Using Disclosure Risk Measures to Set ε in Differential Privacy

Zeki Kazan and Jerome P. Reiter

Duke University

January 25, 2023

Abstract

When agencies release sensitive data, balancing the desire for accuracy in the release with the need to protect individual privacy presents a major challenge. For releases satisfying differential privacy – the gold standard for sensitive data releases – this balance is reflected by the privacy parameter, ε, determining the amount of randomness in the release. In this work, we use Bayesian methods to provide a framework for selecting ε based on an agency's desired statistical disclosure risk profile. We focus on a scenario agencies commonly face, but which has received limited attention: data have already been collected, but the values are too sensitive to release exactly.

Differential privacy (DP) [3] is the gold standard for releasing sensitive data. It is used by large tech companies, like Google [4] and Meta [14], both internally and for public data releases. A variant of DP was used by the U. S. Census Bureau for the release of 2020 redistricting data [1] and will be used for the release of the 2020 demographic and housing characteristics file. Other government agencies, however, have been slow to adopt DP. Jörg Drechsler in *Differential Privacy for Government Agencies — Are We There Yet?* [2] lists the difficulty of setting the privacy parameter, ε, as one of the major reasons why DP may be challenging to use in government contexts. This concern is not unfounded: tuning the privacy parameters for the 2020 redistricting data release took the Census Bureau several years. Even after this lengthy process, their choices faced substantial criticism, for example in [9] and [15].

Despite its importance, the process for choosing ε has received surprisingly little attention in the literature. Prior work on theoretically justifying the setting of ε focuses on scenarios where the sensitive quantity is whether an individual is included in the data release [12] and settings where data have yet to be collected [8]. Many applicationsdo not fit into these categories. The common setting where data already exists and the values in the data are as or more sensitive than an individual's inclusion has received little attention and will be the focus of this article.

The ability to express the privacy guarantees of a DP data release in terms of interpretable quantities would allow agencies to provide improved justification for future choices of privacy parameters. Many argue – for example Hotz et al. [7] – that the most relevant quantities to individuals in data protected with DP are measures of the statistical disclosure risk to an individual due to the release. But the DP definition only relates ε to the amount of randomness in the privacy mechanism, making the setting of ε challenging in practice. The randomness in the privacy mechanism and the risk to individual privacy are, however, intimately related. In this article, we examine this relationship and demonstrate how Bayesian methods of measuring the disclosure risk can be used to select an appropriate ε for a release.

This article is organized as follows. We begin with an overview of DP in Section 1. We then establish the setting and our notation in Section 2. This is followed by a discussion of the process an agency would go through when using our methods in Section 3. We then outline our results in Section 4. Finally, we conclude with some brief commentary on future directions in Section 5.

1 Differential Privacy

Differential privacy is a formal privacy definition used to release statistics of sensitive datasets. Suppose we are interested in releasing a statistic of a dataset, D. The intuition behind DP is that the released value, $f(D)$, should have sufficient noise added such that it will not change substantially if any individual's row is removed from D. This is formalized by following definition from [3].

Definition 1 (Differential Privacy)**.** *For $\varepsilon \in (0, \infty)$, a mechanism f satisfies ε-differential privacy if for input data sets D, D' that differ only in that one row is added or removed and $S \subseteq Range(f)$,*

$$P[f(D) \in S] \leq e^{\varepsilon} P[f(D') \in S]. \tag{1}$$

On a technical note, the above definition is referred to as unbounded DP. An alternative definition of DP, called bounded DP considers datasets D, D' that differ only in that one row is changed (rather than added or removed). A mechanism f satisfying unbounded DP also satisfies bounded DP, with the ε increased by a factor of 2; thus, the following results also apply to mechanisms satisfying bounded DP. See [11] for more details on bounded and unbounded DP.

Commonly-used mechanisms satisfying DP involve releasing the sum of the statistic of interest and noise from an appropriate distribution, such as the Laplace distribution (if the statistic is bounded; [3]) or the two-sided geometric distribution (if the statistic is a count; [5]). These noise distributions are centered at zero and have variance inversely proportional to ε. Thus, for these mechanisms an ε near zero corresponds to more noise: better privacy at the cost of lower accuracy. A large ε, on the other hand, corresponds to less noise: better accuracy at the cost of less privacy.

2 Preliminaries

An agency possesses a sensitive dataset $\mathbf{Y}$ with n individuals (rows) and p variables (columns). We focus on the setting where whether an individual is included in the data is nonsensitive, but the values within the data are too sensitive to be released or by law cannot be released exactly. The agency possessing the data – henceforth referred to as the data holder – wishes to release a function of the data, T. In this article, we assume T has a discrete support (but may be many-dimensional). For the release, the data holder turns to DP and will release T^*, a noisy estimate of T under pure DP. The important question is then: how should the data holder set ε?

Consider now the perspective of a nefarious person – henceforth, the adversary – who desires to learn about individual $i \in \{1, \ldots, n\}$ in the data using the released value $T^* = t^*$. Let $y \in \mathbb{R}^p$ be the true value of Y_i. Before the release of T^*, the adversary could create a model, $\mathcal{M}$, for Y_i based on available public information. We assume that the model $\mathcal{M}$ involves the adversary fixing $Y_{-i} = y_{-i}$, a guess about the values of the other $n - 1$ rows. Let I_i be an indicator for individual i's presence in the data, which we assume to be 1 with probability 1 in this section.

Assuming that the adversary's model puts nonzero probability mass on the correct outcome, their prior can then be written in the form

$$P[Y_i = y \mid I_i = 1, \mathcal{M}] = q_i, \qquad P[Y_i \neq y \mid I_i = 1, \mathcal{M}] = 1 - q_i. \tag{2}$$

For fixed q_i, we can measure the risk of disclosure for individual i in a number of ways. Drawing from [13], one measure of the risk to person i is the relative disclosure risk, $r_i(q_i, t^*)$.

Definition 2 (Relative Disclosure Risk)**.** *For fixed data* $\mathbf{Y}$*, row* $i \in \{1, \ldots, n\}$*, prior model* $\mathcal{M}$*, and released* $T^* = t^*$*, the relative disclosure risk is the posterior-to-prior ratio*

$$r_i(q_i, t^*) = \frac{P[Y_i = y, I_i = 1 \mid T^* = t^*, \mathcal{M}]}{P[Y_i = y, I_i = 1 \mid \mathcal{M}]}. \tag{3}$$

The relative risk is closely related to the differential privacy ratio. Gong and Meng [6] show that ε-DP implies a bound on the relative risk.[1]

Theorem 1 (Theorem 1.3 in [6])**.** *If the release of* $T^* = t^*$ *satisfies* ε*-differential privacy, then for all* $i \in \{1, \ldots, n\}$*,*

$$e^{-2\varepsilon} \leq r_i(q_i, t^*) \leq e^{2\varepsilon}. \tag{4}$$

The relative risk, however, does not tell the full story. Hotz et al. (2022) [7] argue that, in addition to the relative risk, individuals represented in the data should also care about the absolute disclosure risk, $a_i(q_i, t^*)$.

Definition 3 (Absolute Disclosure Risk)**.** *For fixed data* $\mathbf{Y}$*, row* $i \in \{1, \ldots, n\}$*, prior model* $\mathcal{M}$*, and released* $T^* = t^*$*, the absolute disclosure risk is the posterior probability*

$$a_i(q_i, t^*) = P[Y_i = y, I_i = 1 \mid T^* = t^*, \mathcal{M}]. \tag{5}$$

3 Determining Allowable Risk

Return to the data holder's perspective and make no assumptions about q_i (i.e., the data holder does not consider the adversary's prior in their analysis). Theorem 1 implies a method for the data holder to choose ε. They can simply select a desired bound on the allowable relative risk, r_i^*, and set $\varepsilon = \log(r_i^*)$. For most applications, however, this strategy will result in a lower choice of ε than necessary and thus a less accurate release.

The issue with applying Theorem 1 directly is that the allowable relative risk should depend on the adversary's prior. If $q_i = 10^{-8}$, for example, then a relative risk of 100 would imply an absolute risk of 10^{-6}, which is reasonable in many contexts. But if $q_i = 0.01$, a relative risk of 100 would imply an absolute risk of one, which would be disastrous. But making assumptions about what q_i will be seen in practice is unreasonable; one cannot know what information will be available to adversaries in the future. To circumvent this issue, we recommend data holders consider the question: if an adversary were to exist with prior probability q_i, what is the maximum relative risk (or, equivalently, the maximum absolute risk) we would be willing to allow? This can be used to construct a function $r^* : [0, 1] \to [1, \infty]$ of the maximum allowable risk at each possible q_i.

[1] The theorem is adjusted to account for the use of unbounded, rather than bounded DP.

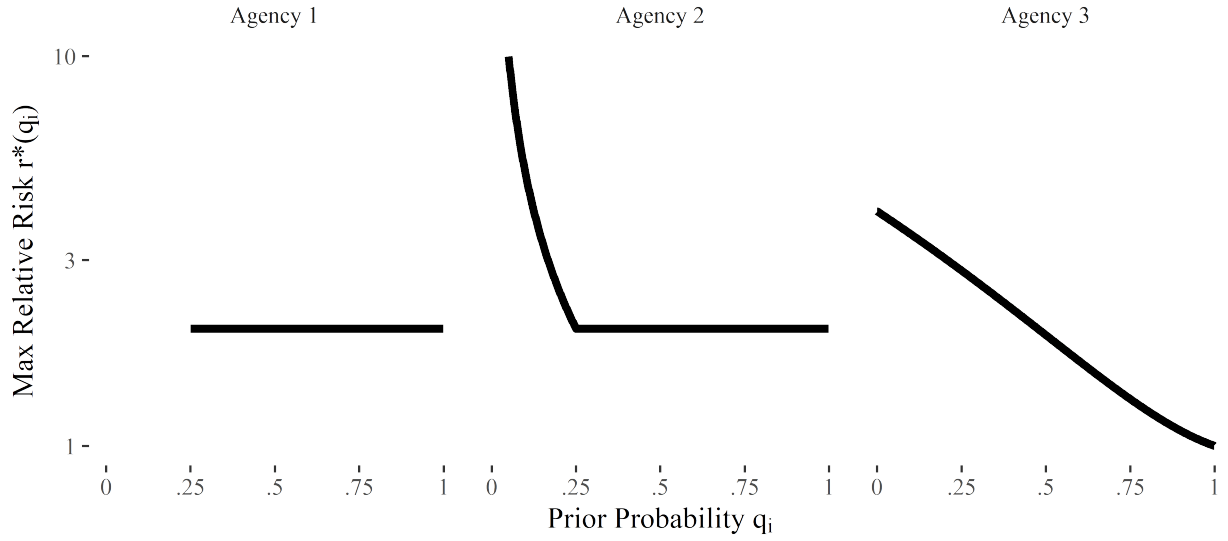


Figure 1: Three examples of the maximum allowable relative risk a data holder may be willing to allow, $r^*(q_i)$, as a function of an adversary's prior probability q_i. Agencies 1 and 2 set $r^*(q_i) = 2$ for $q_i \geq 0.25$. For $q_i < 0.25$, agency 1 sets $r^*(q_i) = \infty$, while agency 2 sets $r^*(q_i) = 0.5/q_i$. Agency 3 sets $r^*(q_i) = 2.3q_i^2 - 5.3q_i + 4$.

Figure 1 considers a few realistic constructions of r^*, corresponding to three agencies with different risk profiles. Agency 1 on the left panel wishes to bound the information gain by a factor of 2, but considers small values of q_i to represent wildly incorrect priors they are comfortable ignoring.[2] Agency 2 on the middle panel has a similar risk profile, but instead of ignoring small q_i, they bound $r^*(q_i) \leq 0.5/q_i$ for $q_i \leq 0.25$.[3] Finally, agency 3 on the right panel has a very different risk profile that does not map to simple bounds on the absolute and relative disclosure risk; they prefer a function r^* that is quadratic in q_i.

4 Results

For a given function r^* selected by the data holder, the ε that should be used for the release can be determined. This is due to the following result.

Theorem 2. *For individual i, fix the adversary's prior probability of disclosure, q_i, and a desired bound on the relative disclosure risk, $r^*(q_i)$. Then any statistic $T^* = t^*$ released under ε-DP with*

$$\varepsilon \leq \frac{1}{2} \log \left(\frac{1 - q_i}{\frac{1}{r^*(q_i)} - q_i} \right) \tag{6}$$

will satisfy $r_i(q_i, t^*) \leq r^*(q_i)$ for any intruder model $\mathcal{M}$.

[2] See Example 7.2 in [10] for a setting where ignoring small q_i might be appropriate.

[3] It can be shown this is equivalent to bounding the absolute disclosure risk at $a_i(q_i, t^*) \leq 0.5$ for $q_i \leq 0.25$

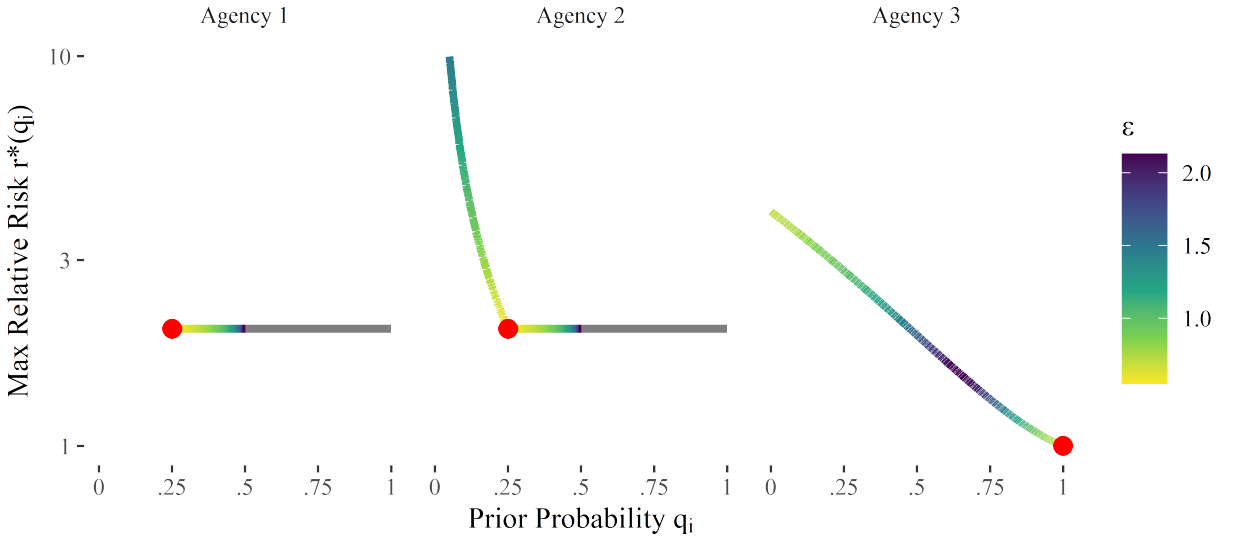


Figure 2: Plot of the ε's given by Equation (6) for the three agencies in Figure 1. The red dots correspond to the minimal ε's from Equation (7) and gray regions imply the ε bound is infinite. For agencies 1 and 2, the minimum is achieved at $q_i = 0.25$ and corresponds to $\varepsilon = \frac{1}{2}\log(3) \approx 0.55$; for agency 3, the minimum is achieved as $q_i \to 1$ and corresponds to $\varepsilon = \frac{1}{2}\log(\frac{10}{3}) \approx 0.60$.

The proof of Theorem 2 is included in Section 6. By Theorem 2, to achieve $r_i(q_i, t^*) \leq r^*(q_i)$ for all q_i, the data holder should set

$$\varepsilon = \min_{q_i \in [0,1]} \frac{1}{2} \log\left(\frac{1 - q_i}{\frac{1}{r^*(q_i)} - q_i} \right). \tag{7}$$

Minimizing this function is straightforward in practice via a search over $q_i \in [0, 1]$. To demonstrate, Figure 2 plots the ε from Equation (6) for each of the agencies described in Section 3; the red dots correspond to the q_i that produces the ε in Equation (7). Recall that for agencies 1 and 2, the simple method via Theorem 1 would recommend $\varepsilon = \frac{1}{2}\log(2) \approx 0.35$. Our method instead recommends $\varepsilon = \frac{1}{2}\log(3) \approx 0.55$ for both, which represents a substantial increase in accuracy. For agency 3, our method recommends $\varepsilon = \frac{1}{2}\log(10/3) \approx 0.60$, whereas the simpler method would have set $\varepsilon \to 0$ as $r^*(q_i) \to 1$.

5 Commentary

This is a partial draft of nearly complete work. The final draft will also include an analysis of the more general setting where whether an individual is present in the data may be sensitive, in addition to the value. It will also include a comparison between our recommended ε from the setting where only inclusion is sensitive and the ε recommended by Lee and Clifton (2011) [12], who examined this problem in the sensitive values only setting (making different assumptions on the adversary than we do). The final draft will also include a discussion of composition, both for performing multiple releases simultaneously and for calculating the risk due to sequential releases.

It will also include more detailed examples of the exact process an agency would go through in constructing their r^* and using the function determine ε.

Ongoing work considers a number of extensions in addition to the above. One important direction involves generalizing what constitutes a privacy violation. If, for example, $\boldsymbol{Y}$ consists of income data, an adversary learning that $Y_i \in [y - \$5{,}000, y + \$5{,}000]$ may constitute a privacy violation (rather than merely an adversary learning that $Y_i = y$). Establishing this generalization then allows for extension to the release of statistics with continuous support, such as means, regression coefficients, etc. More broadly, it is possible similar results hold for (ε, δ)-DP, zero-concentrated DP, and other privacy definitions an agency may wish to use.

References

[1] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, et al. The 2020 census disclosure avoidance system topdown algorithm. *arXiv preprint arXiv:2204.08986*, 2022.

[2] Joerg Drechsler. Differential privacy for government agencies–are we there yet? *arXiv preprint arXiv:2102.08847*, 2021.

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[4] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[5] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6):1673–1693, 2012.

[6] Ruobin Gong and Xiao-Li Meng. Congenial differential privacy under mandated disclosure. In *Proceedings of the 2020 ACM-IMS on foundations of data science conference*, pages 59–70, 2020.

[7] V Joseph Hotz, Christopher R Bollinger, Tatiana Komarova, Charles F Manski, Robert A Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D Spencer. Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, 119(31), 2022.

[8] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin C Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 398–410. IEEE, 2014.

[9] Christopher T Kenny, Shiro Kuriwaki, Cory McCartan, Evan Rosenman, Tyler Simko, and Kosuke Imai. The impact of the us census disclosure avoidance system on redistricting and voting rights analysis. *arXiv preprint arXiv:2105.14197*, 2021.

[10] Daniel Kifer, John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census. *arXiv preprint arXiv:2209.03310*, 2022.

[11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

[12] Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[13] David McClure and Jerome P Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.*, 5(3):535–552, 2012.

[14] Chaya Nayak. New privacy-protected facebook data for independent research on social media's impact on democracy. *Facebook Research*, 2020.

[15] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, volume 109, pages 403–08, 2019.

6 Proof of Theorem 2

Proof. We begin with the expression for the relative disclosure risk and apply Bayes' Theorem.

$$r_i(q_i, t^*) = \frac{P[Y_i = y \mid T^* = t^*, \mathcal{M}]}{P[Y_i = y \mid \mathcal{M}]} = \frac{\frac{P[T^*=t^* \mid Y_i=y]\, P[Y_i=y \mid \mathcal{M}]}{P[T^*=t^* \mid \mathcal{M}]}}{P[Y_i = y \mid \mathcal{M}]} = \frac{P[T^* = t^* \mid Y_i = y]}{P[T^* = t^* \mid \mathcal{M}]}. \tag{8}$$

We now expand the denominator via the law of total probability.

$$r_i(q_i, t^*) = \frac{P[T^* = t^* \mid Y_i = y, \mathcal{M}]}{P[T^* = t^* \mid Y_i = y, \mathcal{M}]\, P[Y_i = y \mid \mathcal{M}] + \sum_{y' \neq y} P[T^* = t^* \mid Y_i = y', \mathcal{M}]\, P[Y_i = y' \mid \mathcal{M}]} \tag{9}$$

$$= \frac{1}{q_i + \sum_{y' \neq y} \frac{P[T^*=t^* \mid Y_i=y', \mathcal{M}]}{P[T^*=t^* \mid Y_i=y, \mathcal{M}]}\, P[Y_i = y' \mid \mathcal{M}]}. \tag{10}$$

By the definition of DP, the ratio in the summation is bounded below by $e^{-2\varepsilon}$ for any t^*. Thus,

$$r_i(q_i, t^*) \leq \frac{1}{q_i + \sum_{y' \neq y} e^{-2\varepsilon}\, P[Y_i = y' \mid \mathcal{M}]} = \frac{1}{q_i + e^{-2\varepsilon}\,(1 - q_i)}. \tag{11}$$

Since $\varepsilon \leq \frac{1}{2} \log\left(\frac{1-q_i}{\frac{1}{r^*(q_i)} - q_i}\right)$, it follows that $e^{-2\varepsilon} \geq \frac{\frac{1}{r^*(q_i)} - q_i}{1-q_i}$. Thus, as desired,

$$r_i(q_i, t^*) \leq \frac{1}{q_i + \frac{\frac{1}{r^*(q_i)} - q_i}{1-q_i}\,(1 - q_i)} = \frac{1}{q_i + \left(\frac{1}{r^*(q_i)} - q_i\right)} = r^*(q_i). \tag{12}$$

□