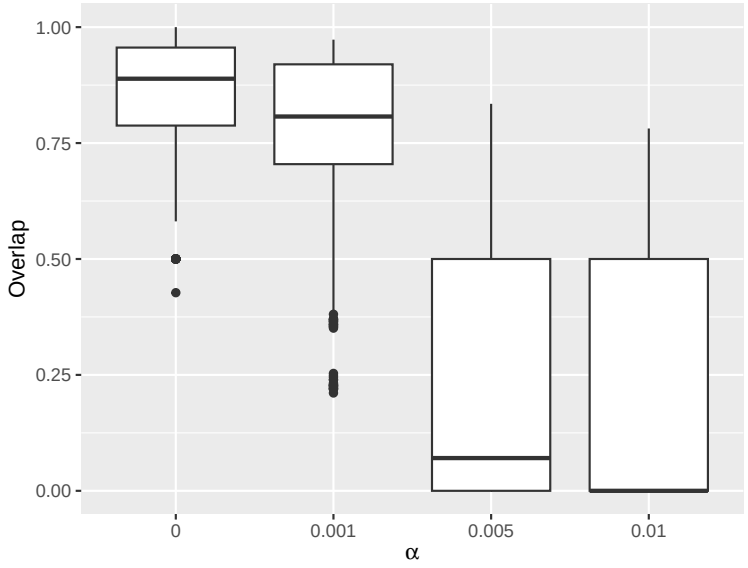

Fig. 3. Boxplots showing confidence interval overlap (utility) for different values of α.

as the negative binomial, are likely to be more favourable; i.e. will give better utility results, while also providing the same DP-type risk guarantees. This is because such distributions would introduce further tuning parameters in addition to α. Previous work suggests that such tuning parameter apply noise in a more efficient fashion [11]. These tuning parameters could be set to obtain certain ϵ or δ values.

We end with an interesting note in relation to DP. Somewhat counterintuitively, the reason why multinomial-based synthesis mechanisms (e.g. the multinomial Dirichlet synthesizer) can satisfy ϵ-DP – but count distributions cannot – is because with multinomial mechanisms there is a maximum synthetic count that any original count can take, namely n. With count distributions, any original count can be synthesized to any non-negative integer. To conceptualise why this causes the DP definition to fail, suppose in an intruder's data set – which, of course, is the actual data set minus the target individual – a certain cell has a count of 1. Then suppose in the synthetic data – generated by simulating from the Poisson – this cell has a count of 5. It is far more likely that this synthetic count originated from a cell with a count of 2 than from a count of 1 (11.7 times more likely), therefore the intruder can infer that this cell is a likely origin of the target. It is interesting therefore that with DP, disclosure risk is deemed to be at its greatest when the scope for potential movement between original and synthetic counts is at its greatest. This largely goes against the objectives of traditional SDC methods, which typically reduce risk by increasing the divergence from the original counts.

References

1. Abowd, J.M., Vilhuber, L.: How Protective Are Synthetic Data? In: Domingo-Ferrer, J., Saygın, Y. (eds.) Privacy in Statistical Databases. pp. 239–246. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
2. Balle, B., Wang, Y.X.: Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 394–403. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/balle18a.html
3. Blanchard, S., Jackson, J.E., Mitra, R., Francis, B.J., Dove, I.: A constructed English School Census substitute (2022), 10.17635/lancaster/researchdata/533
4. Bowen, C.M., Liu, F.: Comparative Study of Differentially Private Data Synthesis Methods. Statistical Science **35**(2), 280 – 307 (2020). https://doi.org/10.1214/19-STS742, https://doi.org/10.1214/19-STS742
5. Charest, A.S.: How can we analyze differentially-private synthetic datasets? Journal of Privacy and Conf. **2**(2) (2011). https://doi.org/10.29012/jpc.v2i2.589, https://journalprivacyconfidentiality.org/index.php/jpc/article/view/589
6. Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). Commun. ACM **64**(7), 33–35 (Jun 2021). https://doi.org/10.1145/3433638, https://doi.org/10.1145/3433638

7. Drechsler, J.: Differential privacy for government agencies—are we there yet? Journal of the American Statistical Association **118**(541), 761–773 (2023). https://doi.org/10.1080/01621459.2022.2161385, https://doi.org/10.1080/01621459.2022.2161385
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
9. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042, http://dx.doi.org/10.1561/0400000042
10. Goetz, M., Machanavajjhala, A., Wang, G., Xiao, X., Gehrke, J.: Publishing Search Logs - A Comparative Study of Privacy Guarantees. IEEE Trans. Knowl. Data Eng. **24**, 520–532 (03 2012). https://doi.org/10.1109/TKDE.2011.26
11. Jackson, J., Mitra, R., Francis, B., Dove, I.: On integrating the number of synthetic data sets m into the a priori synthesis approach. In: Domingo-Ferrer, J., Laurent, M. (eds.) Privacy in Statistical Databases. pp. 205–219. Springer International Publishing, Cham (2022)
12. Jackson, J., Mitra, R., Francis, B., Dove, I.: Using Saturated Count Models for User-Friendly Synthesis of Large Confidential Administrative Databases. Journal of the Royal Statistical Society Series A: Statistics in Society **185**(4), 1613–1643 (08 2022). https://doi.org/10.1111/rssa.12876, https://doi.org/10.1111/rssa.12876
13. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician **60**(3), 224–232 (2006)
14. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: Theory meets practice on the map. In: 2008 IEEE 24th international conference on data engineering. pp. 277–286. IEEE (2008)
15. McClure, D., Reiter, J.P.: Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data. Transactions on Data Privacy **5**(3), 535—-552 (2012)
16. Quick, H.: Generating Poisson-distributed differentially private synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society) **184**(3), 1093–1108 (2021). https://doi.org/https://doi.org/10.1111/rssa.12711, https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12711
17. Rinott, Y., O'Keefe, C.M., Shlomo, N., Skinner, C., et al.: Confidentiality and Differential Privacy in the Dissemination of Frequency Tables. Statistical Science **33**(3), 358–385 (2018)
18. Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. Journal of the Royal Statistical Society: Series A (Statistics in Society) **181**(3), 663–688 (2018)

This preprint has not undergone peer review or any post-submission improvements or corrections. The Version of Record of this contribution is published in Privacy in Statistical Databases. PSD 2024. Lecture Notes in Computer Science, vol 14915, and is available online at https://doi.org/10.1007/978-3-031-69651-0_7 .