Differential Privacy and its Application to Survey Data

Anne-Sophie Charest[1] and Jörg Drechsler[2,3,4]

[1] Université Laval, Canada, anne-sophie.charest@ulaval.ca
[2] Institute for Employment Research, Germany, joerg.drechsler@iab.de
[3] Ludwig-Maximilians-Universität (LMU), Munich, Germany
[4] University of Maryland, USA

Abstract

Differential privacy has emerged as a rigorous and broadly applicable framework for protecting confidential data, offering guarantees that do not depend on unverifiable assumptions. In this paper, we first present the definition of differential privacy and explain how it can be achieved in simple settings using standard mechanisms. We then examine the application of differential privacy to survey data and outline five key issues that complicate its use in this context.

Keywords: differential privacy, confidentiality, surveys, sampling, weighting, imputation.

1 Introduction

Confidentiality for survey and census data has long been a central concern. C*onfidentiality in this context means protecting participants' identities and data by keeping it private and secure, preventing unauthorized access, and only reporting results in aggregated forms to build trust and encourage honest answers, especially for sensitive topics. These goals are often achieved through a combination of techniques such as anonymization, encryption, strict access controls, and clear communication of data usage.* Focusing on anonymization, many techniques have been used to limit the risk of disclosing private information: data suppression, swapping, data perturbation, and, more recently, synthetic data (see for example Hundepool et al., 2012). These approaches aim to reduce disclosure risks but applying them effectively requires deciding when the risk is acceptable. A previous paper in this newsletter (Shlomo, 2022) reviewed traditional disclosure risks, namely identity, attribute, and inferential disclosure, and described how statisticians have estimated these risks over several decades. It also briefly introduced alternative privacy models proposed by computer scientists, such as differential privacy (DP).

DP provides guarantees that differ fundamentally from assumption-based risk estimates. Indeed, traditional disclosure risk metrics depend on unverifiable assumptions regarding the knowledge and capabilities of ill-intended users of the data, henceforth called attackers, that try to learn sensitive information about the units included in the data. Because of these assumptions traditional risk metrics can fail when new external data becomes available, whereas DP offers provable protections that do not rely on assumptions about the attacker's knowledge. Although the idea originated in computer science, an increasing number of statisticians are actively contributing to research in the field, motivated in part by the Census Bureau's adoption of DP to protect data from the 2020 U.S. Census (Abowd, 2018) and by the appeal of its clean theoretical guarantees.

However, the use of DP in the context of surveys is not straightforward. In fact, we will discuss in this paper five key issues that complicate this application. But, first, we present the DP framework in detail and outline a few methods to achieve this guarantee.

2 Differential Privacy

DP was first proposed by Dwork et al. (2006) and has been a very active research area since, particularly in the last ten years or so. The term now encompasses a broad family of definitions, such as (ε, δ)-DP, concentrated DP, and Rényi DP, each designed to address specific analytic or operational needs. We present the original definition in detail and provide references to a few important variants below. In the following section, we will explain how one can achieve DP, for example with the addition of carefully selected noise to a statistic of interest.

2.1 Pure DP

This is the original definition, now referred to as pure DP. It is important to understand that DP is the property of an algorithm, which takes as input a dataset to generate some output (for example, a statistic, a parameter estimate or even a synthetic dataset) and not the property of a specific output. This algorithm is usually called a randomized mechanism because satisfying the DP constraint generally requires the addition of randomness. This randomness plays a crucial role: it ensures that the mechanism's outputs cannot depend too heavily on any single individual's data. More precisely, a randomized mechanism M with output space S satisfies ε-differential privacy for a given privacy parameter $\varepsilon > 0$ if and only if for any two neighboring datasets D and D' and any $A \subseteq S$, we have that $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A]$.

To illustrate, we can look at a class of algorithms that achieve DP by adding independent discrete random noise to the statistic of interest, say the population total, i.e., instead of reporting the true total the algorithm would return a noisy total where the noise is chosen in such a way that the probability that the algorithm returns a specific value $t \in A$ if dataset *D* was used as the input is very close to the probability of returning the same value if dataset *D'* was used as the input. How close these two probabilities need to be is governed by the parameter ε.

Note that neighboring datasets can be defined in different ways, but the key idea is that they differ in the data of a single individual. For a classic tabular dataset where rows represent observations and columns represent variables, datasets D and D' are neighbors if they differ by exactly one row. More precisely, we talk about unbounded DP if D' is obtained by adding or removing one row from D, and bounded DP if D and D' have the same number of rows but differ in the values of one individual's record. There are subtle but important differences between these two definitions; see for example chapter 2 of Li et al. (2017). Other data structures require alternative notions of neighboring datasets. For example, in network data, one may define neighbors by removing or adding a single edge (Nissim et al., 2007) or a single node (Kasiviswanathan et al., 2013).

The guarantee offered by the pure DP definition can be interpreted in several ways. One is plausible deniability: an individual can claim that their data has any value, and the output of a DP mechanism cannot be used to refute that claim, even if an adversary holds as much information as the entire dataset except for that individual. This is because adding any row to this known dataset creates a neighboring dataset, and under DP the mechanism's output must be almost as likely under each of these possibilities. Consequently, no observer can reliably determine which specific values the individual contributed.

Another interpretation, given in Wasserman and Zhou (2010), is in the form of a hypothesis test. Pure DP implies a strict limit on how well any statistical test can distinguish whether the mechanism's output came from D or from D'. Specifically, for any test of level α, the power of that test must be smaller or equal to $e^{\varepsilon}\alpha$, so that the power of the test is very similar to its level. Thus, under ε-differential privacy with sufficiently small values of ε, even the most powerful test cannot reliably determine which of the two neighboring datasets produced the observed output, ensuring privacy for that individual.

Another important aspect of DP is the set of useful properties that follow directly from the definition. First, DP is immune to post-processing, meaning that any computation applied to the output of a DP mechanism will preserve the privacy guarantee. Second, DP composes in a straightforward way: when several DP mechanisms are applied to the same dataset, their privacy losses accumulate in a mathematically quantifiable manner, allowing to keep track of the privacy loss over multiple data releases. For example, k mechanisms that each satisfy ε-differential privacy jointly satisfy at most $k\varepsilon$-differential privacy. Because of this composition, the parameter ε is sometimes also referred to as a privacy budget; it defines the total amount of privacy leakage that is still considered acceptable. Based on the composition property one can then decide how this privacy budget should be spent across several outputs from a single dataset. More details on these properties can be found in Dwork and Roth (2014).

2.2 Approximate DP

Pure DP is a very strict guarantee, concerned with the worst-case scenario, because the inequality $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A]$ must hold for any possible D and D', even very implausible ones. A variant allows the guarantee not to hold when the probabilities of an output are small. More precisely, a randomized mechanism M is said to satisfy (ε, δ)-differential privacy with $\varepsilon > 0$ and $\delta > 0$ if for any two neighboring datasets D and D' and for any $A \subseteq S$ we have that $P[M(D) \in A] \leq e^{\varepsilon} P[M(D') \in A] + \delta$. This variant is more frequently used than pure DP, which is the special case where $\delta = 0$, and often is referred to as simply DP.

2.3 Other variants

Many other variants of differential privacy have been proposed over the years. These alternatives might modify the definition of neighbouring datasets or the way the privacy loss is measured. Desfontaines and Pejó (2020) surveys hundreds of such definitions inspired by DP. A few variants are worth mentioning: Rényi DP (Mironov, 2017), widely used in machine learning and in DP stochastic gradient descent, zero-concentrated DP (Bun and Steinke, 2016**)**, which offers tighter composition bounds and Gaussian DP, which offers an analytically tractable, hypothesis-testing–based framework with tight composition rules (Dong et al., 2022).

Another active line of work focuses on settings with no trusted curator, where privacy must be guaranteed at the user level, that is before the data is stored in a central database (see for example. Kasiviswanathan et al., 2011). This local differential privacy model is used in practice, for example, in large-scale telemetry systems (Apple, 2017).

3 Achieving DP

DP is typically achieved through the addition of randomness. There are a few building block mechanisms, which are often combined to obtain mechanisms for more complex tasks. These are described in Dwork and Roth (2014) and summarized below.

3.1 Noise addition for numeric outputs

Noise addition is the basic building block of many differentially private algorithms. Under pure DP, the standard approach is the Laplace mechanism. Suppose you want to release the output of some function f applied to a dataset D, the Laplace mechanism will add Laplace noise to the value of $f(D)$. The variance of the added noise depends on the global sensitivity of the function, which is the maximum possible change in the value of f when computed on any two neighboring datasets D and D', that is, the largest difference $|f(D) - f(D')|$ over all such pairs. More precisely, the Laplace mechanism for a function f releases $f(D) + X$, where X is drawn from a Laplace distribution with mean 0 and scale equal to the global sensitivity of f divided by ε. The sensitivity must be computed for each function f. For instance, the sensitivity of a counting query is 1, while the sensitivity of the

mean depends on the range of the possible values for the individual values. For a dataset for which the size n can be treated as public knowledge, each observation is in $[a, b]$, then the range is $R = b - a$ and the sensitivity of the mean is R/n. Note that if we cannot provide bounds for the individual values, then the sensitivity will be infinite, and thus it will not be possible to achieve DP. In practice, one may estimate the range from the observed data, but this will require spending some of the privacy budget.

For approximate DP, the standard mechanism is to add Gaussian noise, with variance determined by the global sensitivity of the function and the privacy parameters ϵ and δ. Other variants include adding geometric noise, or discrete or truncated noise distributions. Extensions also exist for multidimensional outputs, with mechanisms designed to handle vector-valued or high-dimensional functions.

3.2 Exponential mechanism for non-numeric outputs

Noise addition works well for numeric outputs, but many tasks require selecting from a set of categorical or structured outcomes. The exponential mechanism is a second fundamental building block that provides a general framework for releasing non-numeric outputs under differential privacy. It selects an output r with probability proportional to $\exp(\varepsilon u(D, r) \,/\, (2\Delta u))$, where $u(D, r)$ is a utility score for reporting r on dataset D and Δu is the sensitivity of this utility score. This mechanism is especially useful when the goal is to select the "best'' option according to a data-dependent criterion, such as choosing a model or a quantile, while ensuring that the choice does not reveal too much about the underlying data. For example, one may use the exponential mechanism to publish the mode of a categorical variable by using the number of observations in dataset D with value equal to r as utility function $u(D, r)$. This utility function has sensitivity $\Delta u = 1$.

3.3 More complicated mechanisms

Most DP mechanisms are constructed from these basic building blocks, together with the composition and post-processing properties of DP. For tasks such as regression, for example, one may add noise directly to the data, perturb the objective function, or add noise to the final output, or even decide to use more robust statistics to reduce the amount of noise required (see for example Alabi et al., 2022). The optimal strategy is problem-dependent, and in many settings remains an active area of research.

In machine learning, using differentially private versions of stochastic gradient descent (DP-SGD) has become the dominant approach for training models with differential privacy. The privacy guarantees of these algorithms rely on privacy accounting. Simple composition is far too loose when models are trained over tens of thousands of gradient steps. Privacy accounting methods provide tight bounds by exploiting subsampling amplification (Balle et al., 2018) and advanced composition frameworks such as Rényi DP (RDP), zero-concentrated DP (zCDP), and Gaussian DP (GDP). Without such accounting techniques, training would appear to consume impractically large privacy budgets, rendering DP-SGD useless in practice. Several accounting methods exist (Abadi et al., 2016; Bun and Steinke, 2016; Mironov, 2017; Dong et al., 2022; Koskela et al., 2020), each trading off accuracy, efficiency, and ease of implementation.

3.4 Practical challenges

Implementing DP in practice raises several important challenges. A first difficulty is choosing the privacy parameter ε. Although DP offers a formal privacy guarantee, it is really only meaningful if ε is relatively small. There is little consensus on what values are acceptable in applied contexts, and existing legal or regulatory frameworks such as the European Data Protection Regulation (GDPR), (Regulation (EU) 2016/679) offer only high-level guidance (Lee and Clifton, 2011).

A second challenge is that theoretical guarantees do not always translate cleanly to real-world implementations. Floating-point arithmetic, numerical clipping, and implementation-level randomness can all introduce deviations from the idealized model. Even subtle issues in pseudorandom number generators can weaken privacy guarantees, as demonstrated in early attacks on DP implementations (Mironov, 2012). Robust software engineering is therefore essential. Although mature libraries such as Google's Differential Privacy library (Google, 2019), OpenDP (OpenDP Project, 2021), and IBM's diffprivlib (Holohan et al., 2019) mitigate many of these risks, ensuring trustworthy and reproducible implementations remains an active area of work.

Finally, dependence within the data and adaptivity in the analysis process introduce additional complexities. Differential privacy is defined for datasets differing in one individual assuming independence between the units, but real datasets may exhibit strong correlations, for instance, between members of the same family, which can increase effective sensitivity and weaken protection (Kifer and Machanavajjhala, 2011). Adaptive analyses, where later queries depend on earlier outputs, also complicate privacy accounting. For example, model diagnostics such as residual checks or comparisons of fit statistics should be handled carefully, and obtaining them under DP will consume additional privacy budget (Dwork et al., 2015).

4 DP for surveys

The following discussions are excerpts from Drechsler and Bailie (2024) and we refer interested readers to this text for a more detailed discussion of DP in the survey context. When working with survey data, there are additional complexities which typically do not arise in other settings. Moreover, the implications of using DP in the context of surveys have received little attention in the DP literature until recently. Overall, there are (at least) five aspects that need to be considered when implementing DP in this context: (i) the multiple stages of the survey pipeline, (ii) limited privacy gains from complex sampling designs, (iii) challenges in computing the privacy guarantees of survey weighted estimates, and consequences of (iv) weighting adjustments and (v) imputations for missing data. We will discuss each of these aspects in the remainder of this section.

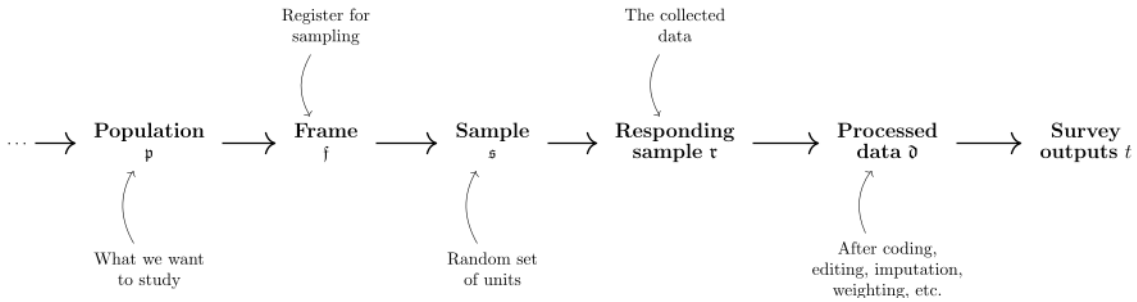

Figure 1: The main steps of the survey data pipeline

4.1 DP and the Survey Data Pipeline

As illustrated in Figure 1, the production of survey data is a complex multistage process. There are two important considerations when integrating a DP mechanism into a data pipeline. Firstly, at what point in the pipeline should the DP mechanism start? And secondly, which of the earlier stages of the data pipeline should be considered invariant? In the DP literature, invariants generally refer to aspects of the input data that remain fixed over neighboring dataset *D* and *D'*. For example, for the Decennial Census 2020, the U.S. Census Bureau decided that several counts must be released unaltered and thus treated them as invariant in their application of DP. With survey pipelines, there are several possible options with respect to the starting point of the mechanism and the decision on which of the earlier stages should be treated as invariant.

In the option most seen in the DP literature, the data-release mechanism starts at the end of the pipeline and performs just the last step – computing the survey outputs from the processed data – and none of the previous steps are taken as invariant. However, a mechanism could conceivably start at any point of the survey pipeline and incorporate all the steps that follow. Furthermore, any of the steps before the mechanism starts could conceivably be taken as invariant. Overall, this leads to up to 15 possible scenarios that need to be considered. Figure 2 highlights ten of these scenarios for illustration (the remaining five options would all start at the responding sample level).

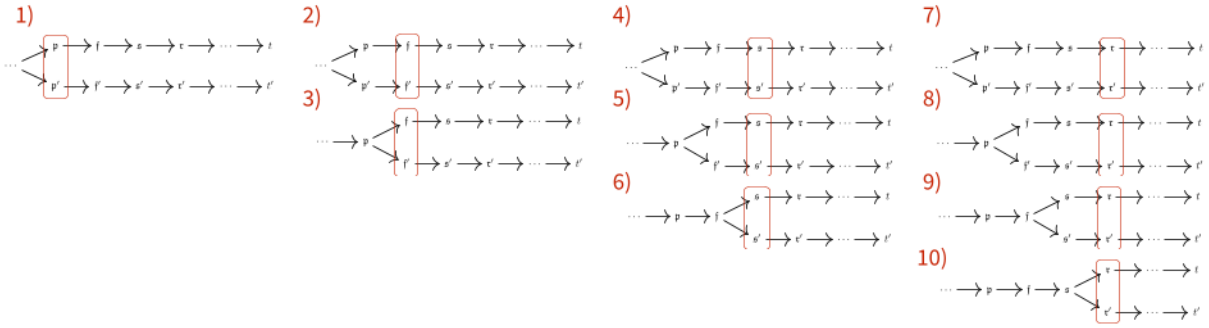

Figure 2: Ten out of fifteen possible settings for a DP mechanism in the survey context (the remaining five settings would all start at the level of the processed data). The red rectangles indicate the starting point of the mechanism.

A few general observations can be made regarding the advantages and disadvantages of the different scenarios (see Drechsler and Baile (2024) for a more detailed discussion): From a privacy perspective, it seems advantageous to start the DP mechanism as early as possible to benefit as much as possible from privacy amplification through subsampling (Balle et al., 2018), further discussed in Section 4.2 below. Note that the data production pipeline consists of three sampling steps: beyond the classical sampling step, nonresponse can be treated as another subsampling step and even the frame can be seen as a random sample from the target population, if we model the probability of inclusion in the frame as a random variable. However, this privacy amplification can be nullified if the attacker knows that the record they are attacking is in the sample, a scenario that statistical agencies often need to consider in practice and the additional privacy amplifications are either small or difficult to quantify (Bailie and Drechsler, 2024). On the other hand, any stage of the survey pipeline that should be part of the DP mechanism must first be fully "algorithmized" (that is, the process by which each of the stage's possible inputs is transformed into one of its outputs must be completely and programmatically specified). A survey pipeline often includes a number of complex, ill-defined and human-intensive tasks, such as building the frame, choosing a sampling design, coding and editing. Because these tasks all usually require a degree of human judgment, they would be difficult to algorithmize.

Another downside of starting the DP mechanism earlier is the fact that it can complicate the computation of the cumulative privacy loss across multiple data-release mechanisms because DP's composition theorems are not applicable when there is dependence between the mechanisms' noise terms (which can happen, for example, when their sampling designs are dependent or when two noisy statistics are computed from the same sample) (Bailie and Drechsler, 2024).

However, even if a data-release mechanism begins later in the survey pipeline so that some steps of the pipeline do not have to be incorporated in the mechanism, implementing DP still requires understanding those steps' effect on the mechanism's input data. For example, with hot deck imputation an individual survey respondent can contribute to multiple records in the post-imputation dataset. This complicates the appropriate definition of neighboring datasets: In the post-imputation dataset, changing a single record does not correspond to changing the data of one entity. In general,

the later the DP mechanism begins, the more difficult it is to determine an appropriate notion of neighbors since steps earlier in the pipeline may introduce dependencies between dataset records.

These complexities demonstrate that there can be conflicting demands in deciding where a DP mechanism should start within the survey pipeline. (See Drechsler and Bailie (2024) for further aspects that need to be considered.) We now return to the question of which steps of the survey pipeline should DP take as invariant. DP assesses the privacy of a data-release mechanism by comparing the survey outputs' distribution under pairs of counterfactual input datasets. By taking some of the steps of the survey pipeline as invariant, DP's counterfactual comparisons are reduced to only those pairs of input datasets which share the same realization of the invariant steps. For example, suppose the steps in the survey pipeline which generate the population and the frame are taken as invariant and the data-release mechanism starts with the responding sample. Then DP only compares those responding samples (i.e., those counterfactual input datasets) which could have come from the same frame. Adding invariants will weaken the privacy guarantees provided by DP (Kifer et al., 2022, Abowd et al., 2022). In general, the later the stage of the pipeline that is kept invariant, the greater the reduction in privacy. However, invariants may be justifiable when the output of the invariant steps can be considered public knowledge (such as if the frame was sourced commercially rather than constructed from confidential information). Moreover, constraining some steps to be invariant has the advantage of reducing the sensitivity of survey weighted estimators and thereby decreasing the noise which must be added for privacy protection as discussed in Section 4.3.

4.2 DP with Complex Sampling Designs

Statistical agencies have been aware for decades that sampling can be a simple and effective strategy to reduce disclosure risks simply because an attacker can no longer be sure whether a specific target record is included in the sample or not. This is the main reason why most statistical agencies only release samples from their censuses as public use micro datasets (they typically also apply additional measures to further increase the level of protection). This idea has been formalized in several papers in the context of DP (Kasiviswanathan et al., 2011, Wang et al., 2016, Bun et al., 2015, Balle et al., 2018, Wang et al., 2019). The authors show that the level of privacy is amplified through sampling, i.e., the actual privacy guarantees are higher than those implied by the chosen privacy loss parameters when protecting the sample output. Specifically, for small sampling rates *r* and small privacy loss parameters ε, applying certain simple sampling designs (simple random sampling with and without replacement, and Poisson sampling) before running an ε-DP mechanism reduces the privacy loss to approximately $r\varepsilon$. However, most surveys conducted by statistical agencies use complex multistage sampling designs, potentially with different sampling strategies at the different stages. Bun et al. (2022) study the amplification effects for complex designs and find that amplification is small for most of the sampling designs used in practice. Their findings can be summarized as follows:

- Cluster sampling using simple random sampling without replacement to draw the clusters offers negligible amplification in practice except for small ε and very small cluster sizes.
- With minor adjustments, stratified sampling using proportional allocation can provide privacy amplification. For small ε, the amplification is still linear in the sampling rate up to a constant factor.

- Data dependent allocation functions such as Neyman allocation for stratified sampling will likely result in privacy degradation. (The effects will depend on the sensitivity of the allocation function.)
- With PPS sampling at the individual level, the privacy amplification will linearly depend on the maximum probability of inclusion (for small ε).
- Systematic sampling will only offer amplification if the ordering of the population is truly random. In all other cases, systematic sampling will suffer from the same effects as cluster sampling, leading to no amplification (assuming the ordering is known to the attacker).

In practice this implies that for many multistage sampling designs, which typically start with (multiple stages of) stratified cluster sampling, amplification effects can generally only be expected from those stages at which individual units or households are selected (typically the last stage of selection).

4.3 DP for Weighted Estimates

As discussed in Section 3.1, the amount of noise that needs to be added to achieve a specific privacy loss ε directly depends on the sensitivity of the statistic of interest. From a utility perspective, this implies that more reliable (less noisy) DP outputs can be expected from statistics with low sensitivity.

When analyzing survey data, it is generally important to take the sampling design into account since the probabilities of selection vary between the units included in the sample. To simplify this task for data users, statistical agencies typically provide survey weights. In practice, these survey weights will also account for nonresponse and other data deficiencies such as undercoverage. (We will address this extra layer of complexity in the next section.)

Using survey weighted estimates raises the question: how (if at all) does the sensitivity of a statistic change when the survey design is taken into account? To illustrate the possible impacts, let us assume the analyst is interested in estimating the mean of some variable Y in the population using the sampled values y_i, $i = 1, \ldots, n$, where n denotes the sample size. If the probabilities of selection were equal for all units, the sample mean would be an unbiased estimate for the population mean and, as indicated in section 3.1, its sensitivity would be R/n, with R denoting the range of possible values for y_i.

When dealing with unequal probabilities of selection, a popular estimator for the population mean is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952): $\widehat{\mu_Y^{HT}} = \sum w_i y_i / N$ where w_i is the weight of unit i, for $i = 1, \ldots, n$ and N is the size of the population. Note that we assume for simplicity that N is known and does not need to be protected and w_i is the design weight, i.e., it only accounts for the sampling design.

If we can treat the weights as fixed, the sensitivity of $\widehat{\mu_Y^{HT}}$ is $\max(w_i)\, R/N$. Whether the maximum is over all units in the frame, over all units in the population, or over all possible counterfactual units, depends on which stages of the survey pipeline are treated as invariant as discussed in Subsection 4.1. Note that for equal-probability designs all $w_i = N/n$ and thus the sensitivity of the Horvitz-Thompson estimator is the same as for the unweighted estimator. If $\max(w_i) > N/n$, the Horvitz-Thompson estimator will have larger sensitivity than the unweighted estimator.

However, these discussions assume that the weights can be treated as fixed, that is, they do not change if a record changes in the database. For most sampling designs used in practice, such an assumption is unrealistic. For example, with sampling proportional to size (PPS), the i-th record's probability of inclusion is given by $\pi_i = (nx_i)/(N\overline{x})$, where x_i is the value for unit i of the measure-

of-size variable X that is used to improve the efficiency of the sampling design, and $\overline{x} = \sum_{i=1}^{N} x_i / N$ is the population mean of X. Changing the value of X for a single record will change the probabilities of inclusion and thus the survey weights for all other records in the sampling frame. Therefore, the sensitivity will be larger compared to the setting with fixed weights as we no longer only need to consider the maximum possible change in a single record's value for Y. We also need to consider the impact of the weight change for all the other records even if their values for Y don't change.

A recently proposed strategy to mitigate this potentially substantial increase in sensitivity is to regularize the weights, as explored by Seeman et al. (2024). (An extreme version of this strategy would set all weights to be equal; this could be justifiable if the increase in the privacy noise due to the weights dwarfs the bias introduced by ignoring the sampling design.) Another possible strategy is to treat the frame or at least the design variables within the frame as invariant as discussed in Figure 2. Frame invariance assumes any two neighboring datasets must always originate from the same frame and so can only differ at the sample level (or later). However, treating the frame as invariant has two additional implications that need to be considered. First, fixing the frame implies that privacy amplification from sampling is no longer possible (we would need to have neighboring datasets at the frame level in order to achieve amplification). However, given the results of Bun et al. (2022), this amplification is likely small in practice and thus the positive effects of reducing the sensitivity will tend to outweigh the negative effects of losing the amplification effect. On the other hand, fixing the frame will restrict the possible counterfactual input datasets to those which are consistent with the realized frame. Because this restriction will fix the survey weights, it might introduce strong constraints on the possible neighboring datasets, depending on the sampling design. As a consequence, the actual privacy guarantees for a frame invariant setting could be significantly weaker than the guarantees under a non-frame-invariant setting even for the same privacy loss parameter. How problematic this reduction in privacy is in real settings is currently an open question for research.

In contrast, if only the design variables are treated as fixed, the data-release mechanism could still start at the frame level, strengthening the privacy guarantees.

4.4 DP and Weighting Adjustments

In practice, two adjustment steps are commonly applied to the design weights to correct for unit nonresponse and other data deficiencies such as over- or undercoverage in the sampling frame: nonresponse adjustments and calibration. How these adjustment steps interfere with differential privacy has not been studied so far. However, both steps are data dependent, that is, they use information from the survey units for the adjustments. This implies that these steps cannot be ignored from a privacy perspective as the adjusted weights leak some personal information. Looking at the impacts on the sensitivity of the final statistic of interest (which uses the adjusted weights), similar problems as those discussed in the previous section will arise: changing one record in the database can potentially change the weight-adjustment factors for all other units in the survey. Thus, it seems imperative to already account for these adjustment steps during data pre-processing. Better results in terms of the privacy-accuracy trade-off might be achieved if the weight-adjustment steps were carried out in a differentially private way. More research is needed to better understand this trade-off. For example, it seems beneficial to identify robust adjustment strategies as less noise would be required to satisfy DP for these strategies.

In the particular case of post-stratification (which is a simple type of calibration), one such robust adjustment strategy has been proposed by Clifton et al. (2023). Another strategy would be to

regularize the nonresponse and calibration weight adjustments. (This would be similar to the survey weight regularization strategy of Seeman et al. (2024) discussed in the previous section.)

4.5 DP and Imputation

All survey data are plagued by item nonresponse as survey respondents are often unwilling or unable to respond to all survey questions especially if they request sensitive information. A common strategy to deal with this problem is to impute the missing values before analyzing the data to avoid biases that might arise when simply discarding incomplete records before the analysis. However, imputations are always data dependent as they typically build a model based on the observed data and use this model to impute the missing values. As a consequence, the implications of imputation on the DP guarantees need to be considered regardless of whether or not the imputation procedure is included inside the data-release mechanism. Some preliminary results for this problem are discussed in Das et al. (2022).

Similar to the problem of weighting adjustments, there are two possible strategies to account for imputation under DP. The first strategy only considers the effects when analyzing the imputed data. The second strategy modifies the imputation routines to ensure that the imputations already satisfy DP. As Das et al. (2022) have shown, the first strategy implies that in the worst case the sensitivity increases linearly with the number of imputed observations. This substantial increase of the sensitivity arises because changing one record in the database can potentially impact all of the imputed values. Whether the worst case applies depends on the analysis of interest and on the selected imputation procedure. Still, for statistical agencies offering pre-imputed datasets for accredited researchers, this strategy is not an option since they cannot anticipate which analyses might be performed on the imputed data.

The second strategy can break the dependence on the number of imputed records at least for certain imputation strategies. The key requirement for breaking the dependence is that the imputation model m can be written as $D_{imp}^{(i)} \sim m\left(D_{obs}^{(i)}, \hat{\theta}\right)$, where $D_{imp}^{(i)}$ and $D_{obs}^{(i)}$ contain the imputed and observed variables for record i and $\hat{\theta}$ denotes the model parameters estimated on the complete data. The model implies that, given $\hat{\theta}$, the imputed values of record i only depend on the observed values of that record and not on any other record. If these requirements are met and the parameters θ of the imputation model are estimated using any suitable differentially private mechanism with privacy loss parameter ε_1, then, given any ε_2 differentially private mechanism used for analyzing the data, the overall privacy loss is given by $\varepsilon_1 + \varepsilon_2$ by the composition property

We note that the conditional independence assumption of the imputation model holds for many imputation methods, for example, parametric imputation models based on linear regression. However, it does not hold for hot-deck imputation, an imputation method commonly applied at statistical agencies.

5. Discussion

Differential privacy provides a formal, elegant framework with strong theoretical guarantees and several appealing properties such as post-processing immunity, clean composition rules, and a clear interpretation of privacy loss. In simple settings, these guarantees are straightforward to compute and the required noise is easy to calibrate. However, real data-analysis workflows are rarely this tidy, and its application to survey data highlights just how complex differential privacy can become in practice. The presence of complex sampling designs, weighting adjustments, imputation steps, and data-dependent decisions means that calibrating privacy loss is rarely straightforward. Each of these operations can interact with DP in subtle ways, and determining how much noise is needed, or even

what the appropriate sensitivity should be, poses challenges that haven't been fully addressed in the literature.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L. (2016, October). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 308-318.

Abowd, J. M. (2018). The US Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2867-2867.

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022). The 2020 Census disclosure avoidance system TopDown Algorithm. In: Harvard Data Science Review (Special Issue 2).

Alabi, D., McMillan, A., Sarathy, J., Smith, A. and Vadhan, S. (2022). Differentially Private Simple Linear Regression. *Proceedings on Privacy Enhancing Technologies*.

Apple Differential Privacy Team. (2017). *Differential privacy technical overview*. Apple Inc. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Bailie, J. and Drechsler, J. (2024). Whose Data Is It Anyway? Towards a Formal Treatment of Differential Privacy for Surveys. In: National Bureau of Economic Research (2024): Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences, Spring 2024, Washington, 33 p. URL: https://conference.nber.org/conf_papers/f194306.pdf.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. In: Advances in Neural Information Processing Systems, 31.

Bun, M., and Steinke, T. (2016) Concentrated differential privacy: Simplifications, extensions, and lower bounds. In: Theory of cryptography conference. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.

Bun, M., Nissim, K., Stemmer, U., and Vadhan, S. (2015). Differentially private release and learning of threshold functions. In: Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science, FOCS '15, Washington, DC, USA, IEEE Computer Society, 634–649.

Bun, M., Drechsler, J., Gaboardi, M., McMillan, A., and Sarathy, J. (2022). Controlling privacy loss in sampling schemes: An analysis of stratified and cluster sampling. In: Foundations of Responsible Computing (FORC 2022), 24 p.

Clifton, C., Dajani, A. N., Hanson, E. J., Clark, S., Merrill, K., Merrill, S., and Rodriguez, R. (2023). Preliminary report on differentially private post-stratification. Working Paper ced-wp-2023-004, U.S. Census Bureau. URL: https://www.census.gov/library/working-papers/2023/adrm/ced-wp-2023-004.html

Das, S., Drechsler, J., Merrill, K., and Merrill, S. (2022). Imputation under differential privacy. Preprint: arXiv:2206.15063.

Desfontaines, D. and Pejó, B. (2020). SoK: Differential privacies. *Proceedings on Privacy Enhancing Technologies*.

Dong, J., Roth, A. and Su, W. J. (2022). Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *84*(1), 3-37.

Drechsler, J., and Bailie J. (2024). The Complexities of Differential Privacy for Survey Data. NBER working paper / National Bureau of Economic Research 32905. To appear in: R. Gong, V. J. Hotz, I. M. Schmutte (Eds.), Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and Their Consequences, 18 p. doi: 10.48550/arXiv.2408.07006.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, *9*(3–4), 211-407.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, *349*(6248), 636-638.

Google. (2019). Google differential privacy library [Computer software]. GitHub. https://github.com/google/differential-privacy

Holohan, N., Braghin, S., Mac Aonghusa, P. and Levacher, K. (2019). Diffprivlib: the IBM differential privacy library. arXiv preprint arXiv:1907.02444.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. In: Journal of the American Statistical Association, 47(260), 663–685.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and De Wolf, P. P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith. A. (2011). What can we learn privately? In: SIAM Journal on Computing, 40(3), 793–826.

Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S. and Smith, A. (2013). Analyzing graphs with node differential privacy. In *Theory of Cryptography Conference*, Berlin, Heidelberg: Springer, 457-476.

Kifer, D., Abowd, J. M., Ashmead, R., Cumings-Menon, R., Leclerc, P., Machanavajjhala, A., Sexton, W., and Zhuravlev, P. (2022). Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 Census. Technical Report: arXiv:2209.03310.

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 193-204.

Koskela, A., Jälkö, J. and Honkela, A. (2020). Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, PMLR, 2560-2569.

Lee, J. and Clifton, C. (2011). How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, Springer Berlin Heidelberg, 325-340.

Li, N., Lyu, M., Su, D. and Yang, W. (2017). *Differential privacy: From theory to practice*. Morgan & Claypool.

Mironov, I. (2012). On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, 650-661.

Mironov, I. (2017) Rényi differential privacy. In: Proceedings of the 30th IEEE computer security foundations symposium (CSF). IEEE, Santa Barbara, 263-275.

Nissim, K., Raskhodnikova, S. and Smith, A. (2007). Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 75-84.

OpenDP Project. (2021) The OpenDP Library (version 0.3) Harvard University. https://opendp.org

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4 May 2016, p. 1–88.

Seeman, J., Si, Y., and Reiter, J. P. (2024). Differentially private population quantity estimates via survey weight regularization. In: National Bureau of Economic Research (2024): Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences, Spring 2024, Washington, 33 p. URL: https://www.nber.org/system/files/chapters/c15023/c15023.pdf

Shlomo, N. (2022). How to measure disclosure risk in microdata?. *The Survey Statistician*, *86*(2), 13-21.

Wang, Y., Balle, B., and Kasiviswanathan, S. P. (2019). Subsampled R´enyi 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, PMLR, 89, 1226–1235. URL: http://proceedings.mlr.press/v89/wang19b.html.

Wang, Y.-X., Lei, J., and Fienberg, S. E. (2016). Learning with differential privacy: Stability, learnability, and the sufficiency and necessity of ERM principle. In: Journal of Machine Learning Research, 17(183),1–40.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, *105*(489), 375-389.