Controlling Privacy Loss in Sampling Schemes: an Analysis of Stratified and Cluster Sampling

Mark Bun
Department of Computer Science at Boston University, USA

Jörg Drechsler
Institute for Employment Research, Germany
Joint Program in Survey Methodology, University of Maryland, USA

Marco Gaboardi
Department of Computer Science at Boston University, USA

Audra McMillan[1]
Apple, USA

Jayshree Sarathy
Harvard John A. Paulson School of Engineering and Applied Sciences, USA

Abstract

Sampling schemes are fundamental tools in statistics, survey design, and algorithm design. A fundamental result in differential privacy is that a differentially private mechanism run on a *simple random* sample of a population provides stronger privacy guarantees than the same algorithm run on the entire population. However, in practice, sampling designs are often more complex than the simple, data-independent sampling schemes that are addressed in prior work. In this work, we extend the study of privacy amplification results to more complex, data-dependent sampling schemes. We find that not only do these sampling schemes often fail to amplify privacy, they can actually result in privacy degradation. We analyze the privacy implications of the pervasive cluster sampling and stratified sampling paradigms, as well as provide some insight into the study of more general sampling designs.

2012 ACM Subject Classification Security and privacy → Privacy protections

Keywords and phrases privacy, differential privacy, survey design, survey sampling

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

1 Introduction

Sampling schemes are fundamental tools in statistics, survey design, and algorithm design. For example, they are used in social science research to conduct surveys on a random sample of a target population. They are also used in machine learning to improve the efficiency and accuracy of algorithms on large datasets. In many of these applications, however, the datasets are sensitive and privacy is a concern. Intuition suggests that (sub)sampling a dataset before analysing it provides additional privacy, since it gives individuals plausible deniability about whether their data was included or not. This intuition has been formalized for some types of sampling schemes (such as simple random sampling with and without replacement and Poisson sampling) in a series of papers in the differential privacy literature [23, 33, 11, 31]. Such *privacy amplification by subsampling* results can provide tight privacy accounting when analysing algorithms that incorporate subsampling, e.g. [32, 1, 21, 28, 19]. However, in practice, sampling designs are often more complex than the simple, data independent sampling schemes that are addressed in prior work. In this work, we extend the study of privacy amplification results to more complex and data dependent sampling schemes.

We consider the setting described in Figure 1. We have a *population* P and a historic or auxiliary data set H which is used to inform the sampling design. We think about the sampling scheme as a

[1] Corresponding author

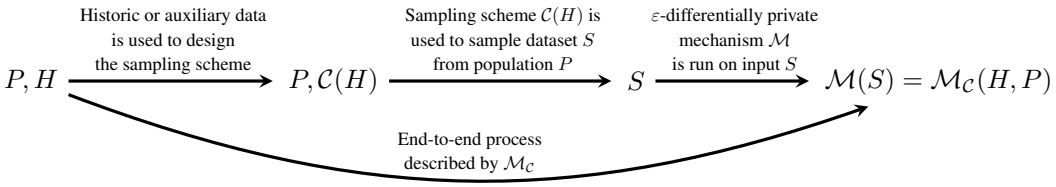

Figure 1 The structure of using a data-dependent sampling scheme.

function $\mathcal{C}(H)$ of the historic or auxiliary data H. Using this sampling scheme, we draw a sample S from the population P, on which we run the differentially private mechanism $\mathcal{M}$. We can think about these multiple steps as comprising a mechanism $\mathcal{M}_{\mathcal{C}}(H, P)$ working directly on the population P and the historic data H whose privacy depends on the privacy of the mechanism $\mathcal{M}$ and on the properties of the sampling scheme $\mathcal{C}(H)$. While this is the general framework for the problem we study, we state the technical results in this paper for the simplified case where $H = P$; see Section 2.1 for further discussion.

1.1 Our contributions

We primarily focus on two classes of sampling schemes that are common in practice: *cluster sampling* and *stratified sampling*. In (single-stage) cluster sampling, the population arrives partitioned into disjoint clusters. A sample is obtained by selecting a small number of clusters at random, and then including all of the individuals from those chosen clusters. In stratified sampling, the population is partitioned into "strata." Individuals are then sampled at different rates according to which stratum they belong to.

For these more complex schemes, we find that privacy amplification can be negligible even when only a small fraction of the population is included in the final sample. Moreover, in settings where the sampling design is data dependent, privacy degradation can occur – some sampling designs can actually make privacy guarantees worse. Intuitively, this is because the sample design itself can reveal sensitive information. Our goal in this paper is to explain how and why these phenomena occur and introduce technical tools for understanding the privacy implications of concrete sampling designs.

Understanding randomised and data-dependent sampling. It is simple to show that deterministic, data-dependent sampling designs do not achieve privacy amplification, and can suffer privacy degradation. Motivated by this observation, we start by studying the privacy implications of randomised and data-dependent sampling, attempting to isolate their effects in the simplest possible setting.

Specifically, we aim to understand sampling schemes of the following form: For a possibly randomised function f (an "allocation rule"), sample $f(P)$ individuals uniformly from P without replacement. In Section 3, we study the case where f is randomised but data-independent, i.e., the number of individuals samples is drawn from a distribution that does not depend on P. We give an essentially complete characterization of what level of amplification is possible in terms of this distribution.

In Section 4, we turn our attention to data-dependent sampling. We identify necessary conditions for allocation rules f to enable privacy amplification by way of a hypothesis testing perspective; intuitively, for f to be a good amplifier, every differentially private algorithm must fail to distinguish the distributions of $f(P)$ and $f(P')$ for neighboring P, P'. We also study a specific natural allocation rule called *proportional allocation* that is commonly applied in stratified sampling. We design a simple randomised rounding method that offers a minor change to the way proportional allocation is generally implemented in practice, but that offers substantially better privacy amplification.

Cluster sampling. In Section 5, we study cluster sampling where a population partitioned into k clusters is sampled by selecting m clusters uniformly at random without replacement. Our results give tradeoffs between the privacy amplification achievable and the sizes of the clusters. In particular, privacy amplification is possible when all of the clusters are small. As the cluster sizes grow, the best achievable privacy loss rapidly approaches the baseline of the privacy guarantee of $\mathcal{M}$. We provide some insight into these results by connecting the privacy loss to the ability of a hypothesis test to determine from a differentially private output which clusters were included in the sample.

Stratified sampling. Building on our randomised rounding method for the "single-stratum" case, we show that stratified sampling with the proportional allocation rule amplifies privacy. Unfortunately, as in the single stratum case, there are natural lower bounds which limit extending this approach to other common allocation rules.

A common goal when choosing an allocation function $\boldsymbol{f}$ (a function which decides how many samples to draw from each stratum) is to minimise the variance of a particular statistic. For example, the popular Neyman allocation is the optimal allocation for computing the population mean. A natural question then is how to define and compute the optimal allocation when privacy is a concern? In this work, we will formulate the notion of an optimal allocation under privacy constraints. This formulation is somewhat subtle since the privacy implications of different allocation methods need to be properly accounted for. Our goal is to initiate the study of alternative allocation functions that may prove useful when privacy is a concern.

1.2 Related work

Several works have studied the privacy amplification of simple sampling schemes. Kasiviswanathan et al. [23] and Beimel et al. [9] showed that applying Poisson sampling before running a differentially private mechanism improves its end-to-end privacy guarantee. Subsequently, Bun et al. [11] analyzed simple random sampling with replacement in a similar way. Beimel et al. [10], Bassily et al. [7], and Wang et al. [34] analyzed simple random sampling without replacement. Imola and Chaudhuri [20] provide lower and upper bounds on privacy amplification when sampling from a multidimensional Bernoulli family, a task which has direct applications to Bayesian inference. Balle et al. [5] unified the analyses of privacy amplification of these mechanisms using the lenses of *probabilistic couplings*, an approach that we also use in this paper. The effects that sampling can have on differentially private mechanisms is also studied from a different perspective in [13]. However, none of the prior works consider the privacy amplification of more complex, data-dependent sampling schemes commonly used in practice. To the best of our knowledge, this paper is the first to do so.

2 Background

2.1 Data-dependent sampling schemes

In the data-driven sciences, data is often obtained by sampling a fraction of the population of interest. This sample can be created in a wide variety of ways, referred to as the sample design. Sample designs can vary from simple designs such as taking a uniformly random subset of a fixed size, to more complex data-dependent sampling designs like cluster or stratified sampling. Data-dependent sampling designs achieve accuracy and meet budgeting goals by using historic or auxiliary data to exploit structure in the population. The privacy implications of simple random sampling are quite well understood from prior work. In this work, we will move beyond simple random sampling to analyse the privacy implications of more complex sampling designs, including data-dependent sampling.

An outline of the schema for data dependent sampling designs is given in Figure 1. There are ostensibly two datasets: H, the historic or auxiliary data that is used to design the sampling scheme

$\mathcal{C}(H)$, and P, the current population that is sampled from. For the remainder of this paper, we make the simplifying assumption that $H = P$. That is, we will not distinguish between the historic or auxiliary data and the "current" data. Even if we only care about maintaining the privacy of the individuals in population P, this assumption is required if we have no information about the relationship between H and P. Thus, we view the function $\mathcal{M}_{\mathcal{C}}(P, H)$ as simply a function of P. We will refer to the size of the sample S as the *sample size*, and the fraction $|S|/|P|$ as the *sampling rate*.

More refined models can be obtained by imposing specific assumptions on the relationship between H and P, for example, by modeling the temporal correlation between historic and current data. We leave this for future work.

2.2 Differential privacy

Differential privacy (DP) is a measure of stability for randomised algorithms. It bounds the change in the distribution of the outputs of a randomised algorithm when provided with two datasets differing on the data of a single individual. We will call such datasets neighboring. In order to formalise what a "bounded change" means, we define (ε, δ)-indistinguishability. Two random variables P and Q over the same probability space are (ε, δ)-*indistinguishable* if for all sets of outcomes E over that probability space,

$$e^{-\varepsilon}(\Pr(Q \in E) - \delta) \leq \Pr(P \in E) \leq e^{\varepsilon} \Pr(Q \in E) + \delta.$$

If $\delta = 0$ then we will say that P and Q are ε-indistinguishable. For any $n \in \mathbb{N}$, let $\mathcal{U}^n$ be the set of all datasets of size n over elements of the data universe $\mathcal{U}$. Let $\mathcal{U}^* = \cup_{n \in \mathbb{N}} \mathcal{U}^n$ be the set of all possible datasets. We discuss two privacy definitions in this work corresponding to two different neighboring relations: *unbounded* differential privacy and *bounded* differential privacy. We will say two datasets are *unbounded neighbors* if one can be obtained from the other by adding or removing a single data point, and *bounded* neighbors if they have the same size, and one can be obtained from the other by changing the data of a single individual.

▶ **Definition 1.** *A mechanism $\mathcal{M} : \mathcal{U}^* \to \mathcal{O}$ is (ε, δ)-unbounded (resp. bounded) differentially private (DP) if for all pairs of unbounded (resp. bounded) neighboring datasets P and P', $\mathcal{M}(P)$ and $\mathcal{M}(P')$ are (ε, δ)-indistinguishable.*

We will use both bounded and unbounded DP throughout the paper as they are appropriate in different settings. When considering which notion to choose, it is important to consider which guarantees are meaningful in context. For example, it will be common in the sample designs we cover for the size of the sample S (see Figure 1) to be data-dependent. When considering these sampling designs, we will focus on mechanisms $\mathcal{M}$ that satisfy unbounded DP since bounded DP does not protect the sample size. However, bounded DP may be more appropriate for the privacy guarantee on $\mathcal{M}_{\mathcal{C}}$ in applications where it is unrealistic to assume that an individual can choose not to be part of the auxiliary dataset or the population. For example, the auxiliary data may be administrative data, data from a mandatory census, or data from a monopolistic service provider. Results and intuition are often similar between unbounded and bounded DP, although care should be taken when translating between the two notions. We note in particular that any ε-unbounded DP mechanism is 2ε-bounded DP.

2.3 Privacy amplification with uniform random sampling

Sampling does not provide strong differential privacy guarantees on its own. But when employed as a pre-processing step in a differentially private algorithm, it can amplify existing privacy guarantees. Intuitively, this is because if the choice of individuals is kept secret, sampling provides data subjects

the plausible deniability to claim that their data was or was not in the final data set. This effect was first explicitly articulated in [29], and a formal treatment of the phenomenon was given in [5]. Three types of sampling are analysed in [5]: simple random sampling with replacement, simple random sampling without replacement, and Poisson sampling. In all three settings the privacy amplification is proportional to the probability of an individual not being included in the final computation. To gain some intuition before we move into the more complicated sampling schemes that are the focus on this paper, let us state and discuss the results from [5].

▶ **Theorem 2.** *[5] Let $\mathcal{C}$ be a sampling scheme that samples m values out of n possible values without replacement. Given an (ε, δ)-bounded differentially private mechanism $\mathcal{M}$, we have that $\mathcal{M}_{\mathcal{C}}$ is (ε', δ')-bounded differentially private for $\varepsilon' = \log(1 + \frac{m}{n}(e^{\varepsilon} - 1))$ and $\delta' = \frac{m}{n}\delta$.*

To consider the implications of this result, notice that $\varepsilon' \leq \varepsilon$ for all values of $m \leq n$ so the sampled mechanism $\mathcal{M}_{\mathcal{C}}$ is strictly more private than the original mechanism $\mathcal{M}$. Further, taking into account the following two approximations which hold for small x,

$$e^x - 1 \approx x \tag{1}$$

$$\log(1 + x) \approx x, \tag{2}$$

we have that for small ε, $\varepsilon' \approx \frac{m}{n}\varepsilon$. So the degree of amplification in both parameters is roughly proportional to the sampling rate m/n.

2.4 How do people use subsampling amplification results?

Suppose we have a dataset that contains n records, and we want to estimate the proportion of individuals that satisfy some attribute in an ε-DP manner. Let us set our target privacy guarantee to be $\varepsilon = 1$. To do this, we can simply compute the proportion non-privately and add Laplace noise with scale $1/n$. But, if we know that the dataset is a secret and simple random sample from a population of $100n$ individuals, then adding Laplace noise with scale $1/n$ as before will actually yield a stronger privacy guarantee of $\varepsilon' = 0.01$ for the underlying population. To get $\varepsilon' = 1$, we will need to add noise with scale only $1/(100n)$. In other words, the secrecy of the sample means that the computation has more privacy inherently, and therefore, we can add less noise in order to achieve the desired privacy guarantee.

Existing DP data analysis tools such as DP Creator [18, 17] employ privacy amplification results to provide better statistical utility. For example, the DP Creator interface prompts the user to input the population size if the data is a secret and random sample from a larger population of known size and take advantage of the resulting boost in accuracy without changing the privacy guarantee.

As we discussed before, privacy amplification results are also used to analyse algorithms that incorporate subsampling as one of their components. Privacy amplification results permit a tighter analysis of the privacy that these algorithm can guarantee. In particular, these algorithms are quite common in learning tasks, e.g. [32, 1, 21, 28, 19].

3 Randomised data-independent sampling rates

While we are ultimately interested in data-dependent sampling designs, we begin with an extension of Theorem 2 to non-constant but data-*independent* sampling rates. Prior results on privacy amplification by subsampling [23, 33, 11, 31, 6] all focus on constant sampling rates where the sampling rate (the fraction of the data set sampled) is fixed in advance. However, we will eventually see that randomising the sample rate is essential to privacy amplification when the target rate is data dependent. To work toward this eventual discussion, we first study the data-independent case to gain intuition for what

properties of the distribution on sampling rates characterize how much privacy amplification is possible.

Suppose that there is a random variable t on $[n]$ and the sampling scheme is as follows: given a dataset P, a sample m is drawn from t, and then m subjects are drawn without replacement from P to form the sample S. In this section we consider unbounded differential privacy[2] for $\mathcal{M}$ and bounded differential privacy for $\mathcal{M}_{\mathcal{C}}$, where the total number of cases, n, is known and fixed. A simple generalisation of Theorem 2 immediately implies that the privacy loss of this randomised scheme is no worse than if t was concentrated on the maximum value in its support. However, prior work does not give insight into what happens when t is concentrated below its maximum or is evenly spread. What property of the distribution characterises its potential for privacy amplification? The following theorem characterizes the privacy amplification of sampling without replacement with data-independent randomised sampling rates.

▶ **Theorem 3.** *Let P be a dataset of size n, let t be a distribution over $\{0, 1, \ldots, n\}$, and let $\mathcal{C} : \mathcal{X} \to \mathcal{U}^*$ be the randomised, dataset-independent sampling scheme that randomly draws $m \sim t$ and samples m records from P without replacement. Define the distribution $\tilde{t}$ on $[n]$ where $\tilde{t}(m) \propto e^{\varepsilon m} \cdot t(m)$ for all $m \in [n]$.*

Upper bound: *Let $\mathcal{M} : \mathcal{U}^* \to \mathcal{O}$ be an ε-unbounded DP algorithm. Then, $\mathcal{M}_{\mathcal{C}}$ is ε'-bounded DP, where*

$$\varepsilon' = \log\left(1 + \frac{1}{n} \cdot \mathbb{E}_{m\sim\tilde{t}}[m] \cdot (e^{\varepsilon} - 1)\right).$$

Lower bound: *There exists neighboring datasets P and P' of size n, and an ε-unbounded DP mechanism $\mathcal{M}$ such that if $\mathcal{M}_{\mathcal{C}}(P)$ and $\mathcal{M}_{\mathcal{C}}(P')$ are ε'-indistinguishable then*

$$\varepsilon' \geq -\log\left(1 - \frac{1}{n} \cdot \mathbb{E}_{m\sim\tilde{t}}[m] \cdot (1 - e^{-\varepsilon})\right)$$

First notice that Theorem 3 comports with the generalization of Theorem 2; as expected, if the support of t is contained within $[0, m']$ then $\mathbb{E}_{m\sim\tilde{t}}[m] \leq m'$, so the randomised scheme is at least as private as if t was concentrated on m'. It also determines that the property of t that determines the privacy amplification is $\mathbb{E}_{m\sim\tilde{t}}[m]$, the expectation of an exponential re-weighting of the distribution that gives more weight to larger sample sizes. When ε is small, the simple approximations $e^x - 1 \approx x$, $1 - e^{-x} \approx x$, and $\log(1 + x) \approx x$ mean that both the upper and lower bounds amount to

$$\varepsilon' \approx \frac{\mathbb{E}_{m\sim\tilde{t}}[m]}{n} \cdot \varepsilon.$$

Due to the exponential re-weighting,

$$\mathbb{E}_{m\sim\tilde{t}}[m] = \frac{\sum_{m=0}^{n} e^{\varepsilon m} \Pr(t = m) m}{\sum_{m=0}^{n} e^{\varepsilon m} \Pr(t = m)}$$

rapidly approaches n as the weight of t on values close to n increases. Intuitively, this means that even a small probability of sampling the entire dataset can be enough to ensure that there is no privacy amplification, even if the mode of t is much smaller than n. Conversely, if t is a light tailed distribution (say, subgaussian) concentrated on a value much smaller than n, then privacy amplification is possible.

For example, suppose that t is a truncated Gaussian on $[0, n]$ with mean $n/2$ and standard deviation σ. If t is highly concentrated then we expect the privacy guarantee of $\mathcal{M}_{\mathcal{C}}$ to be $\approx \varepsilon/2$. As σ grows

[2] Note that we must use the unbounded differential privacy definition for $\mathcal{M}$ in this setting; otherwise, the sample size m would be fixed.

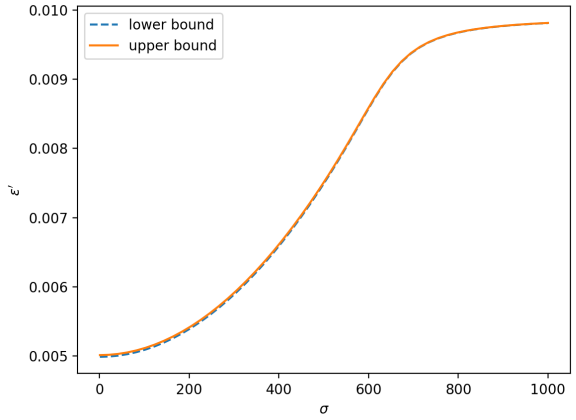

Figure 2 Numerical computation of the upper and lower bounds from Theorem 3 when t is truncated Gaussian supported on $[0, n]$ with mean $n/2$, where $n = 10^4$ and standard deviation σ varies from 1 to 10^3. The privacy parameter of the mechanism $\mathcal{M}$ is 0.01.

we expect the privacy guarantee to tend towards ε as more weight is placed near n. In Figure 2, we illustrate the bounds of Theorem 3 numerically with this Gaussian example. We can see that when $n = 10,000$ and $\sigma \approx 800$, the privacy guarantee of $\mathcal{M}_C$ is already close to $\varepsilon = 0.01$, the privacy guarantee of $\mathcal{M}$.

4 Data-dependent sampling rates

We now turn our attention to sampling schemes where sampling rates may depend on the data. The results in this section are motivated by *stratified sampling*, where the population is stratified into k disjoint sub-populations called strata, and an allocation function is used to determine how many samples to draw from each stratum. We will discuss stratified sampling with $k > 1$ in Section 6, but for simplicity and clarity, we first focus on the "single stratum" case. In this section, we develop tools and statements that we expect to be more broadly useful in understanding complex sampling designs.

Specifically, we consider the sampling design where one selects a number of cases according to a data-dependent function, and then samples that many cases via simple random sampling. That is, let $\tilde{f} : \mathcal{U}^* \to \mathbb{N}$ be a possibly randomised function and let $\mathcal{C}_f$ be the sampling function that on input P samples $f(P)$ data points uniformly without replacement from P. If $\mathcal{M}$ is an ε-DP algorithm, then how private is $\mathcal{M}_{\mathcal{C}_f}$?

4.1 Sensitivity and privacy degradation

We first observe that if the function f used to determine sample size is highly sensitive, then privacy *degradation* may occur. That is, if the number of cases sampled may change dramatically on neighboring populations, then the output of a DP mechanism can immediately be used to distinguish between those populations. For example, suppose P and P' are neighboring populations, and f is a function where $f(P) = m$ and $f(P') = m + \Delta$. (That is, the local sensitivity of f at P is at least Δ.) Consider the ε-DP algorithm $\mathcal{M}^{\text{count}}$ that, on input a sample S, outputs the noisy count $|S| + \text{Lap}(1/\varepsilon)$ of the number of cases in the sample. Then $\mathcal{M}^{\text{count}}_{\mathcal{C}_f}(P)$ is distributed as $m + \text{Lap}(1/\varepsilon)$ whereas $\mathcal{M}^{\text{count}}{}_{\mathcal{C}_f}(P')$ is distributed as $m + \Delta + \text{Lap}(1/\varepsilon)$. When $\Delta \gg 1$, these distributions are far apart; the privacy loss between these two populations is $\Delta \cdot \varepsilon \gg \varepsilon$.

Thus, a *necessary* condition for achieving privacy amplification (rather than degradation) is that the function f has low sensitivity. In the following sections, we explore other conditions on low sensitivity functions that are necessary and sufficient for amplification.

4.2 Data dependent sampling and hypothesis testing

We established in the previous section that using a deterministic function to determine sample size results in privacy degradation. This raises the question: how much randomness is necessary to ensure privacy control? That is, what can we say about a randomised function $\tilde{f} : \mathcal{U}^* \to \mathbb{N}$ with the property that $\mathcal{M}_{\mathcal{C}_{\tilde{f}}}$ is ε'-DP for every ε-DP mechanism $\mathcal{M}$? In this section we establish a connection between the amplification properties of a function $\tilde{f}$ and hypothesis testing.

A simple hypothesis testing problem is specified by two distributions X and Y. A hypothesis test H for this problem attempts to determine whether the samples given as input are drawn i.i.d from X or from Y. If a hypothesis test is only given a single sample then we define the advantage of H to be

$$\text{adv}(H; X, Y) = \Pr_{m \sim X}[H(m) = X] - \Pr_{m \sim Y}[H(m) = X].$$

That is, the advantage is a measure of how likely the hypothesis test H is to correctly guess which distribution the sample was drawn from. The closer the advantage is to 1, the better the test is at distinguishing X from Y.

One common explanation of differential privacy is that an algorithm is differentially private if it is impossible to confidently guess from the output which of two neighbouring datasets was the input dataset. This interpretation can be formalised, following [35], by noting that if $\mathcal{M}$ is ε-DP and P and P' are neighbouring populations then for every hypothesis test H,

$$\text{adv}(H; \mathcal{M}(P), \mathcal{M}(P')) \leq e^{\varepsilon} - 1 \approx \varepsilon.$$

We can establish a similar bound and interpretation of what it means for $\tilde{f}$ to amplify or preserve privacy. Suppose that $\tilde{f}$ is such that $\mathcal{M}_{\mathcal{C}_{\tilde{f}}}$ is ε'-DP for every ε-DP mechanism $\mathcal{M}$. Then in particular, for every ε-DP hypothesis test H, we have that $H(\mathcal{M}_{\mathcal{C}_{\tilde{f}}}(P))$ and $H(\mathcal{M}_{\mathcal{C}_{\tilde{f}}}(P'))$ are ε'-indistinguishable. Now, if we consider only hypothesis tests $H : \mathbb{N} \to \{\tilde{f}(P), \tilde{f}(P')\}$ that simply look at the size of the sample $\mathcal{C}_{\tilde{f}}(\cdot)$, then we can formalise this statement in the following way.

▶ **Proposition 4.** *Suppose $\tilde{f} : \mathcal{U}^* \to \mathbb{N}$ is such that for all ε-DP mechanisms $\mathcal{M}$, we have that $\mathcal{M}_{\mathcal{C}_{\tilde{f}}}$ is ε'-DP. Then for all neighboring datasets P, P', we have*

$$\max \text{adv}(H; \tilde{f}(P), \tilde{f}(P')) \leq e^{\varepsilon'} - 1,$$

where the optimisation is over all hypothesis tests H such that for all $x \in \mathbb{N}$, and $b \in \{0, 1\}$, $e^{-\varepsilon} \Pr(H(x) = b) \leq \Pr(H(x + 1) = b) \leq e^{\varepsilon} \Pr(H(x) = b)$.

This result helps us build intuition for what type of survey designs could possibly amplify privacy. If $\tilde{f}$ results in privacy amplification then for any pair of neighbouring populations P and P', the distributions $\tilde{f}(P)$ and $\tilde{f}(P')$ must be close enough that they can not be distinguished between by any hypothesis test H such that $\log H$ is ε-Lipschitz. From this perspective the result in Section 4.1 follows from the fact that if $\tilde{f}$ is deterministic with high sensitivity then we can define an appropriate hypothesis test with large advantage based on $\mathcal{M}^{\text{count}}$. This is a useful perspective to keep in mind throughout the remainder of the paper.

One consequence of this perspective is a lower bound on how well we can emulate a desired deterministic function f while controlling or amplifying privacy. Suppose that absent privacy concerns, an analyst has determined that they want to use a function f to determine the sample size. However,

to avoid privacy degradation they replace f with a randomised function $\tilde{f}$. How close can $\tilde{f}$ get to f while maintaining or amplifying the original privacy level? We can obtain a lower bound on expected closeness of $f(P)$ and $\tilde{f}(P)$ by relating it to the well studied problem of estimation lower bounds in differential privacy.

▶ **Proposition 5.** *Let $f : \mathcal{U}^* \to \mathbb{R}$ and $\varepsilon, \varepsilon' > 0$. Suppose $\tilde{f} : \mathcal{U}^* \to \mathbb{N}$ is a randomised function such that for all ε-unbounded DP mechanisms $\mathcal{M}$, it holds that $\mathcal{M}_{\mathcal{C}_{\tilde{f}}}$ is ε'-bounded DP. If $\alpha \geq 0$ is such that for every ε'-unbounded DP mechanism $\mathcal{A}$, there exists a dataset P such that $\mathbb{E}[|\mathcal{A}(P) - f(P)|^2] \geq \alpha$, then there exists a dataset P such that*

$$\mathbb{E}[|\tilde{f}(P) - f(P)|^2] \geq \alpha - \left(\frac{1}{\varepsilon}\right)^2.$$

The problem of lower bounding differentially private function estimation is well-studied [30, 4] in the privacy literature. The lower bounds essentially arise from the fact that $\mathcal{A}(P)$ and $\mathcal{A}(P')$ must be similar distributions for neighbouring databases, even if $f(P)$ and $f(P')$ are far apart. Since we know from Proposition 4 that $\tilde{f}(P)$ and $\tilde{f}(P')$ must also be close, we obtain the related lower bound. The slackness of $(1/\varepsilon)^2$ is a result of the fact that while $\mathcal{A}(P)$ and $\mathcal{A}(P')$ must be indistinguishable with respect to *any* hypothesis test, $\tilde{f}(P)$ and $\tilde{f}(P')$ need only be indistinguishable with respect to any ε-DP hypothesis test.

4.3 Privacy amplification from randomised rounding

Many functions used to determine data-dependent sampling rates have high sensitivity, but at least one common sampling method has low sensitivity: proportional sampling. In proportional sampling, a constant, data-independent fraction of the population is sampled independently from each stratum. This method is similar to simple random sampling, but a small amount of data dependence is introduced by the fact that the total number of samples across all strata must be an integer. In this section, we will show that while naïve implementations of proportional sampling can result in privacy degradation, a minor change in the sampling size function results in privacy amplification comparable to that afforded by simple random sampling.

Let $r \in [0, 1]$ and $f(P) = r|P|$ for some constant $r \in (0, 1)$. Since the output space of f is not $\mathbb{N}$, in practice, this is typically replaced with the deterministic function $\tilde{f}_{\text{det},r}(P) = \text{round}(r|P|)$, where $\text{round}(\cdot)$ rounds its input to the nearest integer. Unfortunately, deterministic rounding can be problematic for privacy, as we can see through a simple example. Suppose P and P' are neighbouring populations such that $|P| = 14, |P'| = 15$, and $r = 1/10$. Then, deterministic rounding always results in one case being sampled from P and two cases being sampled from P'. As discussed in Section 4.1, such a data-dependent deterministic function can never result in privacy amplification.

We propose a simple and practical change to the rounding process that *does* guarantee roughly the expected level of privacy amplification. We replace the ideal function f with a randomised rounding function $\tilde{f}_{\text{rand},r}$. That is, let $p = r|P| - \lfloor r|P| \rfloor$ so $\tilde{f}_{\text{rand},r}(P) = \lceil r|P| \rceil$ with probability p, and $\tilde{f}_{\text{rand},r}(P) = \lfloor r|P| \rfloor$ with probability $1 - p$. The following proposition shows that, up to a constant factor, randomised rounding recovers the expected factor of r in privacy amplification.

▶ **Theorem 6** (Privacy Amplification from Randomised Rounding)**.** *Let $r \in (0, 1)$. Then for every ε-unbounded DP mechanism $\mathcal{M}$, the mechanism $\mathcal{M}_{\mathcal{C}_{\tilde{f}_{rand,r}}}$ is ε'-unbounded DP when restricted to datasets of size at least $1/r$, where $\varepsilon' = \log\left(1 + 2r(e^{2\varepsilon} - 1)\right) + \log(1 + r(e^{2\varepsilon} - 1)) \approx 6r\varepsilon$.*

The approximation at the end of the proposition follows from applying (1) and (2), which give that $\log(1 + 2r(\exp(2\varepsilon) - 1)) \approx 2r \cdot 2\varepsilon$ and $\log(1 + r(\exp(2\varepsilon) - 1)) \approx r \cdot 2\varepsilon$. The constant 6 can perhaps be optimized through a more careful analysis. Randomised rounding is a practical modification since

it does not change the size of the sample very much; if traditional proportional allocation would typically assign m samples, then the modified algorithm allocates at most $m + 1$.

5 Cluster sampling

In cluster sampling, the population is partitioned into disjoint subsets, called clusters. A subset of the clusters is sampled and data subjects are selected from within the chosen clusters. If the sampling scheme uses a single stage design, all data subjects contained in the selected clusters will be included in the sample. Otherwise, a random sample of data subjects might be selected from each of the selected clusters (multi-stage design). Cluster sampling produces accurate results when the clusters are mutually homogeneous; that is, when the distributions within each cluster are similar to the distribution over the entire population.

In the survey context, cluster sampling is often performed due to time or budgetary constraints which make sampling many units from a few clusters cheaper and/or faster than sampling a few units from each cluster. A typical example in the survey context is when clusters are chosen to be geographic regions. Sampling a few geographic clusters and interviewing everybody in those clusters saves traveling costs compared to interviewing the same number of people based on a simple random sample from the population. In algorithm design, cluster sampling is often performed to improve the performance and accuracy of classifiers. In this setting, sampling often involves a two-step approach where the data is first clustered, using some clustering classifier, and then a subset of the clusters is selected. Forms of cluster samplings have been applied in several learning areas, for example in federated learning [16] and active learning [27].

5.1 Privacy implications of single-stage cluster sampling with simple random sampling

We focus here on a simple cluster sampling design that is commonly used in survey sampling and which naïvely appears to be a good candidate for privacy amplification: simple random sampling without replacement of clusters. That is, suppose the dataset P is divided into k clusters,

$$P = C_1 \sqcup \cdots \sqcup C_k$$

and the sampling mechanism $\mathcal{C}_\ell : \mathcal{U}^* \to \mathcal{U}^*$ chooses a random subset $I \subset [k]$ of size $\ell < k$, then maps P to $\sqcup_{i \in I} C_i$.

Since simple random sampling at the individual level provides good privacy amplification, one might expect the same to happen when the clusters are sampled in a similar way. In fact, this is true when the size of each cluster is small. However, if the clusters are large this sampling design achieves less amplification than might be expected. This is characterized by the following theorem showing a lower bound in this setting.

▶ **Theorem 7** (Lower Bound on Privacy Amplification for Cluster Sampling)**.** *For any sequence $n_i > 0$ and privacy parameter $\varepsilon > 0$, there exist neighboring populations $P = C_1 \sqcup \cdots C_i \sqcup \cdots \sqcup C_k$ and $P' = C_1 \sqcup \cdots C_i' \sqcup \cdots \sqcup C_k$ (with $|C_i| = n_i$ and $C_i' = C_i \cup \{x\}$ for some $x \in \mathcal{U}$) and an ε-unbounded DP mechanism $\mathcal{M}$ such that if $\mathcal{M}_{\mathcal{C}_\ell}(P)$ and $\mathcal{M}_{\mathcal{C}_\ell}(P')$ are ε'-indistinguishable then*

$$\varepsilon' \geq \ln\left(1 + \frac{\frac{\ell}{k}}{\left(\frac{\ell}{k} + \left(1 - \frac{\ell}{k}\right) e^{-(n_i + n_{\min})\varepsilon}\right)} (e^\varepsilon - 1)\right),$$

where $n_i = |C_i|$ and $n_{\min} = \min_{j \in \{1,\cdots,i-1\} \cup \{i+1,\cdots,k\}} n_j$.

We can compare the expression in the theorem above with the one we have for simple random sampling without replacement (cf. Theorem 14 from [6]):

$$\varepsilon' = \ln\left(1 + \frac{m}{n}(e^{\varepsilon} - 1)\right),$$

where m samples are drawn from a population of size n. We see that the two expressions coincide if $n_i + n_{\min} = 0$, which is an unrealistic corner case. Let us instead consider the case in which all the clusters are small. In this case, the quantity $n_i + n_{\min}$ will also be small, and if $\varepsilon < 1$, we can still expect some privacy amplification. However, as the clusters grow in size, the quantity $n_i + n_{\min}$ will also increase, and the lower bound converges very quickly to ε, giving essentially no amplification.

Next, we present a corresponding upper bound.

▶ **Theorem 8** (Upper Bound on Privacy Amplification for Cluster Sampling). *For any sequence $n_i > 0$, privacy parameter $\varepsilon > 0$, ε-unbounded DP mechanism $\mathcal{M} : \mathcal{U}^* \to \mathcal{O}$, and pair of neighboring populations P and P' such that $P = C_1 \sqcup \cdots C_i \sqcup \cdots \sqcup C_k$ and $P' = C_1 \sqcup \cdots C_i' \sqcup \cdots \sqcup C_k$ (with $|C_i| = n_i$ and $C_i' = C_i \cup \{x\}$ for some $x \in \mathcal{U}$), the mechanisms $\mathcal{M}_{\mathcal{C}_\ell}(P)$ and $\mathcal{M}_{\mathcal{C}_\ell}(P')$ are ε'-indistinguishable where*

$$\varepsilon' \leq \ln\left(1 + \frac{\frac{\ell}{k}}{\left(\frac{\ell}{k} + \left(1 - \frac{\ell}{k}\right) e^{-(n_i + n_{\max})\varepsilon}\right)}(e^{\varepsilon} - 1)\right),$$

and $n_{\max} = \max_{j \in \{1,\cdots,i-1\} \cup \{i+1,\cdots,k\}} n_j$,

Once again it is worth comparing the expression in the theorem above with the one we have for simple random sampling without replacement:

$$\varepsilon' = \ln\left(1 + \frac{m}{n}(e^{\varepsilon} - 1)\right).$$

Similar to the lower bound, the upper bound will quickly approach ε if the quantity $n_i + n_{\max}$ is large. If each cluster contains a single data point, the two bounds are close. This is not surprising since in this case the type of cluster sampling we considered is just simple random sampling without replacement. Note that while ℓ/k is the fraction of clusters included in the final sample and m/n is the fraction of data points, these are approximately the same when the clusters are small. If all the clusters are the same size, then $n_{\max} = n_{\min}$ and the upper and lower bounds we gave above match. The proofs of these results are contained in the Appendix.

5.2 Discussion and hypothesis testing

Privacy amplification by subsampling is often referred to as *secrecy of the sample* due to the intuition that the additional privacy arises from the fact that there is uncertainty regarding which user's data is in the sample. The key intuition then for Theorem 7 is that the larger the clusters are, the easier it is for a differentially private algorithm $\mathcal{M}$ to reverse engineer which clusters were sampled, breaking secrecy of the sample. Intuitively, if the clusters are different enough that a private algorithm can guess which clusters were chosen as part of the sample, then any amplification due to *secrecy of the sample* is negligible. We can formalize this intuition using once again using the lens of hypothesis testing. Note the framing in this section differs slightly from the framing in Section 4, although the underlying idea in both settings is that if a particular hypothesis test is effective, then there is a lower bound on the privacy parameter. In addition, note that privacy is also conserved in this setting, as $\mathcal{M}_{\mathcal{C}_\ell}$ is at least as private as $\mathcal{M}$. The question is: when is $\mathcal{M}_{\mathcal{C}_\ell}$ *more* private than $\mathcal{M}$?

▶ **Theorem 9.** *Let $\varepsilon > 0$, $\ell \in [0, k]$, $\mathcal{M} : \mathcal{U}^* \to \mathcal{O}$ be ε-DP and the sampling mechanism $\mathcal{C}_\ell$ be as defined in Section 5.1. Suppose there exists a hypothesis test $\mathcal{H} : \mathcal{O} \to \{0, 1\}$ such that*

$$\Pr(\mathcal{H}(\mathcal{M}_{\mathcal{C}_\ell}(P)) = 0 \mid C_i \in \mathcal{C}_\ell(P)) \geq e^{\varepsilon'} \Pr(\mathcal{H}(\mathcal{M}_{\mathcal{C}_\ell}(P)) = 0 \mid C_i \notin \mathcal{C}_\ell(P)).$$

Then there exists an event E in the output space of $\mathcal{M}$ such that for any neighboring population P' that differs from P in C_i, if

$$\varepsilon'' = \log \frac{\Pr(\mathcal{M}_{\mathcal{C}_\ell}(P) \in E | C_i \in \mathcal{C}_\ell(P))}{\Pr(\mathcal{M}(\mathcal{C}_\ell(P')) \in E | C_i \in \mathcal{C}_\ell(P'))} \in [0, \varepsilon],$$

and $\mathcal{M}_{\mathcal{C}_\ell}(P)$ and $\mathcal{M}_{\mathcal{C}_\ell}(P')$ are $\tilde{\varepsilon}$-indistinguishable, then

$$\tilde{\varepsilon} \geq \log\left(1 + (e^{\varepsilon''} - 1)\frac{\ell/k}{\ell/k + e^{-\varepsilon'}(1 - \ell/k)}\right).$$

The key take-away of this theorem is that for any ε-DP mechanism $\mathcal{M}$, if there exists a hypothesis test that, when given the output of $\mathcal{M}_{\mathcal{C}_\ell}(P)$, can confidently decide whether cluster C_i was chosen as part of the final sample, then the privacy guarantee of $\mathcal{M}_{\mathcal{C}_\ell}$ is no better than the privacy guarantee would be if we knew for certain that C_i was chosen as part of the sample. That is, in this setting, we gain no additional privacy as a result of secrecy of the sample. The parameter ε' controls how well the hypothesis test can determine whether $C_i \in \mathcal{C}_\ell$. As ε' increases, $\tilde{\varepsilon}$ approaches ε'', the privacy parameter if C_i is known to be part of the sample, so privacy amplification is negligible.

This view is consistent with Theorem 7. Consider a population where only data points in cluster i have a particular property and let $\mathcal{M}$ is an ε-DP mechanism that attempts to count how many data points with the property are in the final sample. If cluster i is large, then it is easy to determine from the output of the mechanism whether C_i is in the final sample. This example required cluster i to be distinguishable from the remaining clusters using a private algorithm. While examples as extreme as the one above may be uncommon in practice, clusters being different enough for a private algorithm to distinguish between them is not an unrealistic assumption.

In Section 5.1, we analysed a single stage design. All subjects contained in the selected clusters were included in the sample. In practice, multi-stage designs are common, where a random sample of subjects are selected from within each chosen cluster. If the sampling within each cluster is sufficiently simple then the privacy amplification from this stage can be immediately incorporated into the upper bound in Theorem 8. For example, if each subject within the chosen clusters is sampled with probability r and $\mathcal{M}$ is ε-DP, i.e., we perform Poisson sampling with probability r, then we immediately obtain an upper bound that is approximately $r\varepsilon$. However, if the sampling within clusters is more complex, then further analysis is required. One can also imagine more complicated schemes for selecting the chosen clusters. If these designs depend on properties of the data, then they are likely to result in privacy degradation. We leave this study for future work.

6 Stratified sampling

Finally, we turn our attention to another common sampling design: stratified sampling. In stratified sampling, the data is partitioned into disjoint subsets, called strata. A subset of data points is then sampled from each stratum to ensure the final sample contains data points from every stratum. Stratified sampling is common in survey sampling where it is used to improve accuracy and to ensure sufficient representation of sub-populations of interest. A classic use case of stratified sampling is business surveys, where businesses are typically stratified by industry and number of employees, or by similar measures of establishment size. Stratification by establishment size results in substantial gains in accuracy compared to simple random sampling, while stratification by industry ensures

that reliable estimates can be obtained at the industry level. Stratified sampling has several other applications; for example it is used in algorithm design to improve performance [2, 24], in private query design and optimization to improve accuracy [8], and to improve search and optimizations [25].

We focus here on *one-stage stratified sampling* using simple random sampling without replacement within each stratum to select samples. We also assume that the stratum boundaries have been fixed in advance. Given a target sample size m, the only design choice in this model is the *allocation function*, which determines how many samples to take from each stratum. Different allocation functions are used in practice. Which method is selected depends on the goals to be achieved (for example, ensuring constant sampling rates across strata or minimizing the variance for a statistic of interest).

Before we describe allocation functions in detail, let us establish some notation for stratified sampling. Suppose there are k strata in the population, and that each data point is a pair (s, x) where $s \in [k]$ denotes which stratum the data subject belongs to, and $x \in \mathcal{U}$ denotes their data. Let $\boldsymbol{f} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k) : ([k] \times \mathcal{U})^* \to \mathbb{N}^k$ denote the allocation rule, so $\boldsymbol{f}_i(P)$ samples are drawn uniformly at random without replacement from the ith stratum, $P_i = \{(s, x) \in P \mid s = i\}$. The final sample S is the union of the samples from all the strata.

An important feature of stratified sampling is that the sampling rates can vary between the strata. This means that data subjects in strata with low sampling rates may expect a higher level of privacy than data subjects in strata with high sampling rates. This leads us to define a variant of differential privacy that allows the privacy guarantee to vary between the strata. This generalisation of differential privacy is tailored to stratified datasets and allows us to state more refined privacy guarantees than the standard definition is capable of.

▶ **Definition 10.** *Let $k \in \mathbb{N}$ and suppose there are k strata. A mechanism $\mathcal{A}$ satisfies $(\varepsilon_1, \cdots, \varepsilon_k)$-stratified bounded differential privacy if for all datasets P, data points (s, x) and (s', x'), $\mathcal{A}(P \cup \{(s, x)\})$ and $\mathcal{A}(P \cup \{(s', x')\})$ are $\max\{\varepsilon_s, \varepsilon_{s'}\}$-indistinguishable. The mechanism $\mathcal{A}$ satisfies $(\varepsilon_1, \cdots, \varepsilon_k)$-stratified unbounded differential privacy if for all datasets P, data points (s, x), $\mathcal{A}(P)$ and $\mathcal{A}(P \cup \{(s, x)\})$ are ε_s-indistinguishable.*

This definition is an adaptation of personalized differential privacy [22, 14, 3]. Note that it protects not only the *value* of an individual's data point, but also which stratum they belong to.

6.1 Optimal allocation with privacy constraints

In this section, we will discuss how to think about choosing an allocation function when privacy is a concern. A common goal when choosing an allocation $\boldsymbol{f}$ is to minimise the variance of a particular statistic. That is, suppose that $\mathcal{C}_{\boldsymbol{f}}$ represents one-stage stratified sampling with allocation function $\boldsymbol{f}$. Then, given a population P and desired sample size m, the *optimal allocation function* $\boldsymbol{f}^*(P)$ with respect to a statistic θ is defined as

$$\boldsymbol{f}^*(P) = \arg\min_{\boldsymbol{f}} \operatorname{var}(\theta_{\mathcal{C}_{\boldsymbol{f}}}(P)), \tag{3}$$

where the randomness may come from both the allocation function and the sampling itself, and the minimum is over all allocation functions such that $\|\boldsymbol{f}(P)\|_1 \leq m$ for all P. [3]

A natural question then is: what is the optimal allocation when one wants to compute the statistic of interest differentially privately? This is a simple yet subtle question. Our results in the previous sections indicate that the landscapes of optimal allocations in the non-private and private settings

[3] As an aside, we note that the notion of optimal allocations implicitly assumes that the historic or auxiliary data, H, used to inform the sampling design and the population data P are the same, or at least similar enough that $\boldsymbol{f}^*(H)$ is a good proxy for $\boldsymbol{f}^*(P)$. This provides further justification for the assumption that $H = P$ in our statements.

may be very different. This is a result of the fact that allocation functions that do not amplify well typically need to add more noise to achieve privacy (see discussion in Section 2.4). The additional noise needed to achieve privacy may overwhelm any gains in accuracy for the non-private statistic. Additionally, it is not immediately obvious how to define the optimal allocation in the private setting.

In this section, we formulate the notion of an optimal allocation under privacy constraints. Our goal is to initiate the study of alternative allocation functions that may prove useful when privacy is a concern. A full investigation of this question is outside the scope of this paper, but we provide some intuition for why this may be an interesting and important question for future work.

Given a statistic θ, we wish to define the optimal allocation for estimating θ privately. Let $\tilde{\theta}^{\lambda}$ be an λ-DP algorithm for estimating θ, so $\tilde{\theta}^{\lambda}(P)$ is an approximation of $\theta(P)$. The smaller λ is, the noisier $\tilde{\theta}^{\lambda}$ is. The scale of λ needed to ensure that $\tilde{\theta}^{\lambda}_{\mathcal{C}_{\boldsymbol{f}}}$ is ε-DP depends on the allocation function $\boldsymbol{f}$. Allocation functions that are very sensitive to changes in the input dataset will require more noise (smaller λ) to mask changes in the allocation. For any allocation $\boldsymbol{f}$, we will define the optimal parameter λ as that which minimises the maximum variance of $\tilde{\theta}^{\lambda}_{\mathcal{C}_{\boldsymbol{f}}}$ over all datasets P, while maintaining privacy:

$$\begin{aligned}\lambda_{\boldsymbol{f}} =& \arg\min_{\lambda>0} \sup_{P} \frac{\text{var}(\tilde{\theta}^{\lambda}_{\mathcal{C}_{\boldsymbol{f}}}(P))}{\text{var}(\theta_{\mathcal{C}_{\boldsymbol{f}}}(P))} \\ & \text{s.t. } \tilde{\theta}^{\lambda}_{\mathcal{C}_{\boldsymbol{f}}} \text{ is } (\varepsilon_1, \cdots, \varepsilon_k)\text{-stratified DP.}\end{aligned} \tag{4}$$

Now, by definition, $\tilde{\theta}^{\lambda_{\boldsymbol{f}}}_{\mathcal{C}_{\boldsymbol{f}}}$ is $(\varepsilon_1, \cdots, \varepsilon_k)$-stratified DP for any allocation function $\boldsymbol{f}$. We minimise the multiplicative increase in variance so that the supremum is not dominated by populations P for which $\text{var}(\theta_{\mathcal{C}_{\boldsymbol{f}}}(P))$ is large. Given privacy parameters $\varepsilon_1, \cdots, \varepsilon_k \geq 0$, we now define the optimal allocation as the allocation function that minimises the maximum variance over all populations P:

$$\boldsymbol{f}^*_{\varepsilon} = \arg\min_{\boldsymbol{f}} \sup_{P} \text{var}(\tilde{\theta}^{\lambda_{\boldsymbol{f}}}_{\mathcal{C}_{\boldsymbol{f}}}(P)). \tag{5}$$

where the minimum again is over all allocations $\boldsymbol{f}$ such that $\|\boldsymbol{f}(P)\|_1 \leq m$ for all P, and the supremum is over all populations of interest. This optimisation function has a different form to Eqn 3, which performs the optimisation independently for each population P. This difference is necessary in the private setting as we need to ensure that the choice of allocation function $\boldsymbol{f}^*_{\varepsilon}$ is not data dependent, since this would introduce additional privacy concerns. We can view the optimal allocation as the optimal balancing between the variance of the non-private statistic, and the scale of the noise needed to maintain privacy.

We believe that examining the difference between the optimal allocation in the non-private setting (Eqn (3)) and in the private setting (Eqn (5)) is an important question for future work. The main challenge is computing the parameter $\lambda_{\boldsymbol{f}}$ for every allocation $\boldsymbol{f}$. Analysing the privacy implications of $\boldsymbol{f}$ in the style of the previous sections gives us an upper bound on $\lambda_{\boldsymbol{f}}$, although this bound may be loose for specific statistics $\tilde{\theta}^{\lambda}$. So, while the previous sections developed our intuition for $\lambda_{\boldsymbol{f}}$, we believe new techniques are required to understand this parameter enough to solve Eqn (5).

6.2 Challenges with optimal allocation

Optimal allocations are defined to perform well for a specific statistic of interest. However, in practice, a wide variety of analyses will be performed on the final sample. The chosen allocation function may be far from optimal for these other analyses. While this problem exists in the non-private setting, it becomes more acute in the private setting. An allocation function that is optimal for one statistic may result in privacy degradation (and hence low accuracy estimates) for another.

We illustrate this challenge using *Neyman allocation*, which is often employed for business surveys. Neyman allocation is the optimal allocation method for the weighted mean [26]:

$$\theta_\mu(S) = \frac{1}{|P|}\sum_{i=1}^{k}\frac{|P_i|}{|S_i|}\sum_{x\in S_i} x,$$

where $|P_i|$ is the size of stratum i, and $S_i = S \cap P_i$. The estimator $\theta_\mu(S)$ is an unbiased estimate of the population mean for any stratified sampling design. Given a desired sample size m, let $\boldsymbol{f}_{\text{Neyman}}$ be the allocation function corresponding to Neyman allocation. Provided each stratum is sufficiently large, $\boldsymbol{f}_{\text{Neyman}}(P) = (m_1, \cdots, m_k)$, where

$$m_i = \frac{|P_i|\sigma(P_i)}{\sum_{j=1}^{k}|P_j|\sigma(P_j)} \cdot m,$$

$\sigma^2(P_i)$ is the empirical variance in stratum i and sufficiently large means that $m_i \leq |P_i|$. Neyman allocation is deterministic and can be very sensitive to changes in the data due to its dependence on the variance within each stratum. So, while it can provide accurate results for some statistics, it provides very noisy results for other statistics of potential interest (e.g. privately computing strata sizes).

To demonstrate the sensitivity of Neyman allocation, we analysed the sensitivity on a real data set. The population is based on the County Business Patterns (CBP) data published by the U.S. Census Bureau [15].[4] Each data point is an establishment and the establishments are stratified by establishment size into $k = 12$ strata. With a target final sample size of $m = 10,000$, and using the weighted mean of the establishment size as the target statistic, the Neyman allocation for this population is $[1261, 621, 517, 1969, 833, 1947, 1058, 762, 257, 248, 306, 225]$. We can find a neighbouring population with Neyman allocation $[1259, 620, 516, 1965, 831, 1943, 1056, 761, 257, 247, 306, 244]$. While these allocations are not wildly different, they do differ by 19 samples in the top stratum, which might not have a large impact on the weighted mean, but could lead to more substantial changes for other statistics. As an illustrative example, we can consider the goal of privately estimating the stratum sizes in the sample, for which this allocation would lead to significant privacy degradation.

6.3 Privacy amplification from proportional sampling

Proportional sampling is an alternative allocation function that is used to provide equitable representation of each sub-population, or stratum. Given a desired sample size $m \in [n]$, proportional sampling samples an $r = \frac{m}{n}$ fraction of the data points (rounded to an integer) from each stratum. Proportional sampling is not an optimal allocation in the non-private setting but, when implemented with randomised rounding, it has good privacy amplification. Now that we consider stratified sampling with number of stratums $k \geq 1$, we can state the following generalisation of Theorem 6.

▶ **Theorem 11** (Privacy Amplification for Proportional Sampling). *Let $r \in [0, 1]$, $\varepsilon > 0$, $\mathcal{M}$ be an ε-DP mechanism, and $P = S_1 \sqcup \cdots \sqcup S_k$ and $P' = S'_1 \sqcup \cdots \sqcup S'_k$ be stratified neighboring datasets that differ on stratum i. If for all $j \in [k]$, $r|S_j| \geq 1$ and $r|S'_j| \geq 1$, then $\mathcal{M}_{\mathcal{C}_{\boldsymbol{f}_{r,\text{prop}}}}$ is ε'-DP where*

$$\varepsilon' \leq \log\left(1 + 2r(e^{2\varepsilon} - 1)\right) + \log(1 + r(e^{2\varepsilon} - 1)).$$

[4] The data released by the U.S. Census Bureau is a tabulated version of the true micro data from the Business Register (BR), a database of all known single and multi-establishment employer companies. The data set we use is micro data generated to be consistent with the tabulated version. Each data point in this population is the size of an establishment in the US. In order to compute the sensitivity, we need to top code the data, we top code the data at 10,000.

Note that given a private statistic $\tilde{\theta}^{\lambda}$ as defined as above, this allows us to set $\lambda_{\boldsymbol{f}_{r,\text{prop}}} \approx \frac{\varepsilon}{6r}$, which is considerably larger than ε for small sampling rates. Thus, while proportional sampling may not minimise the variance of any single statistic, it may be a good choice since it performs reasonably well for *all* statistics.

7 Conclusion

In this paper, we have considered the privacy guarantees of sampling schemes, extending previous results to more complex and data-dependent sampling designs that are commonly used in practice. We find that considering these sampling schemes requires developing more nuanced analytical tools. In this work, we characterize the privacy impacts of randomized and data-dependent sampling schemes. Then, we apply our insights to analyze cluster and stratified sampling and to consider the question of optimal allocations under privacy. To the best of our knowledge, this work is the first to initiate study into these designs. As such, we hope to see future work in three areas. First, future work should tighten and optimize the constants in our theorems. Second, our results should be extended from pure to approximate (and other variants) of differential privacy. Finally, we hope to see further investigation into near-optimal allocations under privacy constraints.

References

1. Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi, editors, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, pages 308–318. ACM, 2016. doi:10.1145/2976749.2978318.
2. Julaiti Alafate and Yoav S Freund. Faster boosting with smaller memory. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/3ffebb08d23c609875d7177ee769a3e9-Paper.pdf.
3. Mohammad Alaggan, Sébastien Gambs, and Anne-Marie Kermarrec. Heterogeneous differential privacy. *Journal of Privacy and Confidentiality*, 7, 04 2015.
4. Hilal Asi and John C. Duchi. Near instance-optimality in differential privacy, 2020. arXiv:2005.10630.
5. Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6280–6290, 2018.
6. Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), Jan. 2020.
7. Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA, 2014. IEEE Computer Society.
8. Johes Bater, Yongjoo Park, Xi He, Xiao Wang, and Jennie Rogers. SAQE: practical privacy-preserving approximate query processing for data federations. *Proc. VLDB Endow.*, 13(11):2691–2705, 2020. URL: http://www.vldb.org/pvldb/vol13/p2691-bater.pdf.
9. Amos Beimel, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. In Daniele Micciancio, editor, *Theory of Cryptography, 7th Theory of Cryptography Conference, TCC 2010, Zurich, Switzerland, February 9-11, 2010. Proceedings*, volume 5978 of *Lecture Notes in Computer Science*, pages 437–454. Springer, 2010. doi:10.1007/978-3-642-11799-2_26.

10 Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In Robert D. Kleinberg, editor, *Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013*, pages 97–110. ACM, 2013. doi:10.1145/2422436.2422450.

11 Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

12 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

13 Hamid Ebadi, Thibaud Antignac, and David Sands. Sampling and partitioning for differential privacy. In *14th Annual Conference on Privacy, Security and Trust, PST 2016, Auckland, New Zealand, December 12-14, 2016*, pages 664–673. IEEE, 2016. doi:10.1109/PST.2016.7906954.

14 Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. *SIGPLAN Not.*, 50(1):69–81, January 2015.

15 Fabian Eckert, Teresa C. Fort, Peter K. Schott, and Natalie J. Yang. Imputing missing values in the us census bureau's county business patterns. Technical report, National Bureau of Economic Research, 2021.

16 Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 3407–3416. PMLR, 2021. URL: http://proceedings.mlr.press/v139/fraboni21a.html.

17 Marco Gaboardi, Michael Hay, and Salil Vadhan. A programming framework for opendp. *Manuscript*, 2020.

18 Marco Gaboardi, James Honaker, Gary King, Jack Murtagh, Kobbi Nissim, Jonathan Ullman, and Salil Vadhan. Psi (ψ): A private data sharing interface. *arXiv preprint arXiv:1609.04340*, 2016.

19 Antonious M. Girgis, Deepesh Data, Suhas N. Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 2521–2529. PMLR, 2021. URL: http://proceedings.mlr.press/v130/girgis21a.html.

20 Jacob Imola and Kamalika Chaudhuri. Privacy amplification via bernoulli sampling. *arXiv preprint arXiv:2105.10594*, 2021.

21 Joonas Jälkö, Antti Honkela, and Onur Dikmen. Differentially private variational inference for non-conjugate models. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler, editors, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL: http://auai.org/uai2017/proceedings/papers/152.pdf.

22 Z. Jorgensen, T. Yu, and G. Cormode. Conservative or liberal? personalized differential privacy. In *2015 IEEE 31st International Conference on Data Engineering*, pages 1023–1034, 2015.

23 Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

24 Wouter Kool, Herke van Hoof, and Max Welling. Estimating gradients for discrete random variables by sampling without replacement. In *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=rklEj2EFvB.

25 Levi H. S. Lelis, Roni Stern, Shahab Jabbari Arfaee, Sandra Zilles, Ariel Felner, and Robert C. Holte. Predicting optimal solution costs with bidirectional stratified sampling in regular search spaces. *Artif. Intell.*, 230:51–73, 2016. doi:10.1016/j.artint.2015.09.012.

26 Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–606, 1934.

27 Hieu Tat Nguyen and Arnold W. M. Smeulders. Active learning using pre-clustering. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff,*

Alberta, Canada, July 4-8, 2004, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004. doi:10.1145/1015330.1015349.

28 Mijung Park, James R. Foulds, Kamalika Chaudhuri, and Max Welling. Variational bayes in private settings (VIPS). *J. Artif. Intell. Res.*, 68:109–157, 2020. doi:10.1613/jair.1.11763.

29 Adam Smith. Differential privacy and the secrecy of the sample, Feb 2010. URL: https://adamdsmith.wordpress.com/2009/09/02/sample-secrecy/.

30 Salil Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, chapter 7, pages 347–450. Springer International Publishing AG, Cham, Switzerland, 2017.

31 Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled renyi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1226–1235. PMLR, 2019. URL: http://proceedings.mlr.press/v89/wang19b.html.

32 Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2493–2502. JMLR.org, 2015. URL: http://proceedings.mlr.press/v37/wangg15.html.

33 Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle. *J. Mach. Learn. Res.*, 17:183:1–183:40, 2016.

34 Yu-Xiang Wang, Jing Lei, and Stephen E. Fienberg. A minimax theory for adaptive data analysis. *arXiv preprint arXiv:1602.04287*, 2016.

35 Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

A Basic facts about indistinguishability

▶ **Definition 12.** *Let the* LCS distance *between two data sets P and P', denoted $d_{LCS}(P, P')$, be the minimal k such that if we let $P = P_0$ and $P' = P_k$, there exist data sets $P_1, P_2, \cdots, P_{k-1}$ where for all $i = 0, \cdots, k-1$, P_i and P_{i+1} are unbounded neighbors.*

▶ **Lemma 13.** *[12] Let X, Y and Z be random variables. For any $\varepsilon, \varepsilon' > 0$, if X and Y are ε-indistinguishable, and Y and Z are ε'-indistinguishable, then X and Z are $\varepsilon + \varepsilon'$-indistinguishable.*

Many of our proofs use couplings so let us briefly describe on the main method we will use to construct a coupling of two random variables. Let X be a random variable taking values in Ω_X and Y be a random variable taking values in Ω_Y. Suppose there exists a (possibly randomised) transformation $f : \Omega_X \to \Omega_Y$ such that $Y = f(X)$. That is, for all $y \in \Omega_Y$,

$$\Pr(Y = y) = \sum_{x \in \Omega_X} \Pr(X = x) \Pr(f(x) = y).$$

Then we can construct a coupling of X and Y by $\mu(x, y) = \Pr(X = x) \Pr(f(x) = y)$. A short calculation confirms that this defines a coupling. Further, notice that $\mu(x, y) \neq 0$ if and only $\Pr(f(x) = y) \neq 0$.

▶ **Lemma 14.** *Let X and Y be random variables taking values in $\mathcal{U}^*$ such that there exists a coupling μ such that if $\mu(x, y) \neq 0$ then the LCS distance between x and y is at most A. Then if $\mathcal{M}$ is ε-unbounded DP then $\mathcal{M}(X)$ and $\mathcal{M}(Y)$ are $A\varepsilon$-indistinguishable.*

Proof.

$$\begin{aligned}
\Pr(\mathcal{M}(X) \in E) &= \sum_{x \in \mathcal{U}^*} \Pr(X = x) \Pr(\mathcal{M}(x) \in E) \\
&= \sum_{x \in \mathcal{U}^*} \sum_{y \in \mathcal{U}^*} \mu(x, y) \Pr(\mathcal{M}(x) \in E) \\
&\leq \sum_{y \in \mathcal{U}^*} \sum_{x \in \mathcal{U}^*} \mu(x, y) e^{A\varepsilon} \Pr(\mathcal{M}(y) \in E) \\
&= e^{A\varepsilon} \sum_{y \in \mathcal{U}^*} \Pr(Y = y) \Pr(\mathcal{M}(y) \in E) \\
&= e^{A\varepsilon} \Pr(\mathcal{M}(Y) \in E).
\end{aligned}$$

◀

▶ **Lemma 15** (Advanced joint convexity, [6])**.** *Let X and Y be random variables satisfying $X = (1-q)X_0 + qX_1$ and $Y = (1-q)Y_0 + qY_1$ for some $q \in [0, 1]$ and random variables X_0, X_1, Y_0 and Y_1. If*

- *X_0 and Y_0 are ε-indistinguishable,*
- *X_1 and Y_1 are $\varepsilon + \varepsilon'$-indistinguishable and*
- *X_1 and Y_0 are $\varepsilon + \varepsilon'$-indistinguishable*
- *X_0 and Y_1 are $\varepsilon + \varepsilon'$-indistinguishable*

then X and Y are $\varepsilon + \log(1 + q(e^{\varepsilon'} - 1))$-indistinguishable.

Proof.

$$\begin{aligned}
\Pr(X \in E) &= (1-q)\Pr(X_0 \in E) + q\Pr(X_1 \in E) \\
&= (1-q)\Pr(X_0 \in E) + qe^{-\varepsilon'}\Pr(X_1 \in E) \\
&\qquad + q(1-e^{-\varepsilon'})(1-q)\Pr(X_1 \in E) + q(1-e^{-\varepsilon'})q\Pr(X_1 \in E) \\
&\le (1-q)e^{\varepsilon}\Pr(Y_0 \in E) + qe^{\varepsilon}\Pr(Y_1 \in E) \\
&\qquad + q(1-e^{-\varepsilon'})(1-q)e^{\varepsilon+\varepsilon'}\Pr(Y_0 \in E) + q(1-e^{-\varepsilon'})qe^{\varepsilon+\varepsilon'}\Pr(Y_1 \in E) \\
&= (e^{\varepsilon} + qe^{\varepsilon}(e^{\varepsilon'}-1))\Pr(Y \in E)
\end{aligned}$$

◀

B Randomized data-independent sampling

▶ **Lemma 16.** *Given $m \in \mathbb{N}$, define $\mathcal{C}_m : \mathcal{U}^* \to \mathcal{U}^m$ be defined as follows: given a dataset P, form a sample S by sampling m data points randomly without replacement from P, then $\mathcal{C}_m(P) = S$. Let P and P' be unbounded neighboring datasets and $m, m' \in \mathbb{N}$, then $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_{m'}}(P')$ are*

$$\left(\log\left(1 + \frac{m}{|P|+1}(e^{2\varepsilon}-1)\right) + |m-m'|\varepsilon\right) \text{ - indistinguishable.}$$

Proof. Let $P' = P \cup \{x\}$. First, let us focus on the case where $m' = m$. Now,

$$\begin{aligned}
\mathcal{M}_{\mathcal{C}_m}(P') &= \frac{\binom{|P|}{m}}{\binom{|P|+1}{m}}\mathcal{M}_{\mathcal{C}_m}(P) + \left(1 - \frac{\binom{|P|}{m}}{\binom{|P|+1}{m}}\right)\mathcal{M}(\mathcal{C}_m(P')|_{x\in S}) \\
&= \left(1 - \frac{m}{|P|+1}\right)\mathcal{M}_{\mathcal{C}_m}(P) + \frac{m}{|P|+1}\mathcal{M}(\mathcal{C}_m(P')|_{x\in S}),
\end{aligned}$$

where $\mathcal{C}_m(P')|_{x\in S}$ denotes the random variable $\mathcal{C}_m(P')$ conditioned on the event that $x \in S$. Now, we can define a coupling of $\mathcal{C}_m(P)$ and $\mathcal{C}_m(P')|_{x\in S}$ by first sampling S from $\mathcal{C}_m(P)$, then replacing a random element of S by x. This coupling has LCS distance at most 2, so by Lemma 14, $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}(\mathcal{C}_m(P')|_{x\in S}))$ are 2ε-indistinguishable. Thus, by Lemma 15, $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_m}(P')$ are

$$\log\left(1 + \frac{m}{|P|+1}(e^{2\varepsilon}-1)\right) \text{ -indistinguishable.}$$

Next, let us consider the case $|m-m'| = 1$ and $P = P'$. We can define a coupling of $\mathcal{C}_m(P)$ and $\mathcal{C}_{m'}(P)$ as follows: first sample S from $\mathcal{C}_m(P)$, then add a random element of $P \backslash S$ to S. This coupling has LCS distance at most 1, so by Lemma 14, $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_{m'}}(P)$ are ε-indistinguishable.

Finally, we'll use Lemma 13 to complete the proof. Note that $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_m}(P')$ are $\log\left(1 + \frac{m}{|P|+1}(e^{2\varepsilon}-1)\right)$-indistinguishable. Then there exist $m_1, \cdots, m_{\ell-1}$ such that if we set $m_0 = m$ and $m_{|m-m'|} = m'$ then for all i, $|m_i - m_{i-1}| \le 1$ and so $\mathcal{M}(\mathcal{C}_{m_{i-1}}(P'))$ and $\mathcal{M}(\mathcal{C}_{m_i}(P'))$ are ε-indistinguishable. Therefore, by Lemma 13, $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_{m'}}(P')$ are

$$\left(\log\left(1 + \frac{m}{|P|+1}(e^{2\varepsilon}-1)\right) + |m-m'|\varepsilon\right) \text{ - indistinguishable.}$$

◀

▶ **Definition 17** (log-Lipschitz functions). *A function* $q : [n] \to \mathbb{R}_{\geq 0}$ *is* ε-log-Lipschitz *if for all* $m \in \{0, 1, \ldots, n-1\}$,

$$|\log q(m) - \log q(m+1)| \leq \varepsilon$$

▶ **Lemma 18.** *Let* $w : [n] \to \mathbb{R}_{\geq 0}$ *be nondecreasing, and let* $p : [n] \to \mathbb{R}_{\geq 0}$ *be any function. Then,*

$$\max_{q:[n]\to\mathbb{R}_{\geq 0} \text{ is } \varepsilon\text{-log-Lipschitz}} \frac{\sum_{m=0}^{n} q(m)w(m)p(m)}{\sum_{m=0}^{n} q(m)p(m)} \leq \frac{\sum_{m=0}^{n} e^{\varepsilon m} w(m)p(m)}{\sum_{m=0}^{n} e^{\varepsilon m} p(m)}$$

Proof. We will show by induction on $k = 0, 1, \ldots, n$ that we can assume w.l.o.g. that the maximizer has the form $q(0) = 1, q(1) = e^{\varepsilon}, \ldots, q(k) = e^{\varepsilon k}$. This holds for $k = 0$ by simply normalizing. Then, assuming it holds for some $k > 0$, and given any ε-log-Lipschitz q such that $q(0) = 1, \ldots, q(k) = e^{\varepsilon k}$, let us define q' as follows.

$$q'(m) = \begin{cases} q(m) & \text{for } m = 0, 1, \ldots, k \\ \frac{e^{\varepsilon} q(k) q(m)}{q(k+1)} & \text{for } m = k+1, \ldots, n \end{cases}$$

By construction, q' is ε-log-Lipschitz. In particular, $q'(k+1) = e^{\varepsilon} q(k) = e^{(k+1)\varepsilon}$. In addition, since q is ε-log-Lipschitz, we have that

$$\frac{e^{\varepsilon} q(k)}{q(k+1)} \geq 1,$$

which means that $q'(k+1) \geq q(k+1)$.

Next, we use the inequality (a slight generalization of the mediant inequality) that for $a, b, c, d > 0$ and $t \geq 1$ such that $a/b \leq c/d$,

$$\frac{a+c}{b+d} \leq \frac{a+tc}{b+td}$$

Let $a = \sum_{m=0}^{k} q(m)w(m)p(m)$, $b = \sum_{m=0}^{k} q(m)p(m)$, $c = \sum_{k+1}^{n} q(m)w(m)p(m)$, and $d = \sum_{k+1}^{n} q(m)p(m)$, and $t = e^{\varepsilon} q(k)/q(k+1)$. By the non-decreasing property of w, we have that

$$a/b \leq w(k) \leq w(k+1) \leq c/d.$$

Therefore, the inequality above, and by definition of q', we have that

$$\begin{aligned} \frac{\sum_{m=0}^{n} q(m)w(m)p(m)}{\sum_{m=0}^{n} q(m)p(m)} &\leq \frac{\sum_{m=0}^{k} q(m)w(m)p(m) + e^{\varepsilon}(q(k)/q(k+1)) \cdot \sum_{m=k+1}^{n} q(m)w(m)p(m)}{\sum_{m=0}^{k} q(m)p(m) + e^{\varepsilon}(q(k)/q(k+1)) \cdot \sum_{m=k+1}^{n} q(m)p(m)} \\ &= \frac{\sum_{m=0}^{n} q'(m)w(m)p(m)}{\sum_{m=0}^{n} q'(m)p(m)}. \end{aligned}$$

So, by induction, we can assume that the maximizer has the form $q(0) = 1, q(1) = e^{\varepsilon}, \ldots, q(n) = e^{\varepsilon n}$, which completes the proof. ◀

Proof of Theorem 3. Let $\mathcal{C}_m : \mathcal{U}^* \to \mathcal{U}^m$ be the sampling scheme that given a dataset P, returns S where S is a uniformly random subset of P of size m (drawn without replacement). Let $y \in \mathcal{O}$ be

any outcome, and let $P \sim P'$ be neighboring datasets. Then, we have that

$$\begin{aligned}
\Pr[\mathcal{M}_{\mathcal{C}}(P) = y] &= \sum_{m=0}^{n} \Pr[\mathcal{M}_{\mathcal{C}_m}(P) = y] \cdot \Pr[|\mathcal{C}(P)| = m] \\
&\leq \sum_{m=0}^{n} \left(1 + \frac{m}{n}(e^{\varepsilon} - 1)\right) \cdot Pr[\mathcal{M}_{\mathcal{C}_m}(P') = y] \cdot t(m) \\
&\leq \frac{\sum_{m=0}^{n} \left(1 + \frac{m}{n}(e^{\varepsilon} - 1)\right) \cdot e^{\varepsilon m} \cdot t(m)}{\sum_{m=0}^{n} e^{\varepsilon m} t(m)} \cdot \sum_{m=0}^{n} \Pr[\mathcal{M}_{\mathcal{C}_m}(P') = y] \cdot t(m) \\
&= \left(1 + \frac{\mathbb{E}_{m \sim \tilde{t}}[m]}{n}(e^{\varepsilon} - 1)\right) \cdot \Pr[\mathcal{M}_{\mathcal{C}}(P') = y]
\end{aligned}$$

where the first inequality follows from Lemma 16. Then, note that $(1 + (m/n)(e^{\varepsilon} - 1))$ is non-decreasing, and that $\Pr[\mathcal{M}_{\mathcal{C}_m}(P')) = y]$ is ε-log-Lipschitz by definition, so the second inequality follows by Lemma 18. After rearranging and simplifying, we obtain the desired result.

Finally, for the lower bound, suppose the data universe $\mathcal{U} = [0, 1]$. Let $P = \{1, \cdots, 1\}$ consist of n 1s and P' be the neighboring dataset $P' = P \backslash \{1\} \cup \{0\}$. Let $\mathcal{M} : \mathcal{U}^* \to \mathbb{R}$ be defined by $\mathcal{M}(S) = \sum_{x \in S} \mathbb{1}\{x = 1\} + \text{Lap}\,(1/\varepsilon)$ so $\mathcal{M}$ is ε-unbounded DP. Then

$$\begin{aligned}
\frac{\Pr(\mathcal{M}_{\mathcal{C}}(P') = n)}{\Pr(\mathcal{M}_{\mathcal{C}}(P) = n)} &= \frac{\sum_{m=0}^{n} \Pr(t = m) \left(\frac{m}{n} e^{-(n-m+1)\varepsilon} + \left(1 - \frac{m}{n}\right) e^{-(n-m)\varepsilon}\right)}{\sum_{m=0}^{n} \Pr(t = m) e^{-(n-m)\varepsilon}} \\
&= \frac{\sum_{m=0}^{n} \Pr(t = m) \left(\frac{m}{n} e^{(m-1)\varepsilon} + \left(1 - \frac{m}{n}\right) e^{m\varepsilon}\right)}{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon}} \\
&= \frac{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon} \left(1 - \frac{m}{n}(1 - e^{-\varepsilon})\right)}{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon}} \\
&= 1 - \frac{1}{n}(1 - e^{-\varepsilon}) \frac{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon} m}{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon}}.
\end{aligned}$$

Thus, taking the reciprocal,

$$\log \frac{\Pr(\mathcal{M}_{\mathcal{C}}(P) = n)}{\Pr(\mathcal{M}_{\mathcal{C}}(P') = n)} = -\log \left(1 - \frac{1}{n}(1 - e^{-\varepsilon}) \frac{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon} m}{\sum_{m=0}^{n} \Pr(t = m) e^{m\varepsilon}}\right).$$

◀

C Data-dependent sampling

Proof of Proposition 4: hypothesis testing perspective. Let $H : \mathbb{N} \to \{0, 1\}$ be the hypothesis test such that for all $x \in \mathbb{N}$, and $b \in \{0, 1\}$, $e^{-\varepsilon} \Pr(H(x) = b) \leq \Pr(H(x + 1) = b) \leq e^{\varepsilon} \Pr(H(x) = b)$. Then $H' : \mathcal{U}^* \to \{0, 1\}$ defined by $H'(S) = H(|S|)$ is ε-unbounded DP. By assumption, $H'_{\mathcal{C}_{\tilde{f}}}$ is ε'-DP. This implies that $H(\tilde{f}(P))$ and $H(\tilde{f}(P'))$ are ε'-indistinguishable. Therefore,

$$\text{adv}(H) = \Pr[H(\tilde{f}(P)) = 0] - \Pr[H(\tilde{f}(P')) = 0] \leq \Pr[H(\tilde{f}(P')) = 0](e^{\varepsilon'} - 1) \leq e^{\varepsilon'} - 1.$$

The result follows from taking the supremum over all ε-DP H. ◀

Proof of Theorem 6: proportional allocation with randomized rounding. Let P be a dataset, x be a data point and $P' = P \cup \{x\}$. Let $m = r|P|$, $m' = r|P'|$, $m^L = \lfloor m \rfloor$, $m'^L = \lfloor m' \rfloor$, $p = m - m^L$ and $p' = m' - m'^L$. Now, $m' - m = r < 1$ so we have two cases, $m^L = m'^L$ or $m^L = m'^L - 1$.

As in Lemma 16, let $\mathcal{C}_m : \mathcal{U}^* \to \mathcal{U}^m$ be the sampling scheme that given a dataset P, returns S where S is a uniformly random subset of P of size m (drawn without replacement). Note that by Theorem 2, for $m, m' \in \mathbb{N}$, $\mathcal{M}_m(P)$ and $\mathcal{M}_{m'}(P)$ are $|m-m'|\varepsilon$-indistinguishable, and $\mathcal{M}_{\mathcal{C}_m}(P)$ and $\mathcal{M}_{\mathcal{C}_{m'}}(P')$ are $\log\left(1+\frac{m}{|P|+1}(e^{2\varepsilon}-1)\right)+|m-m'|\varepsilon$-indistinguishable.

Firstly, suppose $m^L = m'^L$. Let

$$\mu_0 = \frac{1}{1-r}((1-p-r)\mathcal{M}_{\mathcal{C}_{m^L}}(P) + p\mathcal{M}_{\mathcal{C}_{m^L+1}}(P)),$$
$$\mu_0' = \frac{1}{1-r}((1-p-r)\mathcal{M}_{\mathcal{C}_{m^L}}(P') + p\mathcal{M}_{\mathcal{C}_{m^L+1}}(P')),$$
$$\mu_1 = \mathcal{M}_{\mathcal{C}_{m^L}}(P),$$
$$\mu_1' = \mathcal{M}_{\mathcal{C}_{m^L+1}}(P').$$

Notice that

$$\mathcal{M}_{\mathcal{C}_r}(P) = (1-r)\mu_0 + r\mu_1 \quad \text{and} \quad \mathcal{M}_{\mathcal{C}_r}(P') = (1-r)\mu_0' + r\mu_1'.$$

Now, by Lemma 15 and Lemma 14, μ_0 and μ_0' are $\log(1+\frac{m^L+1}{|P|+1}(e^{2\varepsilon}-1))$-indistinguishable. Further, all the pairs (μ_0', μ_1), (μ_1, μ_1') and (μ_0, μ_1') are $\left(\log(1+\frac{m^L+1}{|P|+1}(e^{2\varepsilon}-1))+\varepsilon\right)$-indistinguishable. Therefore, by Lemma 15, $\mathcal{M}_{\mathcal{C}_r}(P)$ and $\mathcal{M}_{\mathcal{C}_r}(P')$ are ε'-indistinguishable where

$$\begin{aligned}\varepsilon' &\le \log\left(1+\frac{m^L+1}{|P|+1}(e^{2\varepsilon}-1)\right) + \log(1+r(e^{\varepsilon}-1)) \\ &\le \log\left(1+\left(r+\frac{1}{|P|+1}\right)(e^{2\varepsilon}-1)\right) + \log(1+r(e^{\varepsilon}-1)).\end{aligned}$$

Next, suppose $m'^L = m^L + 1$. Let $1-q = \min\{p, 1-p'\}$ and

$$\mu_0 = \mathcal{M}_{\mathcal{C}_{m^L+1}}(P),$$
$$\mu_0' = \mathcal{M}_{\mathcal{C}_{m^L+1}}(P'),$$
$$\mu_1 = \frac{1}{q}((p-1+q)\mathcal{M}_{\mathcal{C}_{m^L+1}}(P) + (1-p)\mathcal{M}_{\mathcal{C}_{m^L}}(P),)$$
$$\mu_1' = \frac{1}{q}((1-p'-1+q)\mathcal{M}_{\mathcal{C}_{m^L+1}}(P') + p'\mathcal{M}_{\mathcal{C}_{m^L+2}}(P')).$$

Notice that

$$\mathcal{M}_{\mathcal{C}_r}(P) = (1-q)\mu_0 + q\mu_1 \quad \text{and} \quad \mathcal{M}_{\mathcal{C}_r}(P') = (1-q)\mu_0' + q\mu_1'.$$

Now, by Lemma 2, μ_0 and μ_0' are $\log\left(1+\frac{m^L+1}{|P|+1}\right)$-indistinguishable. Further, all the pairs (μ_0', μ_1), (μ_1, μ_1') and (μ_0, μ_1') are $\left(\log(1+\frac{m^L+1}{|P|+1}(e^{2\varepsilon}-1))+2\varepsilon\right)$-indistinguishable. Also, note that $q \le r$. Then by Lemma 15, $\mathcal{M}_{\mathcal{C}_r}(P)$ and $\mathcal{M}_{\mathcal{C}_r}(P')$ are ε'-indistinguishable where

$$\begin{aligned}\varepsilon' &\le \log\left(1+\frac{m^L+1}{|P|+1}(e^{2\varepsilon}-1)\right) + \log(1+p(e^{2\varepsilon}-1)) \\ &\le \log\left(1+\left(r+\frac{1}{|P|+1}\right)(e^{2\varepsilon}-1)\right) + \log(1+r(e^{2\varepsilon}-1)).\end{aligned}$$

◀

D Cluster sampling

Proof of Theorem 8. Without loss of generality, let $i = 1$. Notice that conditioned on cluster $1 \notin I$, the distribution of outputs of $\mathcal{M}_{\mathcal{C}}(P)$ and $\mathcal{M}_{\mathcal{C}}(P')$ are identical. Let E be a set of outcomes. Then

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E) &= \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E \mid 1 \in I) + \left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E \mid 1 \notin I) \\ &= \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E \mid 1 \in I) + \left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I).\end{aligned}$$

Now, we have that

$$\begin{aligned}\frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E \mid 1 \in I) &= \frac{\ell}{k}\sum_{|I|=\ell, 1\in I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E) \\ &\leq \frac{\ell}{k}\sum_{|I|=\ell, 1\in I}\frac{1}{\binom{k}{\ell}}e^{\varepsilon}\Pr(\mathcal{M}(P'_I) \in E) \\ &= \frac{\ell}{k}e^{\varepsilon}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I),\end{aligned}$$

where the inequality follows from the fact that the LCS distance between P_I and P'_I is 1. Thus,

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E) &\leq \frac{\ell}{k}e^{\varepsilon}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) + \left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I) \\ &= \Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) + \frac{\ell}{k}(e^{\varepsilon} - 1)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I).\end{aligned}$$

Now, we need to relate $\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I)$ to $\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E)$. For a set I such that $1 \notin I$ and index $i \in I$, let $I \cup \{1\}\backslash\{i\}$ be the set where index i has been replaced with 1. Then,

$$\begin{aligned}\left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I) &= \sum_{|I|=\ell, 1\notin I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E) \\ &= \sum_{|I|=\ell, 1\notin I}\sum_{i\in I}\frac{1}{\ell}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E) \\ &\geq \sum_{|I|=\ell, 1\notin I}\sum_{i\in I}\frac{1}{\ell}\frac{1}{\binom{k}{\ell}}e^{-(n_1+n_i)\varepsilon}\Pr(\mathcal{M}(P_{I\cup\{1\}\backslash\{i\}}) \in E) \\ &\geq e^{-(n_1+n_{\max})\varepsilon}\frac{1}{\ell}\sum_{|I|=\ell, 1\notin I}\sum_{i\in I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_{I\cup\{1\}\backslash\{i\}}) \in E),\end{aligned}$$

where the first inequality follows from the fact that the LCS distance between P_I and $P_{I\cup\{1\}\backslash\{i\}}$ is at most $n_1 + n_i$. Now, notice that the sets $I \cup \{1\}\backslash\{i\}$ in the above sum all contain 1, and each index I' such that $|I'| = \ell$ and $1 \in I'$ appears in the sum $k - \ell$ times (corresponding to the $k - \ell$ possible choices for the swapped index i). Therefore, we can rewrite the sum as

$$\begin{aligned}\left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I) &\geq e^{-(n_1+n_{\max})\varepsilon}\frac{k-\ell}{\ell}\sum_{|I|=\ell, 1\in I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E) \\ &= e^{-(n_1+n_{\max})\varepsilon}\frac{k-\ell}{\ell}\frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) \\ &= e^{-(n_1+n_{\max})\varepsilon}\left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I).\end{aligned}$$

Thus,

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) &= \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) + \left(1-\frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I)\\ &\geq \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) + \left(1-\frac{\ell}{k}\right)e^{-(n_1+n_{\max})\varepsilon}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I)\\ &= \left(\frac{\ell}{k} + \left(1-\frac{\ell}{k}\right)e^{-(n_1+n_{\max})\varepsilon}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I).\end{aligned}$$

Finally,

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E) &\leq \Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) + \frac{\ell}{k}(e^{\varepsilon}-1)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I)\\ &\leq \Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) + \frac{\ell}{k}(e^{\varepsilon}-1)\frac{1}{\left(\frac{\ell}{k} + \left(1-\frac{\ell}{k}\right)e^{-(n_1+n_{\max})\varepsilon}\right)}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E)\\ &\leq \left(1 + \frac{\ell}{k}(e^{\varepsilon}-1)\frac{1}{\left(\frac{\ell}{k} + \left(1-\frac{\ell}{k}\right)e^{-(n_1+n_{\max})\varepsilon}\right)}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E)\end{aligned}$$

◀

Now we turn our attention to the lower bound.

Proof of Theorem7. Let $C_1 = \{1, \cdots, 1\}$ and $C_j = \{-1, \cdots, -1\}$ for all $j \in \{2, \cdots, k\}$. Let $C_1' = C_1 \backslash \{1\} \cup \{-1\}$ be the same as C_1 except with one 1 switched to a -1. Let $\mathcal{M}(S) = \sum_{x \in S} x + \text{Lap}(1/\varepsilon)$, so $\mathcal{M}$ is ε-unbounded DP. Notice that $\mathcal{M}$ has the property that if $\sum_{x \in S'} x = \sum_{x \in S} x + a$, for some $a \in \mathbb{R}$ then $\Pr(\mathcal{M}(S) = \sum_{x \in S} x) = e^{|a|\varepsilon}\Pr(\mathcal{M}(S') = \sum_{x \in S} x)$. This equality allows us to tighten many of the inequalities that appeared in the proof of Theorem 8, and give a lower bound.

$$\begin{aligned}\Pr\left(\mathcal{M}_{\mathcal{C}}(P) = n_1 + 1\right) &= \frac{\ell}{k}\Pr\left(\mathcal{M}_{\mathcal{C}}(P) = n_1 + 1 \mid 1 \in I\right) + \left(1-\frac{\ell}{k}\right)\Pr\left(\mathcal{M}_{\mathcal{C}}(P) = n_1 + 1 \mid 1 \notin I\right)\\ &= \frac{\ell}{k}e^{\varepsilon}\Pr\left(\mathcal{M}_{\mathcal{C}}(P') = n_1 + 1 \mid 1 \in I\right) + \left(1-\frac{\ell}{k}\right)\Pr\left(\mathcal{M}_{\mathcal{C}}(P') = n_1 + 1 \mid 1 \notin I\right)\\ &= \Pr\left(\mathcal{M}_{\mathcal{C}}(P') = n_1 + 1\right) + \frac{\ell}{k}(e^{\varepsilon}-1)\Pr\left(\mathcal{M}_{\mathcal{C}}(P') = n_1 + 1 \mid 1 \in I\right).\end{aligned}$$

Now,

$$\begin{aligned}\left(1-\frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I) &= \sum_{|I|=\ell, 1 \notin I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E)\\ &= \sum_{|I|=\ell, 1 \notin I}\sum_{i \in I}\frac{1}{\ell}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E)\\ &= \sum_{|I|=\ell, 1 \notin I}\sum_{i \in I}\frac{1}{\ell}\frac{1}{\binom{k}{\ell}}e^{-(n_1+n_i)\varepsilon}\Pr(\mathcal{M}(P_{I \cup \{1\} \backslash \{i\}}) \in E)\\ &\leq e^{-(n_1+n_{\min})\varepsilon}\frac{1}{\ell}\sum_{|I|=\ell, 1 \notin I}\sum_{i \in I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_{I \cup \{1\} \backslash \{i\}}) \in E)\\ &= e^{-(n_1+n_{\max})\varepsilon}\frac{k-\ell}{\ell}\sum_{|I|=\ell, 1 \in I}\frac{1}{\binom{k}{\ell}}\Pr(\mathcal{M}(P_I) \in E)\\ &= e^{-(n_1+n_{\max})\varepsilon}\left(1-\frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I).\end{aligned}$$

Thus,

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) &= \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) + \left(1 - \frac{\ell}{k}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \notin I) \\ &\leq \frac{\ell}{k}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) + \left(1 - \frac{\ell}{k}\right) e^{-(n_1+n_{\min})\varepsilon}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) \\ &= \left(\frac{\ell}{k} + \left(1 - \frac{\ell}{k}\right) e^{-(n_1+n_{\min})\varepsilon}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I).\end{aligned}$$

Finally,

$$\begin{aligned}\Pr(\mathcal{M}_{\mathcal{C}}(P) \in E) &= \Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) + \frac{\ell}{k}(e^{\varepsilon} - 1)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E \mid 1 \in I) \\ &\geq \Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) + \frac{\ell}{k}(e^{\varepsilon} - 1)\frac{1}{\left(\frac{\ell}{k} + \left(1 - \frac{\ell}{k}\right) e^{-(n_1+n_{\min})\varepsilon}\right)}\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E) \\ &= \left(1 + \frac{\ell}{k}(e^{\varepsilon} - 1)\frac{1}{\left(\frac{\ell}{k} + \left(1 - \frac{\ell}{k}\right) e^{-(n_1+n_{\min})\varepsilon}\right)}\right)\Pr(\mathcal{M}_{\mathcal{C}}(P') \in E).\end{aligned}$$

◀

Proof of Theorem 9. Let $D = \mathcal{C}_\ell(P)$ and $D' = \mathcal{C}_\ell(P')$. For an event $E \in \mathcal{O}$, define the probabilities p, q, p' and q' as follows.

$$\begin{aligned} p &= \Pr(\mathcal{M}(D) \in E | C_1 \in D) \qquad q = \Pr(\mathcal{M}(D) \in E | C_1 \notin D) \\ p' &= \Pr(\mathcal{M}(D') \in E | C_1 \in D') \qquad q' = \Pr(\mathcal{M}(D') \in E | C_1 \notin D') \end{aligned}$$

By the existence of $\mathcal{H}$ described in the lemma statement, there must exist an event E such that $q \leq e^{-\varepsilon'} p$. Since P and P' only differ on C_1, the distributions of $\mathcal{M}(D)|_{C_1 \notin D}$ and $\mathcal{M}(D')|_{C_1 \notin D'}$ are identical, which means that $q = q'$. Then, we can compute a lower bound on the indistinguishability of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ as follows. Without loss of generality, assume $p' > p$

$$\begin{aligned}\frac{\Pr(\mathcal{M}(D') \in E)}{\Pr(\mathcal{M}(D) \in E)} &= \frac{p' \cdot Pr(C_1 \in D') + q \cdot Pr(C_1 \notin D')}{p \cdot Pr(C_1 \in D) + q \cdot Pr(C_1 \notin D)} \\ &= \frac{p' \cdot \frac{\ell}{k} + q \cdot (1 - \frac{\ell}{k})}{p \cdot \frac{\ell}{k} + q \cdot (1 - \frac{\ell}{k})} \\ &= \frac{p \cdot \frac{\ell}{k} + q \cdot (1 - \frac{\ell}{k}) + (p' - p) \cdot \frac{\ell}{k}}{p \cdot \frac{\ell}{k} + q \cdot (1 - \frac{\ell}{k})} \\ &= 1 + \frac{(p' - p)\frac{\ell}{k}}{p \cdot \frac{\ell}{k} + q \cdot (1 - \frac{\ell}{k})} \\ &\geq 1 + \frac{(p' - p)\frac{\ell}{k}}{p \cdot (\frac{\ell}{k} + e^{-\varepsilon'}(1 - \frac{\ell}{k}))} \\ &= 1 + \left(\frac{p'}{p} - 1\right)\frac{\frac{\ell}{k}}{\frac{\ell}{k} + e^{-\varepsilon'}(1 - \frac{\ell}{k})}\end{aligned}$$

where the final inequality follows from the fact that $\mathcal{M}$ is ϵ-DP, so $p'/p \geq e^{\epsilon''}$ by definition. ◀

E Stratified sampling

Proof of Theorem 11: proportional allocation for stratified sampling. Given $\mathcal{M} : ([k] \times \mathcal{U})^* \to \mathcal{Y}$, for all datasets $T_2, \cdots, T_k \in \mathcal{U}^*$, define $\mathcal{M}^{T_2,\cdots,T_k} : \mathcal{U}^* \to \mathcal{Y}$ by $\mathcal{M}^{T_2,\cdots,T_k}(S) =$

$\mathcal{M}(S \sqcup T_2 \sqcup \cdots \sqcup T_k)$. Then since $\mathcal{M}$ was $(\varepsilon, \cdots, \varepsilon)$-stratified unbounded DP, $\mathcal{M}^{T_2,\cdots,T_k}$ is ε-unbounded DP. Let $\mathcal{C}_r$ be as in Lemma 6 so for all S, S' unbounded neighbours such that $r|S| \geq 1$ and $r|S'| \geq 1$, $\mathcal{M}_{\mathcal{C}_r}^{T_2,\cdots,T_k}(S)$ and $\mathcal{M}_{\mathcal{C}_r}^{T_2,\cdots,T_k}(S')$ are ε'-indistinguishable where

$$\varepsilon' \leq \log\left(1 + 2r(e^{2\varepsilon} - 1)\right) + \log(1 + r(e^{2\varepsilon} - 1)).$$

Now, let $P = S_1 \sqcup S_2 \sqcup \cdots \sqcup S_k$ and $P = S_1' \sqcup S_2 \sqcup \cdots \sqcup S_k$ be unbounded stratified neighboring datasets that differ in the first stratum. Since $S_2 \sqcup \cdots \sqcup S_k$ are shared between P and P', and the datasets T_i only dependent on strata S_i, the distribution of $T_2, \cdots, T_k$ are identical given inputs P and P'. Let q be the distribution of $T_2, \cdots, T_k$ so $q(T_2, \cdots, T_k) = \Pr(\mathcal{C}_r(S_2) = T_2, \cdots, \mathcal{C}_r(S_k) = T_k)$. Then given an event E,

$$\begin{aligned}
\Pr(\mathcal{M}_{\mathcal{C}_{\boldsymbol{f}_{\text{prop},r}}}(P) \in E) &= \int_{T_2,\cdots,T_k} q(T_2, \cdots, T_k) \Pr(\mathcal{M}_{\mathcal{C}_r}^{T_2,\cdots,T_k}(S_1) \in E) \\
&\leq \int_{T_2,\cdots,T_k} q(T_2, \cdots, T_k) e^{\varepsilon'} \Pr(\mathcal{M}_{\mathcal{C}_r}^{T_2,\cdots,T_k}(S_1') \in E) \\
&= e^{\varepsilon'} \Pr(\mathcal{M}_{\mathcal{C}_{\boldsymbol{f}_{\text{prop},r}}}(P') \in E).
\end{aligned}$$

◀

F Instantiating the estimation-based lower bound

For any function f, the accuracy to which one can privately estimate f depends on the *sensitivity* of f, i.e., how much changing the input dataset can change the output.

Proof of Proposition 5. Define $\mathcal{M}_{SS} : \mathcal{U}^* \to \mathbb{N}$ as follows. For all $P \in \mathcal{U}^*$,

$$\mathcal{M}(P) = |P| + \text{Lap}(1/\varepsilon).$$

Then $\mathcal{M}$ is ε-unbounded DP. Suppose that $\tilde{f} : \mathcal{U}^* \to \mathbb{N}$ is such that for all ε-unbounded DP mechanisms $\mathcal{A}$, $\mathcal{A}_{\mathcal{C}_{\tilde{f}}}$ is ε'-bounded DP. This implies that

$$\mathcal{M}_{\mathcal{C}_{\tilde{f}}}(P)) = \tilde{f}(P) + \text{Lap}(1/\varepsilon)$$

is ε'-bounded DP. Therefore, by the definition of α, there exists a population P such that

$$\sup_{P \in \mathcal{U}^n} \mathbb{E}[|\mathcal{M}_{\mathcal{C}_{\tilde{f}}}(P)) - f(P)|^2] \geq \alpha.$$

Also

$$\begin{aligned}
\alpha \leq \mathbb{E}[|\mathcal{M}_{\mathcal{C}_{\tilde{f}}}(P)) - f(P)|^2] &= \mathbb{E}[|\tilde{f}(P) + \text{Lap}(1/\varepsilon) - f(P)|^2] \\
&= \mathbb{E}[(\tilde{f}(P) - f(P))^2 + \text{Lap}(1/\varepsilon)(\tilde{f}(P) - f(P)) + (\text{Lap}(1/\varepsilon))^2] \\
&= \mathbb{E}[]|\tilde{f}(P) - f(P)|^2] + (1/\varepsilon)^2.
\end{aligned}$$

After a small amount of rearranging we arrive at the result. ◀

G Experimental Setup

G.1 Figure 2

In this experiment, recall that we let $\mathcal{M}$ be an ε-DP exponential mechanism that returns a noisy count of the number of ones in a dataset from $\{0,1\}^n$, and $\Pr[\mathcal{M}(P) = i] \propto \exp\left(\varepsilon \cdot |i - \#\text{ones in } P|/2\right)$.

Then, we let t be the following randomised function which determines the size of the sample. For a given $\gamma \in (0, 1)$, we define t as follows.

$$t = \begin{cases} 2 & \text{w.p. } 1 - \gamma \\ n & \text{w.p. } \gamma \end{cases}$$

We define neighboring datasets P, P', where P has $n/2$ zeroes and $n/2$ ones, and P' has $n/2$ zeroes and $n/2 + 1$ ones. Let $\mathcal{C}_t(P)$ (resp. $\mathcal{C}_t(P')$) sample $m \sim t$ records randomly without replacement from P (resp. P'). In Figure 2, we vary γ from 10^{-25} and 10^{-15} and compute the empirical privacy loss ε' of $(\mathcal{M} \circ \mathcal{C}_t)(P)$ and $(\mathcal{M} \circ \mathcal{C}_t)(P')$.