Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census

Daniel Kifer[1], John M. Abowd[2], Robert Ashmead[2], Ryan Cumings-Menon[2], Philip Leclerc[2], Ashwin Machanavajjhala[3], William Sexton[3], and Pavel Zhuravlev[2]

[1]U.S. Census Bureau and Penn State University
[2]U.S. Census Bureau
[3]Tumult Labs

September 8, 2022

Abstract

The purpose of this paper is to guide interpretation of the semantic privacy guarantees for some of the major variations of differential privacy, which include pure, approximate, Rényi, zero-concentrated, and f differential privacy. We interpret privacy-loss accounting parameters, frequentist semantics, and Bayesian semantics (including new results). The driving application is the interpretation of the confidentiality protections for the 2020 Census Public Law 94-171 Redistricting Data Summary File released August 12, 2021, which, for the first time, were produced with formal privacy guarantees.

1 Introduction

Differential privacy [29] is the formal privacy framework adopted for disclosure avoidance in the 2020 United States Decennial Census of Population and Housing. Officially invented in 2006, it allows for rigorous mathematical reasoning about confidentiality protections and utility of privacy-protected data. Initially, the term *differential privacy* referred to a single privacy definition [29] (now known as *pure* differential privacy) but it has evolved into a family of related privacy definitions, each with its own parameters and interpretations.

The purpose of this paper is to introduce, to a statistically-oriented audience, the major variations of differential privacy that are particularly relevant to the disclosure avoidance system for the 2020 Census. These are pure differential privacy [29], approximate differential privacy [28], Rényi differential privacy [51], zero-concentrated differential privacy [14], and f-differential privacy [24].[1] We focus primarily on the privacy semantics of these definitions and consider several different viewpoints:

- *Parameter interpretability*: Each privacy definition has its own set of privacy-loss parameters, such as the ϵ in pure differential privacy. Since privacy-loss parameters must be chosen by policymakers, it is important to gain a comprehensible, minimally technical understanding of the parameters that can be used in policy discussions. Thus, where possible, we discuss the interpretations of the different parameters. Not all parameters have minimally technical interpretations, however, and, in particular, the δ parameter of the approximate differential privacy framework is often misinterpreted. We also propose a more interpretable version of approximate differential privacy, which we call probabilistically-bounded differential privacy.

- *Frequentist hypothesis testing*: All variations of differential privacy covered in this paper are based on a concept called the *privacy-loss random variable* (PLRV). The PLRV has close connections to

[1]For other variations of differential privacy, the interested reader can consult the survey article by Desfontaines and Pejó [22].

the likelihood-ratio test and, as a consequence, all these definitions provide protections that can be framed as frequentist hypothesis tests. If an attacker is using the privacy-protected data to perform a hypothesis test related to the confidential information supplied by an individual, the relationship between the significance level and power of *any* such test is governed by the privacy-loss parameters of the variation's definitions.

- *Bayesian semantics:* There are fewer results related to Bayesian semantics owing to the mathematical complexity of analyzing Bayes' rule when differentially private randomization is combined with an attacker's prior information. We explain the Bayesian semantics of pure differential privacy and introduce new results for Rényi and zero-concentrated differential privacy. These results have close connections to the probabilistically-bounded differential privacy definition we introduce in this paper. Traditional Bayesian disclosure avoidance methodology compares what can be learned about a target individual based on a data release to what would have been known without the data release (also known as a *prior-to-posterior* comparison). This comparison is problematic because *anything* learned from a data release (including generalizable knowledge such as smoking causes cancer) is treated as a privacy violation. The literature on differential privacy examines the more nuanced question of what can be learned about an individual based on a data release compared to what could have been learned from the data release had the individual's information been scrubbed beforehand. This comparison is known as a *posterior-to-posterior* comparison and avoids the pitfalls of prior-to-posterior comparisons. Thus the Bayesian semantics we discuss are of the posterior-to-posterior variety.

When discussing these semantics, we also apply them to the production settings of the 2020 Census Disclosure Avoidance System (DAS). We show that different privacy semantics can be fine-grained and tailored to the types of inferences an attacker may try to make. This stands in contrast to the common mistaken belief that only one coarse-grained guarantee, summarized by a single number, is possible.

One might ask why this paper covers six variations of differential privacy. Isn't one enough? It turns out that if we want comprehensive privacy guarantees from the viewpoints of parameter interpretability as well as frequentist and Bayesian statistics, then all six variations are needed. Specifically, pure ϵ-differential privacy [29], the ancestor of all these variations, is the starting point. It provides the strongest privacy protections and is easy to analyze from both the frequentist and Bayesian points of view. In certain cases, it may over-estimate the privacy risks that individuals face even against extremely powerful attackers. Hence, approximate (ϵ, δ)-differential privacy was introduced to better track those privacy risks. Approximate differential privacy represents a continuum of privacy definitions – an (ϵ, δ)-curve [9, 24] – although many articles unfortunately consider just a single point on that curve (like trying to understand an elephant by studying a single hair on its tail). Approximate differential privacy is easy to study from the frequentist point of view [42], but Bayesian results are more limited [43]. Furthermore, the δ parameter is not directly interpretable and is often misinterpreted. To address this issue, we propose (ϵ, δ)-probabilistically-bounded differential privacy, which has more interpretable privacy-loss parameters. When it is treated as a continuum of privacy definitions (i.e., an (ϵ, δ)-curve), pbdp is a reparametrization of f-differential privacy (f-DP) [24] implying that f- DP is a natural consequence of addressing the difficulties in interpreting the δ parameter of approximate differential privacy. In addition to parameter interpretability, f-DP also has natural frequentist semantics because it was specifically developed to manage the trade-off between the level and power of re-identification inferences [24]. Except in special cases, the privacy parameters of f-DP (and probabilistically-bounded differential privacy) are difficult to compute, but Rényi and zero-concentrated differential privacy can be used to compute upper bounds on those parameters and on the frequentist semantics of f-DP and probabilistically-bounded differential privacy. Furthermore, we propose new Bayesian privacy semantics for Rényi and zero-concentrated differential privacy that, except in special cases, are difficult to obtain when working directly with f-DP and probabilistically-bounded differential privacy. Unfortunately, however, the privacy parameters of Rényi and zero-concentrated differential privacy are not easy to interpret. To summarize, parameter and frequentist interpretability is a strength of f-DP and pbdp while Bayesian semantic interpretability and ease of parameter computation are strengths of Rényi and zero-concentrated DP.

We note that there is a natural trade-off between privacy and utility in a data release. It is up to the data collector (e.g., statistical agency) to carefully balance the privacy and utility requirements of their applications. The utility side is dataset-specific, but privacy can be treated in a more general way. Thus,

this paper discusses various ways in which privacy protections can be quantified in order for them to be used in privacy/utility discussions.

This paper is organized as follows. In Section 2, we provide an accessible and informal introduction to what is and isn't considered a confidentiality breach in modern statistical disclosure limitation (SDL). Such a discussion is especially important given that many (peer-reviewed) articles about privacy and confidentiality are based on intuitive but faulty premises – faulty in the sense that the natural conclusions of those premises are contradictory. In Section 3, we introduce the notation used in this paper. In Section 4, we discuss the privacy-loss random variable (PLRV), which is the technical underpinning of all the privacy definitions we consider and has close connections to the likelihood ratio test. We also discuss the desirable technical properties that a "good" formal privacy definition should have. In Section 5, we explain all the privacy definitions, propose probabilistically-bounded differential privacy, and discuss the interpretability of each definition's parameters. Frequentist privacy semantics are discussed in Section 6 and Bayesian semantics are discussed in Section 7, including new semantics for Rényi and zero-concentrated differential privacy. Privacy analyses can be fine-grained, depending on the information an attacker may wish to infer about a person potentially represented in the confidential data. These fine-grained attacks, with applications to the 2020 Census, are discussed in Section 8. Finally, we discuss additional related work in Section 9 and conclude in Section 10.

2 Informal Introduction to Privacy Semantics

Privacy and confidentiality are two different concepts. Informally, privacy relates to a person's ability to control access to information about themselves and confidentiality refers to a third-party's (e.g., a statistical agency's) ability to prevent disclosure of information it has collected. However, in the computer science literature, the word "privacy" is frequently used to mean "confidentiality" (as in the case of differential privacy) and that is also the way the term privacy will be used in this paper.

With this caveat about terminology, we note that there is a common belief that a privacy breach occurs whenever a public dataset (derived from confidential data) is used to make unwanted or potentially harmful inferences about an individual. This belief is quite wrong, as it encompasses situations that are not privacy breaches. One important case is when the unwanted inference depends on generalizable (statistical) knowledge contained in the public dataset rather than on the specific datum that a target individual may or may not have contributed to the underlying confidential database. It is difficult to over-emphasize this distinction, and we take pains here to make it as clear as possible.

Thus, consider the canonical example of the first Cancer Prevention Study, also known as CPS-I [39, 6], discussed by Dwork and Pottenger [26]. This large-scale study, which followed a cohort of volunteers from 1952 to 1972 was the first study to conclusively establish the link between smoking and cancer. Based on the scientific knowledge gained in this study, if you see a Millenial chain smoker, you can be virtually certain that they have a much higher chance of developing lung cancer than if they never smoked. Such an inference is definitely unwanted, as it can result in higher health and life insurance premiums. However, for people born after 1972, like our Millenial, this study cannot possibly be considered a privacy breach because their data could not have been used in the study. For such people, one would say that the inference is purely statistical in nature.

Therefore, an unwanted inference is only a privacy breach if it is specifically *caused* by the inclusion of the individual's information in the dataset from which the inference was made. Consequently, an empirical privacy analysis needs to distinguish between the casual effect on inference due to a particular individual's data being used and the general statistical information provided by the dataset on a group of other individuals. A privacy-protection methodology must work to limit the causal inference about a particular individual (to protect privacy) while leaving intact statistical inferences due to information in the data, which is basically the same as leaving the data fit for statistical purposes [61].

To relate the causal inference interpretation of a privacy breech to concepts that have been used historically, consider the textbook definitions of identity and attribute disclosure. In traditional SDL, an *identity disclosure* occurs when "a data subject is identified from released data" [25, p. 174], an *attribute disclosure* occurs when "information [is disclosed] about a population unit without (necessarily) the identification of the unit within the data set" [25, p. 172], and an *inferential disclosure* occurs when either an identity or

attribute disclosure can be inferred with high probability [18]. The traditional SDL literature has not been careful to distinguish between identity and attribute inferences that depend upon the use of the respondent's data and those that are possible without using the individual's information. The confidentiality-breaching versions of identity, attribute and inferential disclosure all have the same mathematical formulation, which we now illustrate.

The differential privacy framework for SDL methods is designed precisely for the purpose of distinguishing between confidentiality breaches and valid scientific inferences. Typically, one reasons about causality by studying *interventions* [55], which are direct manipulations of a system. In the case of differential privacy, the intervention is to replace an individual's data record with something else and to reason about the effect of this intervention [61].

More concretely, suppose Jessie is a 52-year-old Hispanic Asian woman and let her survey response (e.g., to the decennial census) be denoted by r. Let $\mathcal{D}$ be the collection of responses of everyone else, so that $\mathcal{D} \cup \{r\}$ is the complete dataset (Jessie's record combined with everyone else's records). Let M be an algorithm that applies SDL protections while producing an output ω. Mathematically, $\omega = M(\mathcal{D} \cup \{r\})$. Both ω and the source code of the mechanism M are then made public. One may wonder what does ω (and knowledge of M) reveal about Jessie, and how much of this is due to their record being part of the input to M.

Following the causal view [61], one might ask what would have happened in an alternate world in which the data collector (e.g., the Census Bureau) replaced Jessie's submitted record r with some other fixed, pre-selected record r^* (e.g., a 60-year-old White Non-Hispanic male) before running M. In such a world, the output would be denoted as ω^*. Mathematically, $\omega^* = M(\mathcal{D} \cup \{r^*\})$. Clearly, the specific contents of the record r that Jessie submitted can have no causal effect on ω^* since r was never used. Therefore, ω^* has no causal effect on inferences about Jessie's recorded data. All inferences about Jessie based on ω^* are statistical inferences (e.g., based on the demographic composition of Jessie's community), not privacy-breaching inferences about how Jessie differs from their community.

One could then compare inference about Jessie in the actual world, where the output of $M(\mathcal{D} \cup \{r\})$ is observed, to inference about Jessie in her *privacy-preserving counterfactual world*, in which $M(\mathcal{D} \cup \{r^*\})$ is observed. The way this difference is measured is at the heart of the different commonly-accepted variations of differential privacy [29, 28, 14, 51, 24].

We briefly note that the degree to which an inference is unwanted (i.e., potentially harmful to the particular individual or entity) is also an important consideration. For example, a healthy individual may have fewer concerns about revealing some medical data than an individual who is not. It is possible to add specifications about which types of inference are unwanted in formal privacy definitions (e.g, [47, 41, 48, 35]). However, for simplicity of the explanations, in this document we consider the case where all such inferences are considered potentially sensitive, which is consistent with the statutory framework regulating the Census Bureau and other statistical agencies in the United States (13 U.S. Code §§ 8(b) & 9, 44 U.S. Code §§ 3561(11) & 3563). This also allows individuals to retroactively change their mind about which inferences they would consider harmful.

Now, comparing inferences about Jessie when her record is used (i.e., based on $M(\mathcal{D} \cup \{r\})$ to when her record is not (i.e., $M(\mathcal{D} \cup \{r^*\})$) is also tricky. To see why, suppose M simply outputs the number of 52-year-old females in the input data and suppose, with Jessie's record, there would be 523 of them (i.e., $M(\mathcal{D} \cup \{r\}) = 523$). In Jessie's privacy-preserving counterfactual world, the answer would be 522 (i.e., $M(\mathcal{D} \cup \{r^*\}) = 522$). How does inference about Jessie's age compare to the inference in the counterfactual world? In the counterfactual world, nothing is revealed about Jessie's age. However, in the actual world, it is not clear what the answer 523 reveals – it all depends on what an attacker knows, what additional information is available to the attacker, and how the attacker chooses to perform the statistical inference. There is no general agreement about reasonable choices here – in any room of 10 economists, statisticians, social scientists, or demographers, there would be at least 23 different, and often mutually contradictory, suggestions.

Differential privacy provides a clever way of sidestepping the question of what does 523 reveal about Jessie compared to 522. The main idea is to force M to be randomized: sometimes $M(\mathcal{D} \cup \{r\})$ will produce 523, sometimes it will produce 522, and other times it will produce other numbers. Similarly, sometimes $M(\mathcal{D} \cup \{r^*\})$ will produce 523, sometimes it will produce 522, etc. Thus instead of reasoning about 522 vs. 523, differential privacy reasons about how likely you are to see 523 if the input is $\mathcal{D} \cup \{r\}$ compared to when the input is $\mathcal{D} \cup \{r^*\}$. If the output distribution, when $\mathcal{D} \cup \{r\}$ is the input, is exactly the same as when

$\mathcal{D} \cup \{r^*\}$ is the input, then clearly nothing is revealed as the result of using Jessie's record. If the output distributions are "slightly" different, then it is likely that nothing "statistically meaningful" is revealed as the result of using Jessie's record. This "statistical meaningfulness" is formalized as follows: given an output ω, how well can a statistical procedure determine whether the input was $\mathcal{D} \cup \{r\}$ or $\mathcal{D} \cup \{r^*\}$? The choice of how to measure differences between distributions (to determine whether they are "slightly" different) is what creates the different variations of differential privacy.

This change in perspective – from reasoning about specific outputs to reasoning about output distributions – allows differential privacy to have some appealing hypothesis testing and Bayesian inference interpretations that do not require any assumptions about the knowledge of an attacker. Explaining these guarantees and filling in missing pieces, such as the Bayesian semantics of concentrated differential privacy, are the goals of this paper.

3 Notation

In this section, we summarize the notation used in this paper.

Let $\mathcal{D}$ be a collection of n records (e.g., census or sample survey responses) from individuals. For notational convenience, we assume that all the data an individual provides are collected in one record so that there is a one-to-one correspondence between records and individuals in the data. We let r denote a single record – the data on one individual.

A function that can be computed over a dataset is called a query and denoted as q. One example of a query is the total population in each county in the United States. Note that this is a vector-valued query with one component for each county. We refer to the different components of a vector-valued query using subscripts a follows: $q(\mathcal{D})_i$.

An algorithm or piece of software that applies SDL protections when producing an output is called a *mechanism* and denoted by M. A mechanism can be randomized, so that the output may be different each time it is run. We let $\omega = M(\mathcal{D})$ denote the output of mechanism M, given $\mathcal{D}$ as input. Often we need to refer to a secondary analysis performed on the output of M. We use the notation A to denote a (possibly randomized) algorithm. Secondary analysis of the output of M is denoted as $A(\omega)$ or $A(M(\mathcal{D}))$. The algorithm that applies M to the data and then applies A to the output is denoted by $A \circ M$; that is, $(A \circ M)(\mathcal{D}) \equiv A(M(\mathcal{D}))$.

When an attacker is making inferences about a target individual, we refer to the target individual as Jessie. The dataset with Jessie's record removed is denoted by $\mathcal{D}^-$.

The following symbols are reserved for the parameters of the various privacy definitions we discuss: $\epsilon, \delta, \rho, \alpha, \gamma$. For this reason, when discussing hypothesis tests, we use the following symbols for Type I error probability/significance level (ℓ), Type II error probability (β), and power ($1 - \beta$).

For Bayesian analysis, we let $P_{\mathfrak{A}}$ denote an attacker's prior over datasets and let $\mathcal{X}$ denote the random variable corresponding to the confidential dataset, so that different values of $\mathcal{X}$ correspond to different datasets. $\mathcal{U}$ is the set of possible datasets (universe or domain of $\mathcal{X}$) and R is the random variable corresponding to Jessie's record.

Finally, and importantly, all logarithms in this paper are *natural* logarithms (base e).

4 The Privacy-Loss Random Variables and Technical Privacy Desiderata

As explained in Section 2, the main idea underlying all major variants of differential privacy is the ability to distinguish between pairs of datasets that are formally called *neighbors*. There are two commonly used versions of neighbors: *bounded* and *unbounded*

Definition 4.1 (Bounded Neighbors)**.** *Two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are bounded neighbors if $\mathcal{D}_1$ can be obtained from $\mathcal{D}_2$ by* modifying *the record belonging to a single individual (recall that in this document, we assume all data about an individual is encapsulated in one record). In this setting, all datasets under consideration have the same size: n records, and n is public.*

Definition 4.2 (Unbounded Neighbors)**.** *Two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are unbounded neighbors if $\mathcal{D}_1$ can be obtained from $\mathcal{D}_2$ by* adding or removing *the record belonging to a single individual. In this case, the size of the dataset is not public.*

For example, $\mathcal{D}_1 = \mathcal{D} \cup \{r\}$ and $\mathcal{D}_2 = \mathcal{D} \cup \{r'\}$ are *bounded* neighbors since they only differ on the record supplied by one person, say Jessie, and have the same number of respondents (hence the "bounded" terminology). On the other hand, $\mathcal{D}'_1 = \mathcal{D} \cup \{r\}$ and $\mathcal{D}'_2 = \mathcal{D}$ are unbounded neighbors. Note that a record can contain missing values.

Unbounded neighbors represent the cleanest mathematical formulation of privacy comparisons to a counterfactual world in which an individual's data are not used (i.e., would it be difficult to figure out if the input to the privacy mechanism M was $\mathcal{D} \cup \{r\}$ or just $\mathcal{D}$). For this reason, most theoretical frameworks begin with unbounded neighbors. On the other hand, bounded neighbors model the situation where the *existence* of a survey response can be presumed (as in a full enumeration census), but the *contents* of the response must also be protected. Thus, when the population or sample size is revealed without using any randomization, one must use bounded neighbors. For the 2020 Census, the Census Bureau uses bounded neighbors to provide privacy semantics because the size of the U.S. population as of census day, 331,449,281, was published on April 26, 2021. This is the total number of person records in the confidential Census Edited File (CEF). Thus, in the remainder of this document, we focus on bounded neighbors and refer to them simply as neighbors.

The comparison in inference between pairs of neighbors is formalized by the *privacy-loss random variable*, which can be motivated as follows. Suppose a mechanism M was run with either $\mathcal{D}_1$ or $\mathcal{D}_2$ as the input. We observe the output ω from which we can compute $P(M(\mathcal{D}_1) = \omega)$ and $P(M(\mathcal{D}_2) = \omega)$.[2] Note that all of the randomness here comes from M and not from the data. We now wish to use the information provided by ω to distinguish between whether $\mathcal{D}_1$ or $\mathcal{D}_2$ was the input.

This is a classical problem in statistics for which Neyman and Pearson provide the uniformly most powerful test [53]. Treating, say, $\mathcal{D}_1$ as the null hypothesis and $\mathcal{D}_2$ as the alternative, we have a point-null, point-alternative testing framework. The choice of null or alternative is arbitrary because the definition of neighbor is symmetric. One forms the likelihood ratio test statistic[3] $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ and defines a decision rule using a threshold t and a "tie-breaker" probability c. These two numbers are used as follows:

- If $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)} < t$, then reject the null hypothesis.
- If $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)} = t$, then reject the null hypothesis with probability c.
- Otherwise, fail to reject the null hypothesis.

A privacy-loss random variable [23, 31] is simply the distribution of the log-likelihood ratio test statistic $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ under the null hypothesis – i.e., when ω is obtained from $M(\mathcal{D}_1)$. In other words, we can sample the privacy-loss random variable by first getting the output $\omega = M(\mathcal{D}_1)$ and then computing $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$.

Also note that a privacy-loss random variable can be defined for every pair of datasets $\mathcal{D}_1, \mathcal{D}_2$ that are neighbors of each other, so we use the notation $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ to refer to a specific privacy-loss random variable. To summarize:

$$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} \text{ has the distribution of } \log \frac{P(M(\mathcal{D}_1) = \omega)}{P(M(\mathcal{D}_2) = \omega)} \text{ for } \omega \sim P(M(\mathcal{D}_1)).$$

We emphasize, since it is easy to overlook, that for $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$, the dataset $\mathcal{D}_1$ *is used to sample the output ω and also appears in the top of the log ratio.*

We next present a few examples of privacy-loss random variables and then show how they are used to create privacy definitions.

[2] For absolutely continuous distributions use the density function (Radon-Nikodym derivative) instead of the probability mass function.

[3] As noted in Section 3, log is the natural logarithm.

4.1 Examples of Privacy-Loss Random Variables

For simple mechanisms, the privacy-loss random variable can be computed algebraically. These include *Randomized Response* [62], the Geometric Mechanism [38], and the Gaussian Mechanism [32]. We also illustrate the computation of the privacy loss random variable when a mechanism releases a subset of its data.

Example 4.3 (Randomized Response)**.** *Consider a dataset $\mathcal{D} = (r_1, \ldots, r_n)$ of n records (where n is publicly known). Each record r_i corresponds to one individual and contains a unique identifier, a name, whether the person has cancer, and potentially other information. A target person Jessie may or may not be in the dataset. If Jessie is in the dataset and the corresponding record indicates that they have cancer, we say that Jessie has a cancer record in the dataset. If Jessie's record indicates that they do not have cancer or if Jessie's record is not in $\mathcal{D}$ then we say Jessie does not have a cancer record in the dataset. Alternatively, we can consider the case where Jessie is known to be in the dataset (either way, it does not affect what follows). Consider the following* query *function over a dataset:*

$$q(\mathcal{D}) = \begin{cases} 1 & \text{Jessie has a cancer record in } \mathcal{D} \\ 0 & \text{Jessie does not have a cancer record in } \mathcal{D} \end{cases}.$$

Let M be a mechanism that implements randomized response [62]. That is, given a parameter $\epsilon \geq 0$, M flips the result of q with probability $\frac{1}{1+e^\epsilon}$. Mathematically,

$$M(\mathcal{D}) = \begin{cases} q(\mathcal{D}) & \text{with probability } \frac{e^\epsilon}{1+e^\epsilon} \\ 1 - q(\mathcal{D}) & \text{with probability } \frac{1}{1+e^\epsilon} \end{cases}.$$

There is a privacy-loss random variable $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ for each pair[4] of datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ that are neighbors of each other – one pair of neighbors can differ on Jessie, another pair can differ on Bob, and so on. Each of the privacy-loss random variables summarizes the protections available to different pieces of information, and each has one of the following forms:

- *If $\mathcal{D}_1$ and $\mathcal{D}_2$ differ on Jessie and if Jessie has a cancer record in either $\mathcal{D}_1$ or $\mathcal{D}_2$ (but not both) then:*

  $$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \mathcal{L}^M_{\mathcal{D}_2,\mathcal{D}_1} = \begin{cases} \epsilon & \text{with probability } \frac{e^\epsilon}{1+e^\epsilon} \\ -\epsilon & \text{with probability } \frac{1}{1+e^\epsilon} \end{cases}.$$

  This privacy-loss random variable is relevant to inferences about Jessie's cancer status.

- *If $\mathcal{D}_1$ and $\mathcal{D}_2$ differ on Jessie and either (a) Jessie has no cancer records in both datasets or (b) has cancer records in both datasets, then:*

  $$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \mathcal{L}^M_{\mathcal{D}_2,\mathcal{D}_1} = 0 \text{ with probability 1.}$$

  This privacy-loss random variable is relevant to inferences about whatever is different between $\mathcal{D}_1$ and $\mathcal{D}_2$. Since the cancer status of Jessie is the same in both $\mathcal{D}_1$ and $\mathcal{D}_2$, the privacy-loss random variable $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ is providing information on how some other difference between $\mathcal{D}_1$ and $\mathcal{D}_2$ is being protected.

- *Similarly, if $\mathcal{D}_1$ and $\mathcal{D}_2$ differ on a person other than Jessie, then*

  $$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \mathcal{L}^M_{\mathcal{D}_2,\mathcal{D}_1} = 0 \text{ with probability 1,}$$

  because the differences between $\mathcal{D}_1$ and $\mathcal{D}_2$ would not affect the output distribution of M.

Taken together, the set of all privacy-loss random variables model how well different pieces of information are protected. As an example of the calculations used to derive this result, suppose $\mathcal{D}_1$ and $\mathcal{D}_2$ differ only on Jessie's record. Furthermore, suppose that $q(\mathcal{D}_1) = 1$ (i.e., Jessie's record in $\mathcal{D}_1$ is a cancer record) and

[4]Note that *all* pairs of neighboring datasets are considered, not just datasets that are neighboring to the specific dataset collected by the data curator.

$q(\mathcal{D}_2) = 0$ *(Jessie does not have a cancer record in* $\mathcal{D}_2$*). Then for the output* $\omega = 1$, $P(M(\mathcal{D}_1) = 1) = e^\epsilon/(1+e^\epsilon)$ *and* $P(M(\mathcal{D}_2) = 1) = 1/(1+e^\epsilon)$*. The log of their ratio is* ϵ *and the output* $\omega = 1$ *is produced from* $\mathcal{D}_1$ *with probability* $e^\epsilon/(1+e^\epsilon)$*. Thus* $P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \epsilon) = e^\epsilon/(1+e^\epsilon)$*. Similarly, for the output* $\omega = 0$, $P(M(\mathcal{D}_1) = 0) = 1/(1+e^\epsilon)$ *and* $P(M(\mathcal{D}_2) = 0) = e^\epsilon/(1+e^\epsilon)$*. The log of their ratio is* $-\epsilon$ *and the output 0 is produced from* $\mathcal{D}_1$ *with probability* $1/(1+e^\epsilon)$*. Thus* $P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = -\epsilon) = 1/(1+e^\epsilon)$*.*

Example 4.4 (Geometric Mechanism)**.** *The two-sided geometric random variable, with parameter* ϵ*, is integer-valued and has the probability distribution* $P(k) = \frac{1-e^{-\epsilon}}{1+e^{-\epsilon}} e^{-\epsilon|k|}$*. Consider the mechanism* M *that counts the number of cancer records in the dataset and adds two-sided geometric noise to the result.*[5] *This is known as the Geometric Mechanism [38]. Simple calculations show that the forms of the privacy-loss random variables are exactly the same as in Example 4.3.*

These two examples show that the privacy-loss random variables do not uniquely define the privacy mechanism M.

Example 4.5 (Gaussian Mechanism)**.** *Now consider a variation of Example 4.4 that uses Gaussian noise instead of double geometric. That is,* M *adds* $N(0, \sigma^2)$ *noise to the total number of people with cancer records in the data. This is called the* Gaussian Mechanism *[32]. There is a privacy-loss random variable* $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ *for each pair of datasets* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *that are neighbors of each other and they have the following forms:*

- *Without loss of generality, let Alice be the person whose records differ between* $\mathcal{D}_1$ *and* $\mathcal{D}_2$*. If Alice has a cancer record in either* $\mathcal{D}_1$ *or* $\mathcal{D}_2$ *(but not both) then*

$$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} \sim N\left(\frac{1}{2\sigma^2}, \frac{1}{\sigma^2}\right).$$

- *Let Alice be the person whose records differ between* $\mathcal{D}_1$ *and* $\mathcal{D}_2$*. If Alice has no cancer records in both datasets, then:*

$$\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \mathcal{L}^M_{\mathcal{D}_2,\mathcal{D}_1} = 0 \text{ with probability 1.}$$

We note that the privacy-loss random variables in Example 4.3 and 4.4 are bounded in absolute value by the parameter ϵ used by the mechanisms, while for Example 4.5, it is only concentrated around the mean, with the mean and variance decreasing as σ^2 (the variance of the privacy noise) increases.

Example 4.6 (Composition)**.** *Let* M_1 *and* M_2 *be two mechanisms whose sources of randomness are independent of each other. Let* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *be two neighboring datasets, and let* M^* *be the mechanism that, on input* $\mathcal{D}$*, outputs both* $M_1(\mathcal{D})$ *and* $M_2(\mathcal{D})$ *(i.e., it releases the output of both mechanisms). It is easy to show that:*

$$\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2} = \mathcal{L}^{M_1}_{\mathcal{D}_1,\mathcal{D}_2} + \mathcal{L}^{M_2}_{\mathcal{D}_1,\mathcal{D}_2}.$$

that is, the privacy-loss random variable of a combined data release is the sum of the individual privacy-loss random variables. The reason is that the distribution of $\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2}$ *is the distribution of the random variable* $\log \frac{P(M_1(\mathcal{D}_1)=\omega_1)P(M_2(\mathcal{D}_1)=\omega_2)}{P(M_1(\mathcal{D}_2)=\omega_1)P(M_2(\mathcal{D}_2)=\omega_2)}$ *when* ω_1 *and* ω_2 *are sampled independently from the output distributions* $P(M_1(\mathcal{D}_1))$ *and* $P(M_2(\mathcal{D}_2))$*, respectively, which is the same as the distribution of* $\log \frac{P(M_1(\mathcal{D}_1)=\omega_1)}{P(M_1(\mathcal{D}_2)=\omega_1)} + \log \frac{P(M_2(\mathcal{D}_1)=\omega_2)}{P(M_2(\mathcal{D}_2)=\omega_2)}$*, when* ω_1 *and* ω_2 *are sampled independently from the output distributions* $P(M_1(\mathcal{D}_1))$ *and* $P(M_2(\mathcal{D}_2))$*, respectively. In general, the analysis of the combined privacy cost of several mechanisms is known as* composition *[37].*

Example 4.7 (Random Sampling)**.** *Consider again a dataset* $\mathcal{D}_1 = (r_1, \ldots, r_n)$ *of* n *records, where* n *is publicly known and where each record* r_i *corresponds to one individual. Suppose Jessie has a record in* $\mathcal{D}_1$*, say* r_1*. Let* $\mathcal{D}_2 = (r'_1, r_2 \ldots, r_n)$ *be a neighboring dataset in which there is some other record in place of*

[5] Note that the noisy count may be negative and an end-user might wish to perform post-processing such as truncating negative counts or using proper statistical inference since the distribution of the added noise is known. None of these post-processing steps harm privacy, a property of differential privacy that is known as *postprocessing invariance* and is discussed in Section 4.2.

Jessie's record, but everything else is the same. Consider a mechanism M such that $M(\mathcal{D})$ returns a set of m records uniformly sampled (without replacement) from its input $\mathcal{D}$.

If ω is a set of m records and $\omega \subseteq \{r_2, \ldots, r_n\}$ then $P(M(\mathcal{D}_1) = \omega) = P(M(\mathcal{D}_2) = \omega)$. Note that the log of the ratio of the probabilities is 0 and that ω provides no information at all about Jessie. When $\mathcal{D}_1$ is the input, such an ω is produced with probability $\frac{\binom{n-1}{m}}{\binom{n}{m}} = \frac{n-m}{n}$.

If $r_1 \in \omega$ and the rest of the records in ω are a subset of $\{r_2, \ldots, r_n\}$ then $P(M(\mathcal{D}_1) = \omega) > 0$ while $P(M(\mathcal{D}_2) = \omega) = 0$, so the ratio of the probabilities is ∞. In this situation, anyone can clearly tell that the input to M could not have been $\mathcal{D}_2$. When $\mathcal{D}_1$ is the input, such an even happens with probability $1 - \frac{n-m}{n} = \frac{m}{n}$.

If $r_1' \in \omega$ and the rest of the records in ω are a subset of $\{r_2, \ldots, r_n\}$ then $P(M(\mathcal{D}_1) = \omega) = 0$ while $P(M(\mathcal{D}_2) = \omega) > 0$. However, if $\mathcal{D}_1$ is the input, such an event happens with probability 0. Putting all this together, the privacy loss random variable $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ has the distribution:

$$P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = 0) = \frac{n-m}{n}$$
$$P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} = \infty) = \frac{m}{n}$$

This privacy loss random variable says that the output ω either provides no information to an attacker trying to distinguish between $\mathcal{D}_1$ and $\mathcal{D}_2$ or allows the attacker to distinguish between them perfectly. The former situation happens with probability $\frac{n-m}{n}$ and the later with probability $\frac{m}{n}$.

4.2 From Privacy-Loss Random Variables to Privacy Definitions and Privacy Accounting Frameworks

The set of privacy-loss random variables $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}$ (one for each pair of neighboring datasets) associated with a mechanism M capture the privacy properties of the mechanism. *Privacy definitions* are statements of desirable properties that the mechanism M and its privacy-loss random variables should have.

For example, let us consider the case of *pure differential privacy*, also known as ϵ-differential privacy:

Definition 4.8 (ϵ-differential privacy/pure differential privacy)**.** *Given an $\epsilon \geq 0$, a mechanism M satisfies ϵ-differential privacy if for all pairs of neighbors $\mathcal{D}_1, \mathcal{D}_2$ and all (measurable) sets S,*

$$P(M(\mathcal{D}_1) \in S) \leq e^\epsilon P(M(\mathcal{D}_2) \in S). \tag{1}$$

Equivalently, a mechanism M satisfies ϵ-differential privacy if for any pair $(\mathcal{D}_1, \mathcal{D}_2)$ that are neighbors of each other, the corresponding privacy-loss random variable satisfies: $\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} \leq \epsilon$ with probability 1. Note that the probabilities in this definition are taken only with respect to M and not with respect to any randomness in the data.

First, we note the symmetry in the definition. If $(\mathcal{D}_1, \mathcal{D}_2)$ is a pair of neighbors, then so is $(\mathcal{D}_2, \mathcal{D}_1)$ and therefore Definition 4.8 also requires $P(M(\mathcal{D}_2) \in S) \leq e^\epsilon P(M(\mathcal{D}_1) \in S)$. This kind of symmetry holds for all the privacy definitions we discuss.

From Definition 4.8, we see that pure differential privacy summarizes the properties of M using a single number ϵ, referred to as a *privacy loss* (or privacy cost). When ϵ is close to 0, it guarantees that $P(M(\mathcal{D}_1) \in S) \approx P(M(\mathcal{D}_2) \in S)$ so that changing the content of any record barely affects the probability of any outcome. In terms of the privacy-loss random variable, it guarantees that the log-likelihood ratio test statistic is always bounded by ϵ (we cover these semantics in greater detail in Sections 6 and 7).

The general pattern in the formal privacy literature is that some property of the privacy-loss random variables is treated as a summary of the privacy loss of a mechanism. For example, the maximum value achievable by the privacy-loss random variables is the privacy-loss ϵ in pure differential privacy. Such an assignment of cost is also known as a privacy accounting framework. However, not all properties of the privacy-loss random variables are useful measures of privacy cost. There are certain additional properties that a privacy accounting framework should also have. These are called *transparency*, *post-processing invariance*, and *composition*.

Transparency means that the accounting framework must assume that the attacker knows how M works (i.e., may have access to its source code). This is a crucial property for data quality – SDL methods that are *not* transparent must hide their source code and other key details, meaning that a statistician would have no way of adjusting inferences to account for SDL perturbations. An example of a non-transparent system is data-swapping, which was used for the 2010 Census and for which exact details, such as the swap rate and likelihood of different households to be swapped, were kept confidential. On the other hand, differential privacy and its variants support transparency because they can be translated into properties of the privacy-loss random variable which, by construction, makes use of knowledge of M and the relations between its inputs and outputs.

Another reason transparency is important for data quality is that it lets the public audit the software used for the production of official statistics. There is strong evidence that disclosure avoidance implementation mistakes have happened in the past. For example, Alexander et al. [5] analyzed data from the American Community Survey (ACS), Current Population Survey (CPS), and the 2000 Census and found errors of up to 15% in statistics for men and women over the age of 65. The errors were attributed by them to mistakes in the disclosure avoidance system used at the time. The errors were detectable for the 2000 Census and ACS data because the disclosure avoidance was applied using different methods for published tabular summaries and microdata [4]. Had identical methods been used, as was the case for the CPS, significant data quality issue could have remained undetected. More generally, if code and algorithmic details are not provided, coding mistakes can have a significant impact on data quality and be difficult, if not impossible, to catch.

Post-processing invariance means that secondary analysis of the output of M is permissible (i.e., secondary analysis does not increase privacy risks). For example, if the input dataset is $\mathcal{D}$, the output $M(\mathcal{D})$ could take the form of privacy-protected microdata or noisy tabulations. An analysis function A could be applied to these results – for example, it could calculate the age distribution in each state, or it could combine the output of M with additional data sources to produce a hierarchical data model. We use the notation $A \circ M$ to denote the process of running M on the data and A on the result. A privacy-accounting framework should *not* claim that $A(M(\mathcal{D}))$ has a higher privacy cost than $M(\mathcal{D})$ because A has no direct access to $\mathcal{D}$, and A can be performed by external users once the output of $M(\mathcal{D})$ is published. Simply put, post-processing invariance means that the privacy loss attributed to $A \circ M$ should not be larger than that of M. Violations of post-processing invariance can be likened, in the financial world, to hidden costs or the practice of surprise billing.

Post-processing invariance is a fundamental and useful property because it rules out seemingly intuitive quantities as the basis for privacy-loss accounting. For instance, it is well known that the tail probabilities of privacy-loss random variables are not post-processing invariant [15, 50]. That is, for a fixed t, $P(\mathcal{L}^{M_1}_{\mathcal{D}_1,\mathcal{D}_2} > t)$ may increase or decrease with postprocessing as the following example shows:

Example 4.9. *Let M_1 be the randomized response mechanism from Example 4.3 (that asks if Jessie has cancer) and let M_2 be a probabilistic copy of M_1 (that is, M_2 is the same as M_1 except that its source of randomness is independent of M_1). Let M^* be the mechanism that, on input $\mathcal{D}$, outputs the results of both mechanisms: $M^*(\mathcal{D}) = (M_1(\mathcal{D}),\ M_2(\mathcal{D}))$. From Examples 4.3 and 4.6, if Jessie has a cancer record in $\mathcal{D}_1$ but not in $\mathcal{D}_2$ then:*

$$\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2} = \begin{cases} 2\epsilon & \text{with probability } \frac{e^{2\epsilon}}{(1+e^\epsilon)^2} \\ 0 & \text{with probability } 2\frac{e^\epsilon}{(1+e^\epsilon)^2} \\ -2\epsilon & \text{with probability } \frac{1}{(1+e^\epsilon)^2} \end{cases}$$

Now consider two post-processing functions A_1 and A_2. The function A_1 ignores its input and always outputs the number 1. The function A_2 takes the output of M^* and ignores the second part (so that $A_2(M^*(\mathcal{D})) = M_1(\mathcal{D})$). This results in the following privacy-loss random variables:

$$\mathcal{L}^{A_1\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2} = 0 \text{ with probability } 1$$

$$\mathcal{L}^{A_2\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2} = \begin{cases} \epsilon & \text{with probability } \frac{e^\epsilon}{1+e^\epsilon} \\ -\epsilon & \text{with probability } \frac{1}{1+e^\epsilon} \end{cases}$$

From these equations, we can see that $P(\mathcal{L}^{A_1\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon/2) = 0$, which is less than $P(\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon/2) = \frac{e^{2\epsilon}}{(1+e^\epsilon)^2}$, which is less than $P(\mathcal{L}^{A_2\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon/2) = \frac{e^\epsilon}{1+e^\epsilon}$ and thus different choices for post-processing can

increase or decrease the tail probabilities. Note that the mechanisms and the post-processing in this example continue to satisfy Definition 4.8. The privacy-loss random variables $\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2}$, $\mathcal{L}^{A_1\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2}$, *and* $\mathcal{L}^{A_2\circ M^*}_{\mathcal{D}_1,\mathcal{D}_2}$ *are all bounded by* 2ϵ *with probability 1, so the mechanisms* M^*, $A_1 \circ M^*$, *and* $A_2 \circ M^*$ *all satisfy* (2ϵ)*-differential privacy. This illustrates that pure differential privacy satisfies post-processing invariance, but the tail-areas of privacy-loss random variables do not.*

Composition. Composition refers to the combined privacy loss due to the release of the outputs of multiple mechanisms $M_1, \ldots, M_k$ and hence covers, as a special case, the privacy loss due to the release of multiple privacy-protected datasets. For example, if $M_1, \ldots, M_k$ satisfy pure differential privacy with parameters $\epsilon_1, \ldots, \epsilon_k$, respectively, then their combined privacy cost is at most the sum of the individual costs: $\sum_{i=1}^{k} \epsilon_i$ – in other words, the mechanism that releases the outputs of $M_1, \ldots, M_k$ satisfies $(\sum_{i=1}^{k} \epsilon_i)$-differential privacy (although often the actual privacy parameter is even smaller than this) [29]. This is a desirable property because it shows that differential privacy can account for the interactions from output released using different mechanisms. Note that this composition is linear and can be likened to an actual budget of $\sum_{i=1}^{k} \epsilon_i$ that is allocated across different mechanisms.

On the other hand, many SDL methods including k-anonymity [57], cell suppression [25, Ch. 4], and other legacy disclosure avoidance methods do not have suitable composition properties. That is, each data release $M_i(\mathcal{D})$ may *individually* satisfy some SDL confidentially requirements, but the combination of all of the data releases can fail to satisfy any of those requirements (under any parameter settings), in many cases completely revealing confidential information such as individual records [37]. This failure of composition is similar to the idea that three linear equations on three variables can determine them uniquely even though any individual equation retains a large amount of uncertainty about their exact values. On the other hand, formal privacy methods consider future possible interaction between the outputs of different mechanisms.

There are several important types of composition to consider.

- *Independent composition.* Suppose $M_1, \ldots, M_k$ are mechanisms whose only inputs are the data $\mathcal{D}$ (they are unaware of the outputs of the other mechanisms). Independent composition refers to their overall privacy cost. For example, as discussed earlier, if each M_i satisfies ϵ_i-differential privacy, then the combined release of the outputs of $M_1, \ldots, M_k$ satisfies ϵ-differential privacy for an ϵ that is guaranteed to satisfy $\epsilon \leq \sum_i \epsilon_i$. This is a linear composition property. Some privacy definitions support a sublinear composition property, which would be referred to as *advanced composition* [32, 34, 42].

- *Adaptive Composition.* Next, suppose the input to each M_i is the dataset $\mathcal{D}$ along with the output ω_{i-1} of the previous mechanism. This allows a mechanism to adapt its behavior based on the history of the previous outputs. Now, for each fixed ω_{i-1}, we can view $M_i(\omega_{i-1}, \cdot)$ to be a mechanism whose input is $\mathcal{D}$ (the dot is a placeholder for where to insert the data). Now, suppose that for each i, and for each ω_{i-1}, the mechanism $M_i(\omega_{i-1}, \cdot)$ satisfies ϵ_i-differential privacy. Adaptive composition refers to the overall privacy cost of releasing those outputs $\omega_1, \ldots, \omega_k$. For the case of pure differential privacy, again we have the guarantee that the privacy cost ϵ satisfies $\epsilon \leq \sum_i \epsilon_i$ [32], so pure differential privacy also has adaptive composition.

- *Fully Adaptive Composition.* Next, suppose further that each mechanism can adjust its privacy cost based on the output history. That is the privacy cost of $M_i(\omega_{i-1}, \cdot)$ may differ from that of $M_i(\omega'_{i-1}, \cdot)$ for $\omega_{i-1} \neq \omega'_{i-1}$. Fully adaptive composition refers to the overall privacy cost of this scenario [56, 36]. It turns out that under pure differential privacy, if for all possible sequences $\omega_1, \ldots, \omega_k$, the sum of the individual privacy costs is at most some value ϵ^* with probability 1, then the combined release of all of the outputs satisfies ϵ^*-differential differential privacy, so that pure differential privacy satisfies fully adaptive composition [56]. This is a highly desirable property because it allows policy makers to make their future decisions on data releases and privacy budgets based on what has already been released.

- *Convexity.* Convexity is a degenerate case of adaptive composition that is useful as an easy-to-prove sanity check for proposals of new privacy accounting frameworks. Suppose M_1 and M_2 satisfy the same privacy definition with the same privacy-loss parameters. Consider the mechanism M^* such that $M^*(\mathcal{D})$ outputs $M_1(\mathcal{D})$ with probability p, and otherwise outputs $M_2(\mathcal{D})$ (in other words, it randomly chooses which mechanism to run). Does M^* also satisfy the same privacy definition with the same parameters? If yes, the privacy definition is convex. If not, it is not convex [44]. This is a special

case of adaptive composition in which an initial mechanism ignores the data (hence has 0 privacy cost) and outputs 0 or 1, and a subsequent mechanism uses that bit to decide whether it acts like M_1 or M_2. Note that this definition of convexity is applicable when M_1 and M_2 satisfy a privacy definition under the same parameters, it is not a statement about what happens when the privacy parameters are different.

Such composition properties of formal privacy accounting frameworks allow policymakers to treat privacy cost as a resource, controlled by something called the *privacy-loss budget*. The privacy-loss budget can be allocated across the different mechanisms $M_1, \ldots, M_k$ – just like the overall ϵ in pure differential privacy can be split into parts, with each part ϵ_i allocated to a mechanism M_i. Furthermore, the privacy-loss budget and its allocation across different mechanisms can be turned into semantic privacy guarantees, which we explain in Sections 6, 7, and 8. But first, in the Section 5, we explain the major variants of differential privacy that provide the semantic guarantees.

5 Major Variants of Differential Privacy and Parameter Interpretability

Pure differential privacy (Definition 4.8) was used in Section 4 to illustrate how privacy definitions can be formed from privacy-loss random variables and to explain the desirable properties of formal privacy definitions. In this section, we explain the other major variations of differential privacy used in this paper, their relation to the privacy-loss random variables, and discuss the added value they provide. We also, where possible, attempt to provide a minimally technical interpretation of the privacy parameters that can be used for policy discussions.

The starting point is the discussion of the ϵ parameter in pure differential privacy. Let us consider the situation where an attacker is trying to guess a sensitive property of a target individual Jessie (e.g., does Jessie have cancer or not) based on the output ω of a disclosure avoidance algorithm M. Even if the attacker has complete information about everyone else, and even if the attacker knows everything except Jessie's cancer status (so that the only question is whether the input was $\mathcal{D}_1$ in which *Jessie* has cancer or its neighbor $\mathcal{D}_2$, in which Jessie does not have cancer), the odds that the attacker correctly guesses the cancer status of Jessie's record is at most[6] e^ϵ times the odds they would be correct if Jessie's record were replaced with a different record before being processed by M. In other words, participation in the data changes the odds of a correct guess by a factor of at most e^ϵ.

This is a strong guarantee, but in many cases the actual odds are significantly lower than the maximum bound.

For instance, consider the mechanism M^* of Example 4.9 which uses randomized response to output Jessie's cancer status (resulting in the first output ω_a) followed by another randomized response on Jessie's cancer status (resulting in output ω_b). When $\omega_a = \omega_b = 1$, then the odds of an attacker guessing correctly change by a factor of $\frac{P(M^*(\mathcal{D}_1)=(1,1))}{P(M^*(\mathcal{D}_2)=(1,1))} = e^{2\epsilon}$. However, if $\omega_a \neq \omega_b$ then ω_a and ω_b contradict each other, and the odds do not change at all (e.g., $\frac{P(M^*(\mathcal{D}_1)=(1,0))}{P(M^*(\mathcal{D}_2)=(1,0))} = 1$). Finally, if $\omega_a = \omega_b = 0$, the odds *decrease* by a factor of $e^{2\epsilon}$ (i.e., get multiplied by $e^{-2\epsilon}$) making the attacker more likely to guess incorrectly. What we see is that the change in odds is a random variable (because it depends on ω, which is random) and the actual changes in odds are often lower than the worst-case $e^{2\epsilon}$.

This discussion is captured by the privacy-loss random variable $\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2}$ of Example 4.9, which takes the value 2ϵ (corresponding to an increase in odds of $e^{2\epsilon}$) with probability $\frac{e^{2\epsilon}}{(1+e^\epsilon)^2}$, takes the value 0 (no change in odds) with probability $2\frac{e^\epsilon}{(1+e^\epsilon)^2}$, and -2ϵ (e.g., the odds decrease by $e^{2\epsilon}$) with the remaining probability. In other words, there is a good chance that the privacy loss random variable does not hit its maximum value.

This situation is common when M^* is composed of many mechanisms (instead of just *two* mechanisms), in which case the chance that ω causes the worst-case change in odds can become vanishingly small. Thus it is helpful to track the rest of the distribution of the privacy-loss random variable, and not just its extreme value.

[6] I.e., the odds would change by $\frac{P(M(\mathcal{D}_1)=1)}{P(M(\mathcal{D}_2)=1)} \leq e^\epsilon$

5.1 Approximate Differential Privacy

Approximate differential privacy [28], also known as (ϵ, δ)-differential privacy, is a relaxation of pure differential privacy. Its aim is to allow a mechanism to produce a bad outcome ω (for which the change in odds is large) as long as the probability of bad outcomes is vanishingly small (i.e., harder than winning the lottery). It is formally defined as follows:

Definition 5.1 ((ϵ, δ)-differential privacy [28])**.** *Given privacy parameters $\epsilon \geq 0$ and $\delta \in [0, 1]$, a mechanism M satisfies (ϵ, δ)-differential privacy if for all pairs $\mathcal{D}_1, \mathcal{D}_2$ that are neighbors of each other and for all measurable sets S,*

$$P(M(\mathcal{D}_1) \in S) \leq e^{\epsilon} P(M(\mathcal{D}_2) \in S) + \delta. \tag{2}$$

Equivalently [11], M satisfies (ϵ, δ)-differential privacy if

$$P(\mathcal{L}^{M}_{\mathcal{D}_1, \mathcal{D}_2} \geq \epsilon) - e^{\epsilon} P(\mathcal{L}^{M}_{\mathcal{D}_2, \mathcal{D}_1} \leq -\epsilon) \leq \delta. \tag{3}$$

This relation uses two *privacy loss random variables: $\mathcal{L}^{M}_{\mathcal{D}_1, \mathcal{D}_2}$ for the first probability and its reverse $\mathcal{L}^{M}_{\mathcal{D}_2, \mathcal{D}_1}$ for the second probability.*

The condition in Equation 2 has an important interpretation. When $\epsilon = 0$ then δ is the total variation distance between the output distributions of $M(\mathcal{D}_1)$ and $M(\mathcal{D}_2)$. Thus, approximate differential privacy is an interpolation between the differential privacy equations (when $\delta = 0$) and total variation distance (when $\epsilon = 0$). These conditions on the privacy-loss random variables were discovered after approximate differential privacy was proposed. It is worth noting that Equation 3 does not use the privacy-loss random variable in an intuitive way and portends the difficulties in interpreting the δ parameter.

The δ parameter of Definition 5.1 is often *incorrectly* interpreted as an upper bound on the probability that a privacy-loss random variable will exceed ϵ (i.e., an upper bound on the probability that the odds of correctly guessing a sensitive piece of information will change by at least e^{ϵ}). However, this is known to be false [15], as can be seen from Equation 3. However, δ is an upper bound on something else – the probability that the privacy-loss random variable is infinite: $\delta \geq P(\mathcal{L}^{M}_{\mathcal{D}_1, \mathcal{D}_2} = \infty)$. The privacy-loss random variable can only be infinite if there is an output ω such that $P(M(\mathcal{D}_1) = \omega) > 0$ and $P(M(\mathcal{D}_2) = \omega) = 0$, in which case $\log \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)} = \infty$. However, Equation 2 guarantees that this output is only seen with probability at most δ when $\mathcal{D}_1$ is the input.

Infinite privacy loss is also known as a catastrophic failure. Two well-known $(0, \delta)$-differentially private mechanisms for which this can happen are (1) the mechanism that, with probability δ, releases the entire dataset and (2) the mechanism that releases a random person's record when the input dataset has $n \geq 1/\delta$ people (see Example 4.7). For this reason, it is often recommended that δ be a very small quantity, much smaller than $1/n$ [43].

In practice, however, most (ϵ, δ)-differentially private mechanisms cannot have infinite privacy loss, and so the catastrophic-failure interpretation of δ is not relevant. One example of a mechanism that never catastrophically fails is the Gaussian mechanism from Example 4.5. This mechanism satisfies (ϵ, δ)-differential privacy for infinitely many combinations of ϵ and δ. Specifically, the Gaussian mechanism satisfies (ϵ, δ)-differential privacy for every point above the red curve in Figure 1 (we explain the other curves later in this section).

Every mechanism has its own Pareto curve of (ϵ, δ) values and the privacy semantics depend on the entire curve [9, 24, 58] – a single point on the curve is not very informative. We note that as long as this curve is not bounded from below by some positive constant, then there can be no catastrophic output with infinite privacy loss.

One important reason for considering the entire curve is composition. Using only one point on the (ϵ, δ) curve results in weaker composition properties than those achievable when the entire curve is used. If one uses only individual points on the curve, approximate differential privacy satisfies adaptive composition [32]: if $M_1, \ldots, M_k$ individually satisfy approximate differential privacy with parameters $(\epsilon_1, \delta_1), \ldots, (\epsilon_k, \delta_k)$, the mechanism that releases all of their outputs satisfies $(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i)$-differential privacy. This naive composition result can be substantially improved [42]. However, these results are generic in the sense that they only acknowledge a single point on the (ϵ, δ) curve of each mechanism. If the entire curve of each

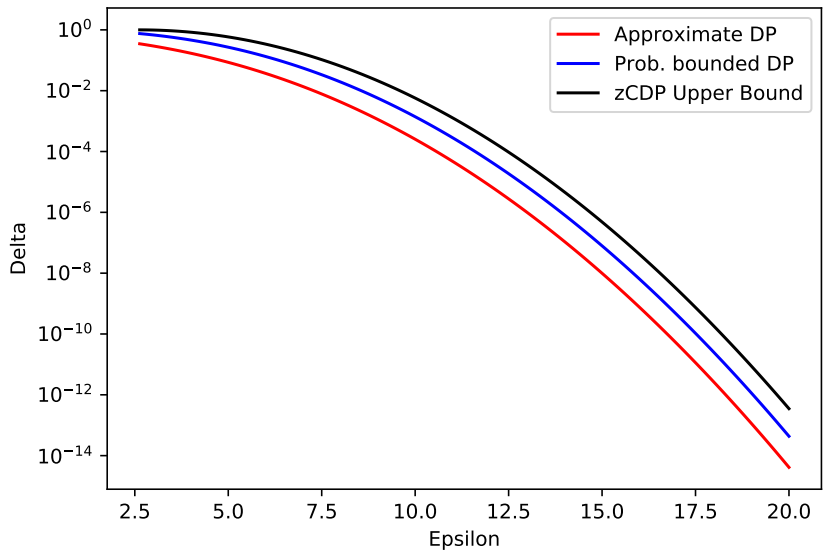

Figure 1: The (ϵ, δ) curves under Approximate Differential Privacy, probabilistically-bounded differential privacy, and the zCDP upper bound, had continuous (instead of discrete) Gaussian noise been used in the 2020 redistricting data. This is equivalent to using Gaussian noise variance $\sigma^2 = 1/2.63$ in Example 4.5.

mechanism is used, tighter composition results are achievable. This is the motivation for Rényi differential privacy (RDP) [51], zero-concentrated differential privacy (zCDP) [14], and f-DP [24] that we discuss later. Privacy accounting is dramatically simplified when using RDP and zCDP, and their parameters can be used to upper bound the parameters of other definitions as needed (e.g., they can generate an upper bound on the (ϵ, δ) curve).

Given a value for ϵ, one interpretation of δ, due to Bun and Steinke [14, 52], is the following. If $\mathcal{D}_1$ is the input to M and there is an output ω such that $\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)} \leq e^{\epsilon}$ then outputting ω was a good event. On the other hand, if $\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)} > e^{\epsilon}$ then this is not necessarily a bad event. Instead, one flips a coin that has $P(\text{heads}) = e^{\epsilon} \Big/ \frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$. If that lands heads, it is a good event, and it if lands tails, then it is a bad event. In other words, the probability that observing ω is considered bad depends on by how much the ratio $\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ exceeds e^{ϵ}. Then, conditional on a bad event not happening, the probability ratio is $\leq e^{\epsilon}$, and the overall probability of a bad event is $\leq \delta$ [14, 52]. Given the fairly complex nature of a "bad" event and the conditional probabilities involved, this interpretation is difficult to explain to policy makers. Desfontaines [20] gives another interpretation of δ as "the mass of all possible bad events, weighted by how likely they are and how bad they are" that is also fairly technical.

The main difficulty with interpreting δ is that it appears to be a parameter that was chosen for mathematical convenience. The goal of the next section is to start with a more intuitive quantity and then see where the math leads.

5.2 Probabilistically-bounded differential privacy and f-DP

Executive decision-makers charged with setting privacy policy need interpretable privacy parameters to support their decisions. This has led us to create an alternative to approximate differential privacy called (ϵ, δ)-probabilistically-bounded differential privacy (pbdp). To the best of our knowledge, this privacy definition has not been proposed before, but it turns out to have natural connections to and serve as good motivation for f-DP, zCDP, and RDP. Its δ parameter is much more interpretable, and the privacy definition is post-processing invariant. As with approximate differential privacy, a mechanism satisfies *pbdp* for a

continuum of (ϵ, δ) values. If one acknowledges only one point on this curve, the privacy definition turns out to be not convex (see Section 4.2). However, when the entire curve is considered, not only does the privacy definition become convex, but it also becomes equivalent to a re-parametrization of f-DP. Thus, we view our result as a demonstration that f-DP is a natural consequence of addressing the interpretability issues of approximate differential privacy.

Hence, the ultimate goal of this section is to define pbdp as a way of explaining and motivating f-DP. Later, in Section 5.3, we discuss zCDP and RDP, which solve a computational problem – the exact privacy parameters of pbdp and f-DP are difficult to compute – but are easy to upper bound with zCDP and RDP.

The prerequisite to explaining *pbdp* to policy makers is to first have them understand the meaning of the odds $\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ – i.e., first they need to understand pure differential privacy.

We begin by considering a flawed approach for modifying approximate differential privacy to make δ interpretable – setting δ to be the tail probability of the privacy-loss random variables. This is the same as the probability of observing an ω for which the odds $\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ are too large (greater than e^ϵ). Formally, suppose we require that for all pairs of neighbors $\mathcal{D}_1$ and $\mathcal{D}_2$, that $P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) \leq \delta$. This version of δ is the upper bound on the probability that the odds of guessing a sensitive attribute correctly change by more than e^ϵ. Another way of saying this is that such a δ is a bound on the probability that the ϵ-differential privacy constraints (Equation 1) fail to hold.[7] However, as noted in [50] and shown in Example 4.9 the tail probability $P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon)$ is not post-processing invariant and so this first attempt is not viable. Post-processing invariance can be added as follows, by considering all possible post-processing functions.

Definition 5.2 (Probabilistically-bounded differential privacy)**.** *Given privacy parameters $\epsilon > 0$ and $\delta \in [0, 1]$, a mechanism M satisfies (ϵ, δ)-pbdp if for all (possibly randomized) post-processing functions A (whose domain is the range of M) and all pairs $\mathcal{D}_1, \mathcal{D}_2$ of datasets that are neighbors of each other,*

$$P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) \leq \delta.$$

Definition 5.2 is post-processing invariant by construction. In fact, Definition 5.2 provides the smallest δ that is both post-processing invariant and an upper bound on the privacy-loss tail probabilities. One can think of the post-processing algorithms A as secondary analyses of the output of M and the definition ensures that no secondary analysis can raise the odds (of correctly guessing a piece of sensitive information) more than e^ϵ, except with probability bounded by δ. However, even though this δ now has a very concrete interpretation, Definition 5.2 is not easy to work with because it is defined directly in terms of privacy-loss random variables. One can perform a simplification that removes the reference to the privacy-loss random variable:

Definition 5.3 (Probabilistically-bounded differential privacy, version 2)**.** *Given privacy parameters $\epsilon > 0$ and $\delta \in [0, 1]$, a mechanism M satisfies (ϵ, δ)-pbdp if for all (possibly randomized) post-processing functions A whose domain is the range of M and whose output is 0 or 1, and for all pairs $\mathcal{D}_1, \mathcal{D}_2$ of datasets that are neighbors of each other,*

$$\Big(P(A(M(\mathcal{D}_1)) = 1) > e^\epsilon P(A(M(\mathcal{D}_2)) = 1)\Big) \;\Rightarrow\; \Big(P(A(M(\mathcal{D}_1)) = 1) \leq \delta\Big).$$

Theorem 5.4. *Definitions 5.2 and 5.3 are equivalent for $\delta \in [0, 1]$.*

The binary-valued post-processing algorithms A in Definition 5.3 can be thought of as attack algorithms (does the algorithm predict Jessie has some property or not), or, equivalently, as a test between two hypotheses (H_0 : Jessie has the property vs. H_1 : Jessie does not).

The next simplification is the easiest to work with:

Definition 5.5 (Probabilistically-bounded differential privacy, version 3)**.** *Given privacy parameters $\epsilon > 0$ and $\delta \in (0, 1]$, a mechanism M satisfies (ϵ, δ)-pbdp if for all (possibly randomized) post-processing functions A whose domain is the range of M and whose output is 0 or 1, and for all pairs $\mathcal{D}_1, \mathcal{D}_2$ of datasets that are neighbors of each other,*

$$\Big(P(A(M(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta\Big) \;\Rightarrow\; \Big(P(A(M(\mathcal{D}_2)) = 1) \leq \delta\Big).$$

[7] This is equivalent to something that is called probabilistic differential privacy [50].

Remark 5.6. *Note that Definition 5.5 excludes the case of $\delta = 0$ while the others do not. The first reason is that setting $\delta = 0$ would be equivalent to saying that if $M(\mathcal{D}_1)$ cannot produce a certain output, then neither can $M(\mathcal{D}_2)$ and vice versa. This is clearly not equivalent to ϵ-differential privacy. However, it can be shown (see the Appendix) that if a mechanism M satisfies Definition 5.5 for a fixed ϵ and all $\delta > 0$, then it satisfies pure ϵ-differential privacy and vice versa.*

Theorem 5.7. *Definitions 5.3 and Definition 5.5 are equivalent for $\delta \in (0, 1]$.*

Even though the form of Definition 5.5 is now very different from the original motivation, the chain of equivalences maintains the original post-processing-invariant interpretations of ϵ and δ that, with probability at most δ, pbdp M produces a bad output ω (i.e., one that changes the odds of correctly guessing sensitive information by more than e^{ϵ}). Nevertheless, there is still one drawback. Namely, if one only considers a single (ϵ, δ) pair instead of the entire curve, then, as the following example shows, the privacy definition is not convex.

Example 5.8. *Let M_1 be a variation of the randomized response mechanism such that:*

$$P(M_1(\mathcal{D}_1) = 2) = \frac{e}{1+e}$$
$$P(M_1(\mathcal{D}_1) = -2) = \frac{1}{1+e}$$
$$P(M_1(\mathcal{D}_2) = 2) = \frac{1}{1+e}$$
$$P(M_1(\mathcal{D}_2) = -2) = \frac{e}{1+e}$$

Let M_2 be the mechanism such that:

$$P(M_2(\mathcal{D}_1) = 1) = 0$$
$$P(M_2(\mathcal{D}_1) = 0) = 1$$
$$P(M_2(\mathcal{D}_2) = 1) = 0.01$$
$$P(M_2(\mathcal{D}_2) = 0) = 0.99$$

First we claim and show that M_1 and M_2 both satisfy $(\epsilon = 1.01, \delta = 0.02)$-pbdp.

M_1 satisfies $(\epsilon = 1.01, \delta = 0.02)$-pbdp because (1) it satisfies 1-differential privacy and so does $A \circ M_1$ for any (binary-valued) postprocessing algorithm A, (2) this means $P(A(M_1(\mathcal{D}_1)) = 1) \geq e^{-1}P(A(M_1(\mathcal{D}_2)) = 1)$, so (3) if $P(A(M_1(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta$ then $e^{-1}P(A(M_1(\mathcal{D}_2)) = 1) \leq e^{-\epsilon}\delta$ and so $P(A(M_1(\mathcal{D}_2)) = 1) \leq \delta$ for $\epsilon = 1.01$. The same happens with the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ reversed.

In the case of M_2, it also satisfies $(\epsilon = 1.01, \delta = 0.02)$-pbdp. To see why, first we note that a binary-valued randomized algorithm A here would be characterized by two numbers $x \equiv P(A(1) = 1)$ and $y \equiv P(A(0) = 1)$. Then $P(A(M_2(\mathcal{D}_1)) = 1) = y$ and $P(A(M_2(\mathcal{D}_2)) = 1) = 0.99y + 0.01x$, so if $y = P(A(M_2(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta$ then $P(A(M_2(\mathcal{D}_2)) = 1) \leq 0.99e^{-\epsilon}\delta + 0.01x \leq \delta$ when $\epsilon = 1.01$ and $\delta = 0.02$ (no matter what x is). Furthermore, $P(A(M_2(\mathcal{D}_2)) = 1) \geq 0.99P(A(M_2(\mathcal{D}_1)) = 1)$ so if $P(A(M_2(\mathcal{D}_2)) = 1) \leq e^{-\epsilon}\delta$ then $P(A(M_2(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta/0.99 \leq \delta$ for our choice of ϵ and δ. Hence M_2 also satisfies the privacy definition with these parameters.

Now, consider algorithm M^* that runs M_1 with probability $1/2$ and otherwise runs M_2. Consider the postprocessing algorithm A such that $A(1) = 1$ with probability 1, $A(-2) = 1$ with probability 0.05 and in all other settings, A outputs 0. Then:

$$\begin{aligned} P(A(M^*(\mathcal{D}_1)) = 1) &= \frac{1}{2}P(M_1(\mathcal{D}_1) = -2)P(A(-2) = 1) + \frac{1}{2}P(M_2(\mathcal{D}_1) = 1)P(A(1) = 1) \\ &= \frac{1}{2} * \frac{1}{1+e} * 0.05 < 0.0068 < e^{-1.01} * 0.02 = e^{-\epsilon}\delta \end{aligned}$$

$$P(A(M^*(\mathcal{D}_2)) = 1) = \frac{1}{2}P(M_1(\mathcal{D}_2) = -2)P(A(-2) = 1) + \frac{1}{2}P(M_2(\mathcal{D}_2) = 1)P(A(1) = 1)$$
$$= \frac{1}{2} * \frac{e}{1+e} * 0.05 + \frac{1}{2} * 0.01 > 0.023 > \delta$$

and so M^* does not satisfy $(\epsilon = 1.01, \delta = 0.02)$-pbdp.

Convexity of the privacy definition can be fixed by considering a curve, which we can do by introducing a function f. Then the conditions of Definition 5.5 can be turned into the conditions:

$$\forall a: \; P(A(M*(\mathcal{D}_1)) = 1) \leq a \Rightarrow P(A(M*(\mathcal{D}_2)) = 1) \leq 1 - f(a)$$

Note that $a \leq 1 - f(a)$ since that also occurs in Definition 5.5 because $e^{-\epsilon}\delta \leq \delta$. This type of generic constraint was first studied by Kifer and Lin [45, Theorem 2.1.3] who showed that for postprocessing invariance and convexity, a necessary condition on $1 - f$ is that it is concave and non-decreasing (so f is convex and non-increasing). Throwing in continuity at 0 for f (continuity elsewhere is guaranteed by convexity and monotonicity), this becomes equivalent to f-DP, recently proposed by Dong et al. [24], which can be defined as:

Definition 5.9 (f-DP [24])**.** *Let $f : [0,1] \to [0,1]$ be a continuous, convex, non-increasing function such that $f(x) \leq 1-x$. A mechanism M satisfies f-DP if for pairs of neighboring datasets $\mathcal{D}_1, \mathcal{D}_2$ and all binary-valued post-processing algorithms A (whose domain contains the range of M),*

$$1 - f(P(A(M(\mathcal{D}_1)) = 1)) \geq P(A(M(\mathcal{D}_2)) = 1).$$

The privacy parameter for f-DP is this function f (selecting it is an area of current research [24]). It is known as a trade-off function [24], a name that comes from the interpretation of A as a hypothesis test of H_0 : the input is $\mathcal{D}_1$ vs. H_1 : the input is $\mathcal{D}_2$, performed on the output of M. Here A outputs 1 to reject the null hypothesis and 0 to fail to reject. Under this interpretation, the expression $P(A(M(\mathcal{D}_1)) = 1)$ is the probability of a Type I error (significance level) and $P(A(M(\mathcal{D}_2)) = 1)$ is the power of the test. The function f provides the trade-off of the maximum power one can achieve for a given significance level.

From such a function f, we can recover the ϵ and δ parameters as follows. For a chosen δ, the corresponding ϵ is $\log \frac{\delta}{f^{-1}(1-\delta)}$, where, to account for the possibility that f is not one-to-one, f^{-1} is defined[8] as in [24] (i.e., $f^{-1}(z) = \inf\{y \in [0,1] \; : \; f(y) \leq z\}$). The continuity and non-increasing property of f implies that $f(f^{-1}(z)) \geq z$ with equality when f is invertible.

The (ϵ, δ) parameter conversion works because $P(A(M(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta \equiv f^{-1}(1-\delta)$ implies $P(A(M(\mathcal{D}_2)) = 1) \leq 1 - f(P(A(M(\mathcal{D}_1)) = 1)) \leq 1 - f(e^{-\epsilon}\delta) = 1 - f(f^{-1}(1-\delta)) \leq \delta$. This (ϵ, δ) curve also has a Bayesian interpretation that we discuss in Section 7.

Remark 5.10. *f-DP is postprocessing invariant and the adaptive composition results of Dong et al. 5.9 show that it is also convex. Hence, aside from continuity at 0, the conditions on f are both necessary and sufficient (this also follows from [45, Theorem 2.1.4]). The function f can even be replaced with the function $\max(f, f^{-1})$ without altering the privacy definition [24]. Furthermore, Dong et al. [24] showed that for every $\mathcal{D}_1, \mathcal{D}_2$ there exists a mechanism M and a worst-case post-processing algorithm A for which the inequality in Definition 5.9 is satisfied with equality.*

Remark 5.11. *Following an observation of Desfontaines and Pejó [22], one can use the convexity of f to eliminate A from Definition 5.9 and require, for all measurable sets S, $1 - f(P(M(\mathcal{D}_1)) \in S) \geq P(M(\mathcal{D}_2) \in S)$. The definition then becomes very similar to a previously introduced convex, post-processing invariant, abstract (and harder to interpret) version differential privacy [45].*

It is interesting to note that the chosen direction for resolving an interpretability issue with approximate differential privacy has f-DP as its consequence.

[8] Essentially, recalling that f is non-increasing, $f^{-1}(z)$ selects the smallest pre-image of z

5.2.1 Computing the privacy parameters

Computing the (ϵ, δ)-pbdp parameters of a mechanism M (or, equivalently, computing the trade-off function f) is not always easy. Here we discuss how to do this for the Gaussian Mechanism and in Section 5.3 we explain how to use RDP and zCDP to obtain a conservative bound for more complex mechanisms.

Example 5.12 (Gaussian Mechanism and Gaussian Differential Privacy [24])**.** *The Gaussian Mechanism [32, 11, 64, 24] can be applied more generally than was done in Example 4.5. Let q be a function whose input is a dataset and whose output is a vector (i.e., a collection of query answers). Suppose multivariate Gaussian noise with a* diagonal *covariance matrix Σ is added to the query answers. That is, the mechanism is $M(\mathcal{D}) = q(\mathcal{D}) + N(\vec{0}, \Sigma)$. This gives us a noisy vector as the output. The privacy loss random variable for component i of this vector, denoted $\mathcal{L}^{q_i}_{\mathcal{D}_1,\mathcal{D}_2}$, can be shown to have the distribution $N(\frac{(q(\mathcal{D}_1)_i - q(\mathcal{D}_2)_i)^2}{2\sigma_i^2}, \frac{(q(\mathcal{D}_1)_i - q(\mathcal{D}_2)_i)^2}{\sigma_i^2})$, where the σ_i^2 are the diagonal elements of Σ. By composing the privacy-loss random variables (see Example 4.6), the overall privacy-loss random variable has the univariate Gaussian distribution $N(\frac{\mu_{1,2}^2}{2}, \mu_{1,2}^2)$, where $\mu_{1,2}^2 = (q(\mathcal{D}_1) - q(\mathcal{D}_2))^T \Sigma^{-1} (q(\mathcal{D}_1) - q(\mathcal{D}_2))$. Let μ be the supremum of $\mu_{1,2}$, taken over all pairs of neighboring datasets $\mathcal{D}_1$,$\mathcal{D}_2$. Then:*

- *The exact (ϵ, δ)-differential privacy parameters can be computed as: $\delta = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^{\epsilon}\Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right)$ [11, 24], where Φ is the CDF of the standard normal distribution.*
- *The trade-off function f is $f(x) = \Phi(\Phi^{-1}(1-x) - \mu)$. When such a trade-off function is used with f-DP, one calls it* Gaussian Differential Privacy*.*
- *The (ϵ, δ)-pbdp parameters can then be computed as:*

$$\epsilon = \log \frac{\delta}{f^{-1}(1-\delta)} = \log \frac{\delta}{1 - \Phi(\Phi^{-1}(1-\delta) + \mu)} = \log \frac{\delta}{\Phi(-\Phi^{-1}(1-\delta) - \mu)}.$$

This result can be extended to covariance matrices Σ that are not diagonal by using standard techniques that diagonalize general Gaussian distributions (see, for example, [64]).

For other mechanisms, such as the Discrete Gaussian Mechanism used by the TopDown Algorithm in the 2020 Census Disclosure Avoidance System [2], it is very difficult to compute the exact (ϵ, δ) curves. Instead, we use zCDP accounting to conservatively approximate (upper-bound) the privacy parameters for pbdp (see Section 5.3).

For the purposes of comparison, consider the hypothetical situation in which the (continuous) Gaussian Mechanism is used in the production settings of the 2020 redistricting data. The μ parameter (of Example 5.12) would have been $\sqrt{2\rho}$ (where $\rho = 2.63$ is the final production ρ-zCDP parameter for the redistricting application of the TopDown Algorithm) [2]. Under this setting, Figure 1 plots the (ϵ, δ)-curves for approximate differential privacy (which is difficult to interpret) and the corresponding curve for pbdp/f-DP (which has a more natural interpretation). It also plots the upper bound on the pbdp parameter δ using ρ-zCDP privacy-loss accounting. We believe, based on the results in Section 6, that the pbdp curve of the *Discrete* Gaussian Mechanism is almost identical to that of the continuous Gaussian Mechanism (instead of being close to the ρ-zCDP upper bound).

We conclude this part of the discussion with an open question: how does one choose a trade-off function f or an (ϵ, δ)-curve? We do not have a definitive answer to this question, but the approach taken by the Disclosure Avoidance Team for the 2020 Census was to use the curve that was implicitly provided by the ρ parameter of zCDP and to base recommendations on the semantic consequences of different ρ-values (discussed in Sections 5.3, 6, 7, and 8). Afterwards, we derived these tighter guarantees based on pbdp.

5.3 Zero-Concentrated and Rényi Differential Privacy

Concentrated differential privacy was first introduced by Dwork and Rothblum [33] and then refined by Bun and Steinke [14] using Rényi divergences to create what is now called zero-Concentrated Differential Privacy (zCDP). Rényi Differential Privacy (RDP), featuring similar ideas, was developed nearly concurrently by Mironov [51].

RDP and zCDP have several important properties:

- they have very straightforward adaptive composition rules,
- they allow one to compute conservative privacy parameters for other definitions in situations where exact privacy parameter computation may be intractable (e.g., the Discrete Gaussian Mechanism),
- they are instrumental in deriving the Bayesian semantic guarantees in Section 7 (those Bayesian guarantees look remarkably similar to pbdp).

However, their privacy parameters are not very interpretable. This leads us naturally to the definitions that follow.

Definition 5.13 (α-Rényi divergence)**.** *The α-Rényi divergence of discrete distribution Q_1 from discrete distribution Q_2, both of whose supports are contained in a set Ω, is*

$$\mathfrak{D}_\alpha(Q_1||Q_2) = \frac{1}{\alpha-1}\log\sum_{x\in\Omega} Q_1(x)\frac{Q_1(x)^{\alpha-1}}{Q_2(x)^{\alpha-1}},$$

with summation replaced by integration in the continuous case. Note that the definition is not symmetric for the two input distributions.

(α,γ)-Rényi Differential Privacy[9] is simply a bound on the Rényi divergence for a specific α:

Definition 5.14 ((α,γ)-Rényi Differential Privacy [51])**.** *Given an $\alpha > 1$ and $\gamma \geq 0$, a mechanism M satisfies (α,γ)-RDP if for all pairs of neighboring datasets $\mathcal{D}_1, \mathcal{D}_2$, the α-Rényi divergence of the output distribution $M(\mathcal{D}_1)$ from the distribution $M(\mathcal{D}_2)$ is at most γ (or, equivalently,$E[e^{(\alpha-1)\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}}] \leq e^{(\alpha-1)\gamma}$).*[10]

A mechanism can satisfy RDP for many (α,γ) pairs and so, in practice, one keeps track of a set of such pairs and then converts them into more interpretable privacy parameters [1] (we will explain this process for pbdp).

The idea behind zero-Concentrated Differential Privacy is that for the Gaussian mechanism M in Example 4.5 with variance σ^2, the α-Rényi divergence of the output distribution $M(\mathcal{D}_1)$ from the distribution $M(\mathcal{D}_2)$ equals $\frac{\alpha}{2\sigma^2}$. Therefore, the definition requires that the α-Rényi divergence be proportional to α for all $\alpha > 1$:

Definition 5.15 (ρ-zero-Concentrated Differential Privacy [14])**.** *A mechanism M satisfies ρ-zCDP if for all pairs of neighboring datasets $\mathcal{D}_1, \mathcal{D}_2$ and all $\alpha > 1$, the α-Rényi divergence between the output distributions of $M(\mathcal{D}_1)$ and $M(\mathcal{D}_2)$ is at most $\alpha\rho$ (or, equivalently, $E[e^{(\alpha-1)\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}}] \leq e^{\alpha(\alpha-1)\rho}$).*

Clearly satisfying ρ-zCDP is equivalent to satisfying $(\alpha,\alpha\rho)$-RDP for all $\alpha > 1$. Both ρ-zCDP and (α,γ)-RDP are invariant under post-processing [14, 51]. Furthermore, they satisfy fully adaptive composition [36]. That is, in the case of RDP, for a fixed α, the γ values of different mechanisms add up linearly. In the case of zCDP, the ρ values add up linearly (just like the ϵ parameters in pure differential privacy). The Rényi divergences for many different mechanisms have already been computed [14, 51]. Both definitions provide a straightforward privacy-loss accounting framework.

RDP and zCDP can be used to obtain upper bounds on the (ϵ,δ) curve of approximate differential privacy [16, 8]. Existing results can even be leveraged to produce upper bounds on the (ϵ,δ)-curve of pbdp. Specifically, we know that for any point (ϵ^*,δ^*) on the (ϵ,δ) curve for *pbdp* of a mechanism M, the value of δ^* is the smallest post-processing invariant quantity such that:

$$P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} > e^{\epsilon^*}) \leq \delta^*.$$

Bun and Steinke (Lemma 3.5, [14]) proved that:

- if M satisfies (α,γ)-RDP with $\alpha > 1$, then for $\epsilon > \gamma$, $P(\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2} > e^{\epsilon}) \leq e^{(\alpha-1)(\gamma-\epsilon)}$;

[9] In the original paper [51], the privacy parameters were α and ϵ, but to avoid confusion with the ϵ in pure differential privacy, we use the symbol γ instead.

[10] Since $E[e^{(\alpha-1)\mathcal{L}^M_{\mathcal{D}_1,\mathcal{D}_2}}] = \sum_\omega P(M(\mathcal{D}_1)=\omega)\exp\left((\alpha-1)\log\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}\right) = \sum_\omega P(M(\mathcal{D}_1)=\omega)\left(\frac{P(M(\mathcal{D}_1)=\omega)}{P(M(\mathcal{D}_2)=\omega)}\right)^{\alpha-1}$ in the discrete case.

- if M satisfies ρ-zCDP, then for $\epsilon > \rho$, $P(\mathcal{L}^{M}_{\mathcal{D}_1,\mathcal{D}_2} > e^{\epsilon}) \leq e^{-(\epsilon-\rho)^2/(4\rho)}$.

Both of these upper bounds are functions of the privacy parameters and are post-processing invariant – post-processing can only decrease ρ and γ (while α stays the same), resulting in smaller probabilities. In other words, post-processing cannot increase the estimates of the tail probabilities. Therefore, they are both upper bounds on the δ parameter in *pbdp*. Thus, if we know that M satisfies (α, γ)-RDP for several different (α, γ) pairs, we can compute the upper bound using each pair and take the smallest one. Alternatively, if we know M satisfies ρ-zCDP, we can use the zCDP conversion to get an upper bound on δ.

Because of these results, both zCDP and RDP clearly provide a great deal of convenience for developing privacy-loss accounting frameworks, but it should be noted that some precision gets lost in the conversion to the privacy parameters of other definitions. For example, Figure 1 shows the (ϵ, δ) curve for pbdp of the Gaussian mechanism (blue) and the upper bound on the curve obtained from the zCDP conversion formula (black).

6 Frequentist Hypothesis Testing Semantics

The hypothesis testing interpretation of differential privacy is based on how well an attacker could succeed in the following experiment: $\mathcal{D}_1$ and $\mathcal{D}_2$ are arbitrary datasets that differ on the contents of the record of one individual (say, Jessie). One of them is selected as an input to M and the output ω is provided to the attacker. The success of an attacker in deciding between $\mathcal{D}_1$ and $\mathcal{D}_2$ is then equivalent to the success of the attacker in inferring a piece of sensitive information about Jessie (success strictly due to the use of Jessie's record, since everything else is held the same). While a hypothesis test does not directly quantify the causal contribution of the use of Jessie's record to make inferences about Jessie (i.e., it doesn't tell us what the inference about Jessie would have been if the record had been scrubbed), it does measure a related quantity – to what extent could an attacker determine that the record had been scrubbed.

In the frequentist hypothesis testing framework, one dataset would be designated as the null hypothesis: $H_0 = \mathcal{D}_1$ and the other as the alternative hypothesis: $H_1 = \mathcal{D}_2$. Due to symmetry in the differential privacy definitions, the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ can also be switched.

The uniformly most powerful test in this case is the likelihood ratio test [53]. One would compare the likelihood ratio test statistic $\log \frac{P(M(\mathcal{D}_1))=\omega)}{P(M(\mathcal{D}_2)=\omega)}$ to a threshold t. If the test statistic is smaller, one rejects the null hypothesis. If it is larger, one fails to reject the null hypothesis, and if the test statistic is equal to t, then one can randomize the decision rule by rejecting the null hypothesis with probability $c \in [0, 1]$.

The important properties of this (or any other) statistical test are:

- **Power**, denoted by $1 - \beta$, which is the probability of correctly rejecting the null hypothesis (i.e., rejecting the null hypothesis when the true input to M was $\mathcal{D}_2$).
- **Type II error probability**, denoted by β, is the probability of failing to reject the null hypothesis when the alternative hypothesis is true.
- **Significance level**, denoted by ℓ is the probability of incorrectly rejecting the null hypothesis (i.e., rejecting the null hypothesis when the true input to M was $\mathcal{D}_1$). Significance level is also known as the probability of a Type I error.

We note, from Section 4, that the distribution of the privacy-loss random variable $\mathcal{L}^{M}_{\mathcal{D}_1,\mathcal{D}_2}$ has the distribution of the likelihood ratio test statistic under the null hypothesis. Similarly, the distribution of $-\mathcal{L}^{M}_{\mathcal{D}_2,\mathcal{D}_1}$ has the corresponding distribution under the alternate hypothesis. Thus one can write expressions for the power $1 - \beta$ and significance level ℓ of the likelihood ratio test that uses the threshold t and tie-breaking probability c as follows:

$$\ell = cP(\mathcal{L}^{M}_{\mathcal{D}_1,\mathcal{D}_2} \leq t) + (1 - c)P(\mathcal{L}^{M}_{\mathcal{D}_1,\mathcal{D}_2} < t)$$
$$1 - \beta = cP(\mathcal{L}^{M}_{\mathcal{D}_2,\mathcal{D}_1} \geq -t) + (1 - c)P(\mathcal{L}^{M}_{\mathcal{D}_2,\mathcal{D}_1} > -t)$$

This discussion shows that there is a deep connection between differential privacy and hypothesis testing. Each definition provides a trade-off between the power and significance level of *any* hypothesis test of $H_0 : \mathcal{D}_1$

vs. $H_1 : \mathcal{D}_2$ that is based on the output of a mechanism satisfying the privacy definition. Furthermore, the trade-off can be directly computed from the privacy parameters. That is, the power for any significance level ℓ for (say) pure ϵ-differential is upper bounded by a function of ℓ and ϵ. This upper bound applies to any M that satisfies ϵ-differential privacy. However, some of these mechanisms might have a tighter trade-off (i.e., lower power at a given significance level). Computing that tighter tradeoff for a specific mechanism M would require reasoning about the privacy loss random variables $\mathcal{L}^{M}_{\mathcal{D}_1,\mathcal{D}_2}$ for all neighbors $\mathcal{D}_1, \mathcal{D}_2$.

6.1 Frequentist semantics of pure differential privacy

One can think of a hypothesis test as a randomized algorithm A whose input is ω and whose output is either 1 (reject the null hypothesis) or 0 (fail to reject). Thus the significance level is $\ell = P(A(M(\mathcal{D}_1)) = 1)$ and the power is $1 - \beta = P(A(M(\mathcal{D}_2)) = 1)$. If M satisfies ϵ-differential privacy then so does $A \circ M$ and so the pure differential privacy constraints apply to both $P(A(M(\mathcal{D}_1)) = 1)$ and $P(A(M(\mathcal{D}_2)) = 1)$. Noting the symmetry of $\mathcal{D}_1$ and $\mathcal{D}_2$ in the definition of differential privacy, and noting that if A is a hypothesis test, then so is $1 - A$, then we obtain the following result due to Wasserman and Zhou [63]:

Theorem 6.1 ([63])**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be neighboring datasets, let M be an ϵ-differential privacy mechanism, and let ω denote the output of M. Then, any hypothesis test of $H_0 = \mathcal{D}_1$ vs. $H_1 = \mathcal{D}_2$ that is a function of ω has the following relation between power $1 - \beta$, Type II error probability β and significance level ℓ:*

$$\begin{aligned} \ell &\leq e^{\epsilon}(1-\beta) \\ 1-\beta &\leq e^{\epsilon}\ell \\ (1-\ell) &\leq e^{\epsilon}\beta \\ \beta &\leq e^{\epsilon}(1-\ell) \end{aligned}$$

Combining these inequalities, we get the following explicit limitations on power at a given significance level for any hypothesis testing procedure:

$$\max(e^{-\epsilon}\ell, 1 - e^{\epsilon}(1-\ell)) \leq 1-\beta \leq \min(e^{\epsilon}\ell, 1 - e^{-\epsilon}(1-\ell)). \tag{4}$$

We note that a non-informative test that cannot distinguish between H_0 and H_1 at all would be rejecting the null hypothesis at the same rate regardless of the truth of H_0 or H_1; the power and significance levels would always be equal to each other in such a test. From Equation 4, we see that as ϵ approaches 0, the power approaches the significance level. When $\epsilon = 0$ only the non-informative tests are possible. Table 1 uses the power-bounds formula above to plot the maximum achievable power in inferring personal information for a given significance level and ϵ parameter.

	ϵ				
Significance Level	0.1	0.5	1	2	4
0.01	0.011	0.016	0.027	0.074	0.550
0.05	0.055	0.820	0.136	0.370	0.983
0.10	0.111	0.165	0.272	0.739	0.984

Table 1: Maximum power for inferring personal information under pure differential privacy at a given significance level, for different values of ϵ.

We note that at $\epsilon = 4$, these frequentist guarantees do not at first appear meaningful because of the large achievable power. However, this is the power that is achieved by mechanisms that use all of their privacy-loss budget to try to reveal Jessie's personal information. Since most differentially private algorithms would spread a large privacy-loss budget over many queries, it is often possible to use additional properties of the mechanism to reduce these upper bounds. For example, one can consider the entire (ϵ, δ) curve of a mechanism under f-DP/pbdp, which we demonstrate later in this section. Another approach is to apply the ideas developed in this section while examining how the algorithm allocates privacy-loss budget among queries, which we consider in Section 8.

6.2 Frequentist Semantics of Approximate Differential Privacy

To get frequentist guarantees for approximate differential privacy, Kairouz et al. [42] used ideas similar to those used for pure differential privacy to get the following result:

Theorem 6.2 ([42]). *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be neighboring datasets, let M be an (ϵ, δ)-differential privacy mechanism, and let ω denote the output of M. Then, any hypothesis test of $H_0 = \mathcal{D}_1$ vs. $H_1 = \mathcal{D}_2$ that is a function of ω has the following relation between power $1-\beta$, Type II error probability β, and significance level ℓ:*

$$\begin{aligned} \ell &\leq e^{\epsilon}(1-\beta) + \delta \\ 1-\beta &\leq e^{\epsilon}\ell + \delta \\ (1-\ell) &\leq e^{\epsilon}\beta + \delta \\ \beta &\leq e^{\epsilon}(1-\ell) + \delta. \end{aligned}$$

Combining these inequalities, we get the following explicit limitations on power at a given significance level for any hypothesis testing procedure:

$$\max(e^{-\epsilon}(\ell-\delta), 1-e^{\epsilon}(1-\ell)-\delta) \leq 1-\beta \leq \min(e^{\epsilon}\ell+\delta, 1-e^{-\epsilon}(1-\ell-\delta)).$$

However, we note that a mechanism M satisfies approximate differential privacy for a curve C of (ϵ, δ) values, and so the upper bound on the power is actually $\min_{(\epsilon,\delta)\in C} \min(e^{\epsilon}\ell+\delta, 1-e^{-\epsilon}(1-\ell-\delta))$. This is another reason to use the entire (ϵ, δ) curve rather than rely on a single point and the ensuing inaccurate summary of the privacy protections.

6.3 Frequentist semantics of pbdp, f-DP, RDP, and zCDP

Frequentist semantics for the remaining privacy definitions can be derived in the same manner as those of pure and approximate DP. We group them here so that we can compare them using the production settings of the 2020 Census redistricting data.

We start with the significance level/power trade-offs for zCDP and RDP, which were studied by Balle et al. [10]. Then, we follow the reasoning used for pure and approximate DP, combined with the definition of the Rényi divergence to complete the semantics.

Theorem 6.3 ([10]). *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be neighboring datasets, let M satisfy (α, γ)-RDP. Let ω denote the output of M. Then, any hypothesis test of $H_0 = \mathcal{D}_1$ vs. $H_1 = \mathcal{D}_2$ that is a function of ω has the following relation between power, $1-\beta$, Type II error probability, β, and significance level, ℓ:*

$$\begin{aligned} \ell^{\alpha}(1-\beta)^{1-\alpha} + (1-\ell)^{\alpha}\beta^{1-\alpha} &\leq e^{\gamma(\alpha-1)} \\ (1-\beta)^{\alpha}\ell^{1-\alpha} + \beta^{\alpha}(1-\ell)^{1-\alpha} &\leq e^{\gamma(\alpha-1)}. \end{aligned}$$

Since satisfying ρ-zCDP is equivalent to satisfying $(\alpha, \rho\alpha)$-RDP for all $\alpha > 1$, the relations for ρ-zCDP are:

$$\begin{aligned} \ell^{\alpha}(1-\beta)^{1-\alpha} + (1-\ell)^{\alpha}\beta^{1-\alpha} &\leq e^{\rho\alpha(\alpha-1)} \quad \text{for all } \alpha > 1 \\ (1-\beta)^{\alpha}\ell^{1-\alpha} + \beta^{\alpha}(1-\ell)^{1-\alpha} &\leq e^{\rho\alpha(\alpha-1)} \quad \text{for all } \alpha > 1. \end{aligned}$$

Because the constraints are nonlinear, it is not possible to present an exact algebraic bound on the power as a function of level and ρ (or γ and α). However, an upper bound on power can be estimated numerically by considering a large number of α values and finding points on the boundaries of the constraints.

The noise-generating component of the disclosure avoidance system for the 2020 redistricting data satisfies ρ-zCDP with $\rho = 2.63$ [2]. In Figure 2 we plot the numerically estimated upper bound on power for each significance level, and in Table 2 we list the upper bounds for a selection of significance levels.

We note that these upper bounds on power appear quite high, but this is because the upper bounds are guarantees for *all* possible mechanisms that satisfy ρ-zCDP for $\rho = 2.63$ (not just the TopDown Algorithm). In particular, this guarantee would also bind mechanisms that specifically target Jessie and use their entire

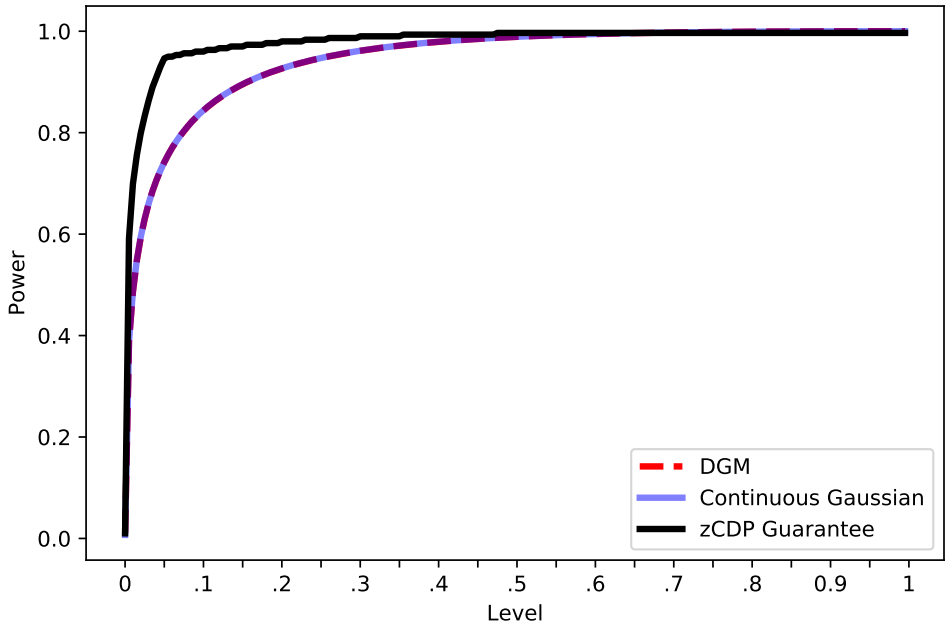

Figure 2: Level (x-axis) vs. power (y-axis) curves for (1) the Gaussian mechanism, which is the same as the test of $H_0 = N(0, \frac{1}{2\rho})$ vs. $H_1 = N(1, \frac{1}{2\rho})$ for $\rho = 2.63$, (2) the likelihood ratio test of the discrete Gaussian noisy queries having the same privacy-loss budget allocation as in the production settings for 2020 Census redistricting data.

.

Significance Level	Power (Gaussian)	Power (DGM)	zCDP Upper Bound
0.01	0.49	0.49	0.70
0.05	0.74	0.74	0.95
0.10	0.84	0.84	0.96

Table 2: Likelihood ratio test significance level/power trade-off for 2020 Census redistricting data production privacy-loss budget allocations (1) if Gaussian noise is used, (2) if discrete Gaussian noise is used, (3) guaranteed upper bound if an arbitrary ρ-zCDP mechanism with $\rho = 2.63$ is used.

privacy-loss budget in an attempt to reveal Jessie's personal information. Furthermore, even though ρ-zCDP can produce an entire (ϵ, δ) curve for f-DP/pbdp, the number of curves one can obtain by changing ρ is limited and hence may not be the tightest representation of a mechanism.

A more precise analysis and tighter bounds can be achieved by considering f-DP (for the purposes of this section, the reparametrization offered by pbdp does not affect the frequentist semantics). The guarantees can be read directly off Definition 5.9 by noting that the randomized algorithm A it uses can be viewed as a hypothesis test (in fact, f-DP was originally introduced directly as the trade-off between the Type I and Type II error probabilities [24]). Thus we get:

$$1 - \beta \leq 1 - f(\ell)$$
$$\ell \leq 1 - f(1 - \beta)$$

In the case of the Gaussian Mechanism (Example 5.12), the function f is known exactly and so we can compare its trade-off with the upper bound provided by zCDP as follows. Let us consider some query q such that $q(\mathcal{D}_1)$ differs from $q(\mathcal{D}_2)$ by at most 1 for any pair of neighboring $\mathcal{D}_1$, $\mathcal{D}_2$. It is known that the mechanism M defined as $M(\mathcal{D}) = q(\mathcal{D}) + N(0, \sigma^2 = \frac{1}{2\rho})$ satisfies ρ-zCDP [14] and has the trade-off function given in Example 5.12. We plot the upper bound on power for each significance level for this mechanism

in Figure 2 (blue line) and list some of the upper bounds for a selection of significance levels in Table 2 in the column labeled "Power (Gaussian)". We note that the f-DP trade-off function is the exact trade-off one would get from the likelihood ratio test [24]. We see that the actual upper bound on power for the continuous Gaussian is less than what zCDP semantics provide for $\rho = 2.63$.

But what about the actual TopDown Algorithm that produced redistricting data? Its noise addition mechanism is analogous to the Gaussian Mechanism of Example 5.12 but it adds discrete Gaussian (instead of continuous Gaussian) noise to millions of queries. The details of how noise was allocated to different queries appear in [2] and are reproduced in Section 8. The trade-off function f for the composition of discrete Gaussian mechanisms is not known, but can be estimated using Monte Carlo simulation and the following observations:

- the maximum possible power for a given significance level is given by the maximum power of the likelihood ratio test (maximized over all pairs of neighboring datasets);
- one can, without loss of generality, examine only pairs of neighbors $\mathcal{D}_1$, $\mathcal{D}_2$ for which the maximal number of queries have different answers;
- for the TopDown algorithm, it turns out that all of these maximal neighboring pairs are equivalent for this purpose – the likelihood ratio test gives exactly the same power for a given significance level.

We generated one million samples from the null and one million samples from the alternate hypothesis to empirically estimate the significance level/power trade-off. The results are shown as the dashed red line in Figure 2 and in the column labeled "Power (DGM)" in Table 2. We see that the results are nearly identical to the continuous Gaussian. Thus, this significance level/power curve represents one of the interpretations of the privacy guarantees of the TopDown Algorithm. We examine Bayesian guarantees in Section 7. We also consider fine-grained privacy semantics that examine the noise added to different queries in Section 8.

6.4 A Generic Extension of Frequentist Semantics

While the previous results focussed on significance/level power curves for distinguishing between pairs of neighbors $\mathcal{D}_1$ and $\mathcal{D}_2$, it is possible to extend the mathematical statements of all of the results to something that more closely matches the informal description of "can an attacker detect that a record has been altered"?

Let $\mathcal{D}^-$ refer to the set of records belonging to everyone other than Jessie and let F be an *arbitrary* distribution for $\mathcal{D}^-$ (e.g., those records do not have to be independent). Given F and two possible candidate records r_0, r_1 for Jessie, we can form the hypothesis test of H_0: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_0\}$ is the input to M vs. H_1: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_1\}$ is the input to M.

We show in this section that if a privacy definition is post-processing invariant, convex (see Section 4.2), and also satisfies a minor regularity condition called *symmetry* then the significance level vs. power trade-off curve is the same as for the earlier cases where one considered pairs of neighboring datasets. Since all the privacy definitions we considered have these properties, this extension applies to them all.

The symmetry property can be described follows. Let ν be a 1-to-1 and onto function on records that does not alter the id (but the behavior of ν can depend on the id). For example, μ might toggle the ethnicity attribute for Alice, toggle both ethnicity and gender for Jessie, add 5 to Bob's age, etc. One can apply ν to an entire dataset by defining $\nu(\mathcal{D})$ to be the result of applying ν to each record in $\mathcal{D}$. We say that a privacy definition has symmetry if for all mechanisms M that satisfy the privacy definition and all such functions ν, then M and $M \circ \nu$ both satisfy the privacy definition for exactly the same privacy parameter settings (note that $M \circ \nu$ is the function that first applies the recoding ν and then runs M on the result). This symmetry property essentially states that no dataset gets extra protections compared to another dataset. Furthermore, these ν functions preserves the neighborhood structure of datasets: if $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbors then the pair $\nu(\mathcal{D}_1)$ and $\nu(\mathcal{D}_2)$ are neighbors and same with the pair $\nu^{-1}(\mathcal{D}_1)$ and $\nu^{-1}(\mathcal{D}_2)$. Symmetry allows us to restrict attention to just one (arbitrary) pair of neighbors $\mathcal{D}_1, \mathcal{D}_2$ when examining the privacy properties of a definition.

The first result in this section is that symmetric, convex, postprocessing invariant privacy definitions have a concave significance level vs. power trade-off. This is useful for proving the main result of this section and for diagnosing if a privacy definition might violate convexity.

Theorem 6.4. *Given a transparent, symmetric, convex and postprocessing invariant privacy definition with privacy parameters θ, let $\mathfrak{M}$ denote the set of all mechanisms that satisfy the privacy definition at those privacy parameter settings. Define the significance level vs. power trade-off function g as follows. For any significance level ℓ, $g(\ell)$ is the supremum of achievable power in testing $H_0 : \mathcal{D}_1$ vs. $H_1 : \mathcal{D}_2$ for all neighboring pairs $\mathcal{D}_1, \mathcal{D}_2$, for all mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level is at most ℓ. Then g is a concave function.*

Next is the main result of this section, which shows that generalizing the hypothesis testing setup does not change the significance level vs. power curve.

Theorem 6.5. *Given a transparent, symmetric, convex and postprocessing invariant privacy definition with privacy parameters θ, let $\mathfrak{M}$ denote the set of all mechanisms that satisfy the privacy definition at those privacy parameter settings.*

- *Define the* pointwise *significance level vs. power tradeoff function g as follows. For any significance level ℓ, $g(\ell)$ is the supremum of achievable power in testing $H_0 : \mathcal{D}_1$ vs. $H_1 : \mathcal{D}_2$ for all neighboring pairs $\mathcal{D}_1, \mathcal{D}_2$, for all mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level as at most ℓ.*
- *Define the* compound *significance level vs. power tradeoff function g^* as follows. For any significance level ℓ, $g^*(\ell)$ is the supremum of achievable power in testing H_0^*: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_0\}$ is the input to M vs. H_1^*: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_1\}$ is the input to M, maximizing over all distributions F, pairs of records r_1, r_2, mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level is at most ℓ.*

Then, $g = g^*$.

7 Bayesian Posterior-to-Posterior Semantics

In this section, we present the Bayesian privacy semantics for pure differential privacy, zCDP, and RDP. The semantics for zCDP and RDP are new and resemble those derivable from pbdp. We also provide a direct Bayesian interpretation of the (ϵ, δ) curve of pbdp. While some results about the Bayesian semantics for approximate differential privacy are known [43], we do not review the exact details here because they only apply to a small subset of privacy-loss parameters (ϵ must be smaller than 1/4) and do not use the full (ϵ, δ) curve.

We consider an attacker using a prior $P_{\mathfrak{A}}$ over possible datasets $\mathcal{D}_1, \mathcal{D}_2, \ldots$. This prior is independent of the randomness in M (since the attacker does not have access to the bits of M). Otherwise, we make no assumptions about this prior – in particular, records do not have to be independently and identically distributed (i.i.d.). Thus, in the attacker's view, the dataset is a random variable $\mathcal{X}$ having distribution $P_{\mathfrak{A}}$ (we often use notation such as $P_{\mathfrak{A}}(\mathcal{D})$ and $P_{\mathfrak{A}}(\cdot \mid \mathcal{D})$ as shorthand for $P_{\mathfrak{A}}(\mathcal{X} = \mathcal{D})$ and the conditional distribution $P_{\mathfrak{A}}(\cdot \mid \mathcal{X} = \mathcal{D})$). A mechanism M operates on the data and produces an output ω (e.g., privacy-protected microdata or tabulations). This output ω along with all details of M, except for the specific values of the random bits it uses, are released publicly and are known to the attacker. The attacker is trying to infer information about the record belonging to a specific target individual Jessie (for records, we use the notation R to denote the random variable and r to represent a specific value). What can be guaranteed about the posterior $P_{\mathfrak{A}}(R \mid M(\mathcal{X}) = \omega)$? Before answering this question, it is important to note that we use $P_{\mathfrak{A}}$ for expressions in which the attacker's belief plays a role and P for probabilities in which only the randomness in M plays a role. In other words, $P(M(\mathcal{D}) = \omega)$ is the probability of ω when the input $\mathcal{D}$ is given and $P_{\mathfrak{A}}(M(\mathcal{X}) = \omega)$ is the probability of ω when the input $\mathcal{X}$ is also random.

7.1 The Unsuitability of Prior to Posterior Comparisons

Traditional disclosure avoidance methodology attempts to provide semantic guarantees based on comparisons between the prior $P_{\mathfrak{A}}(R)$ and posterior $P_{\mathfrak{A}}(R \mid M(\mathcal{X}) = \omega)$ distributions. However, in light of privacy research over the past two decades that studies the consequences of prior-posterior comparisons, the suitability of

these comparisons has been thrown into question. First, the difference between the prior and posterior can be arbitrarily large (rendering the comparison vacuous) unless assumptions are made about the prior [27, 30, 46]. Obtaining consensus about the right prior(s) to use will always remain an unsettled issue, and the approach will never be future-proof – knowledge learned in the future may result in a different prior from which a different semantic conclusion will be reached. Second, even more importantly, prior-posterior comparisons do not distinguish between changes that are due to gaining scientific knowledge (such as learning that smoking causes cancer) and changes that are specifically due to participation in the data [26]. In particular, prior-posterior comparisons may show large changes even for people whose information is not recorded in the data.

As discussed in Section 2, gaining generalizable scientific/statistical knowledge is the whole point of data release. Disclosure avoidance methodologies that penalize a data release for providing statistical knowledge run the risk of severely damaging the usefulness of privacy-protected data. However, an alternative approach [43], known as *posterior-posterior* comparisons provides a clean way of separating inference about individuals due to generalizable knowledge from inference that is possible specifically because an individual's record is in the input to M. Such posterior-posterior semantics are the focus of this section.

7.2 The Posterior-Posterior framework

The main idea behind the posterior-to-posterior framework is to compare (1) what an attacker would learn about Jessie if she participated in the data collection process (e.g., submitted survey responses) to (2) what would be learned about Jessie in a counterfactual world in which Jessie's data were not collected or not used. This counterfactual world represents the privacy-preserving gold standard for Jessie. Furthermore, unlike the frequentist setting, the attacker knows if we are operating in the actual world or the counterfactual world (since the goal is not to detect whether a change has occurred, but to directly compare how the contents of a record affect inference about it). On a technical note, the attacker must use the same prior in both worlds, as the hypothetical choice of the world would not depend on the records in the actual data.

There are several ways of formalizing the counterfactual world.

- **Option 1**: Drop Jessie's record from the dataset before running M. Although this is a very appealing option, it is not always applicable. For example, when using a differential privacy variant with *bounded neighbors*, the size n of the input to M is not allowed to change; therefore, the counterfactual world must respect that constraint and also use a dataset of size n.

- **Option 2**: Replace Jessie's record with a pre-determined default record [43]. This option creates a counterfactual world in which the *contents* of Jessie's survey responses are not used at all. There are many possible choices for default records, and the default record could even be sampled from a distribution that does not depend on Jessie's actual responses. There is one specific choice that is very appealing both mathematically and conceptually, and that is the basis of the next option.

- **Option 3**: Replace Jessie's record with a random sample from the posterior distribution the attacker would use, if the attacker knew the records of everyone else. This posterior distribution is arguably the statistical information that the attacker would learn from the rest of the data records. Hence, this option can be thought of as replacing Jessie's actual record with statistical information.[11] This is the option we adopt for the posterior-to-posterior semantics in this section.

We now formalize the posterior-to-posterior comparisons using Option 3 as the counterfactual world. For any dataset $\mathcal{D}$, let $\mathcal{D} \cup \{r\}$ be the result of adding record r to $\mathcal{D}$. Let $\mathcal{D}^-$ be the result of removing the record of the target person Jessie from $\mathcal{D}$. Let $\text{psample}[\mathcal{D}]$ represent the result of replacing Jessie's record with a random record sampled from $P_{\mathfrak{A}}(R \mid \mathcal{X}^- = \mathcal{D}^-)$, which is the same as creating $\mathcal{D}^- \cup \{r\}$ where $r \sim P_{\mathfrak{A}}(R \mid \mathcal{X}^- = \mathcal{D}^-)$.[12] Finally, let $\mathcal{U}$ represent the set of possible datasets (assumed discrete to simplify the discussion) that include survey information from Jessie, and let $\mathcal{U}^-$ represent the datasets in $\mathcal{U}$ after Jessie's record has been removed. The goal is to compare the posterior distribution in the actual world after the output ω of M has been observed (i.e., $P_{\mathfrak{A}}(R \mid M(\mathcal{X}) = \omega)$) to the corresponding posterior in the

[11] This is like saying that in the actual world, the record used for Jessie provides information about how Jessie differs from her community, but the substituted record in the counterfactual world does not.

[12] In other words, $\text{psample}[\mathcal{D}]$ is a random variable that equals $\mathcal{D}^- \cup \{r'\}$ with probability $P_{\mathfrak{A}}(r' \mid \mathcal{X}^- = \mathcal{D}^-)$.

counterfactual world: $P_{\mathfrak{A}}(R \mid M(\text{psample}[\mathcal{X}]) = \omega)$. The formulas for these posterior distributions are given below.

Actual World:

$$P_{\mathfrak{A}}(R = r \mid M(\mathcal{X}) = \omega) = \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}.$$

Counterfactual World:

$$\begin{aligned} P_{\mathfrak{A}}(R = r \mid M(\text{psample}(\mathcal{X})) = \omega) &= \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\text{psample}(\mathcal{D}^- \cup \{r\})) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\text{psample}(\mathcal{D}^- \cup \{r'\})) = \omega)} \\ &= \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big)} \\ &= \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) \sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega)}. \end{aligned}$$

We are interested in the ratio of the posterior in the actual world to the one from the counterfactual world, with a ratio close to one indicating that the two posteriors are similar. Noting that the denominators in both expressions are the same, the ratio of the posteriors is:

$$\frac{P_{\mathfrak{A}}(R = r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(R = r \mid M(\text{psample}(\mathcal{X})) = \omega)} = \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big)}.$$

The ratio of posteriors is a function of ω and can be viewed as the Bayesian analogue of the privacy-loss random variable. We next consider what can be said about this ratio of posteriors when M satisfies different versions of differential privacy.

7.3 Posterior-to-Posterior Semantics of Pure Differential Privacy

It is easy to bound the posterior-to-posterior ratio under pure differential privacy. For all priors and all ω, the ratio of actual to counterfactual posteriors is always between $e^{-\epsilon}$ and e^{ϵ}.

Theorem 7.1. *If M satisfies ϵ-differential privacy, then for any attacker prior $P_{\mathfrak{A}}$ and all ω and every r that has nonzero probability,*

$$e^{-\epsilon} \leq \frac{P_{\mathfrak{A}}(R = r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(R = r \mid M(\text{psample}(\mathcal{X})) = \omega)} \leq e^{\epsilon}.$$

It is worth noting that Theorem 7.1 is quite strong: no matter what prior the attacker uses (even if it is completely wrong), and no matter what output M produces, any posterior probability about Jessie (or any other person) is within a factor of e^{ϵ} compared to purely statistical inference about that person.[13]

Clearly, if an attacker's prior is completely unreasonable (e.g., all people have 3 arms) then any inference performed with such a prior is suspect. A large difference between an actual posterior and counterfactual posterior would only be a *perceived* privacy breach, not a real one. Pure differential privacy always provides protections even against perceived privacy breaches. However, this is not always the case for other variants of differential privacy. To see why, we adapt a well-known example discussed by Kasiviswanathan and Smith and credited to Dwork and McSherry [43].

[13] Prior work (e.g, [43]) and folklore results in which the counterfactual world is obtained by replacing Jessie's record with a pre-specified default record can achieve similar guarantees but with a factor of $e^{2\epsilon}$ instead of e^{ϵ}.

Example 7.2. *Suppose the true dataset $\mathcal{D}$ contains records from Jessie and 99 other individuals, with everyone in the data being Hispanic, but the attacker (incorrectly) believes that the other 99 individuals are not Hispanic, that Jessie's ethnicity is independent of the other people, and that $P($Jessie is Hispanic$) = 0.01$. Thus, in the attacker's view, there are only 2 possible datasets and*

$$P_{\mathfrak{A}}(\mathcal{X} = \textit{Only Jessie is Hispanic}) = 0.01$$
$$P_{\mathfrak{A}}(\mathcal{X} = \textit{No one is Hispanic}) = 0.99$$

Let M be a mechanism that adds $N(0,1)$ noise to the number of Hispanic individuals in the dataset (such a mechanism would satisfy approximate differential privacy, zCDP, RDP, and f-DP under appropriate parameter settings). Thus $M(\mathcal{D})$ outputs $100 + N(0,1)$. Suppose the realized value is $\omega = 101.3$.

In the actual world (i.e., Jessie's record is part of the input to M), the attacker believes that $N(0,1)$ was either added to 1 or to 0, resulting in the unlikely outcome 101.3 and the posterior inference:

$$P_{\mathfrak{A}}(\textit{Jessie is Hispanic} \mid \omega = 101.3) = \frac{0.01\frac{1}{\sqrt{2\pi}}e^{-(101.3-1)^2/2}}{0.01\frac{1}{\sqrt{2\pi}}e^{-(101.3-1)^2/2} + 0.99\frac{1}{\sqrt{2\pi}}e^{-(101.3-0)^2/2}}$$
$$= \frac{1}{1 + 99e^{-101.3+1/2}} \approx 1.$$

In the counterfactual world, Jessie's record would be replaced with a Hispanic record with probability 0.01 and a non-Hispanic record with probability 0.99. Simple computation shows that:

$$P_{\mathfrak{A}}(\textit{Jessie is Hispanic} \mid M(\text{psample}(\mathcal{X})) = \omega) = 0.01.$$

Intuitively, in the counterfactual world, ω only depends on the rest of the records, but the attacker believes the rest of the people are all non-Hispanic with probability 1 and so ω carries no new information. Thus, the ratio of the two posteriors is almost 100, and we could make it as large as we want simply by choosing a prior where Jessie has an even smaller probability of being Hispanic (regardless of the privacy parameters associated with M).

The key in Example 7.2 is to note that under the attacker's wildly incorrect prior, the output ω that caused this perceived privacy breach is virtually impossible and so, in such a situation, the attacker should question the correctness of his prior. Thus, Kasiviswanathan and Smith [43] recommend examining how posteriors in the actual and counterfactual worlds differ when the attacker's prior is correct (i.e., what is the probability of a large difference when ω is generated according to the randomness of M and the prior). This idea is the basis for our semantic results for zCDP and RDP.

7.4 Posterior-to-Posterior Semantics of RDP and zCDP

We consider two types of Bayesian posterior-to-posterior semantics for RDP and zCDP. In the first case, the attacker knows $\mathcal{D}^-$ (i.e., has complete information about everyone else in the dataset). In the second type of semantics, the attacker is uncertain about $\mathcal{D}^-$. This situation is more general, but the bounds we can prove are slightly weaker. In both settings, we get an (ϵ, δ) curve with the guarantee that for any (ϵ, δ) pair on that curve, the probability that the ratio of posteriors is at least e^ϵ is bounded by δ. Furthermore, this (ϵ, δ) curve is computed using Rényi divergences and, hence, has desirable properties under post-processing and composition.

7.4.1 Bayesian semantics when the rest of the data are known.

In this setting, the true dataset is $\mathcal{D}$ and the attacker knows $\mathcal{D}^-$, so that $P_{\mathfrak{A}}(\mathcal{X}^- = \mathcal{D}^-) = 1$. Under the assumption that the prior is correct, the observed output ω is a random variable whose distribution is a function of both $P_{\mathfrak{A}}$ and the randomness in M. The data generating model for ω in the actual world is:

1. sample $R \sim P_{\mathfrak{A}}(R \mid \mathcal{D}^-)$,
2. set $\mathcal{D} = \mathcal{D}^- \cup \{R\}$,

3. produce $\omega = M(\mathcal{D})$.

In the counterfactual world, $M(\mathcal{D})$ is replaced with $M(\text{psample}(\mathcal{D}))$.

In the actual world, the attacker's posterior belief about r, conditioned on ω, is:

$$P_{\mathfrak{A}}(R = r \mid \mathcal{X}^- = \mathcal{D}^-, M(\mathcal{X}) = \omega) = \frac{P_{\mathfrak{A}}(R = r \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}.$$

As noted earlier, it is important to remember that we use $P_{\mathfrak{A}}$ for expressions in which the attacker's belief plays a role and P for probabilities in which only the randomness in M plays a role. In the counterfactual world, the posterior is:

$$\begin{aligned} P_{\mathfrak{A}}(R = r \mid \mathcal{X}^- = \mathcal{D}^-, M(\text{psample}(\mathcal{X})) = \omega) &= \frac{P_{\mathfrak{A}}(R = r \mid \mathcal{D}^-)\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \\ &= P_{\mathfrak{A}}(R = r \mid \mathcal{D}^-), \end{aligned}$$

which is the same as the prior, since the attacker has complete certainty about the rest of the data. Thus the actual posterior to counterfactual posterior ratio turns out to be the same as the actual posterior to prior ratio and equals:

$$\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}.$$

We are interested in the probability of generating an ω for which this ratio is large. Note that the randomness in the numerator is only with respect to M.

Theorem 7.3. *Let $P_{\mathfrak{A}}$ be an attacker's prior such that $P_{\mathfrak{A}}(\mathcal{X}^- = \mathcal{D}^-) = 1$ for some $\mathcal{D}^-$. Let M be a mechanism and let P_ω be the distribution of the output ω of M under the assumption that the attacker's prior $P_{\mathfrak{A}}$ is correct (i.e., a record for the target individual is sampled $R \sim P_{\mathfrak{A}}(R \mid \mathcal{D}^-)$ and then $\omega = M(\mathcal{D}^- \cup \{R\})$ is produced). Then for any r,*

- *If M satisfies (α, γ)-RDP for $\alpha > 1$, then $P_\omega\left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \geq e^\epsilon\right) \leq e^{-(\epsilon-\gamma)\alpha-\gamma}$ (i.e., the probability of seeing an ω for which the posterior-to-posterior ratio exceeds e^ϵ is at most $\delta = e^{-(\epsilon-\gamma)\alpha-\gamma}$).*
- *If M satisfies ρ-zCDP and $\epsilon > \rho$, then $P_\omega\left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \geq e^\epsilon\right) \leq e^{-(\epsilon+\rho)^2/(4\rho)}$.*

We note that the probabilities that the ratios exceed e^ϵ only depend on M through the RDP privacy parameters (α, γ) or the zCDP privacy parameter ρ. Post-processing M cannot increase γ or ρ, and the upper bounds given in Theorem 7.3 are non-increasing in ρ and γ, thus making these post-processing invariant upper bounds. In other words, we can replace M with $A \circ M$, for any A whose domain contains the range of M, and the bounds will still hold. The Bayesian (ϵ, δ) curve for the semantics provided by Theorem 7.3 for the production ρ setting for the 2020 Census redistricting data is shown in Figure 3. More fine-grained results are shown in Section 8.

We can sharpen this result even more by considering how the posterior-to-posterior ratio changes for the true record. That is, we consider the ratio for a record r when ω is drawn from the distribution $P_{\mathfrak{A}}(M(\mathcal{X}) = \omega \mid R = r)$. In this case, it is easy to see that $P_{\mathfrak{A}}(M(\mathcal{X}) = \omega \mid R = r) = P(M(\mathcal{D}^- \cup \{r\}) = \omega)$ where the randomness comes only from M. So, the curve of posterior-to-posterior semantics coincides with the (ϵ, δ)-pbdp/f-DP curve,[14] as shown by the following theorem.

Theorem 7.4. *Let M be a mechanism that satisfies f-DP. Let $P_{\mathfrak{A}}$ be an attacker's prior such that $P_{\mathfrak{A}}(\mathcal{X}^- = \mathcal{D}^-) = 1$ for some $\mathcal{D}^-$. Let $P_{\omega \mid r}$ be the distribution of the output ω of M under the assumption that the attacker's prior $P_{\mathfrak{A}}$ is correct and the true record is value r. Define $f^{-1}(z) = \inf\{y \in [0,1] \ : \ f(y) \leq z\}$. Then for any r and $\delta \in [0,1]$, if $\epsilon \geq \log \frac{\delta}{f^{-1}(1-\delta)}$, then*

$$P_{\omega|r}\left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-)P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} > e^\epsilon\right) \leq \delta.$$

Furthermore, this same bound holds if M is replaced by $A \circ M$.

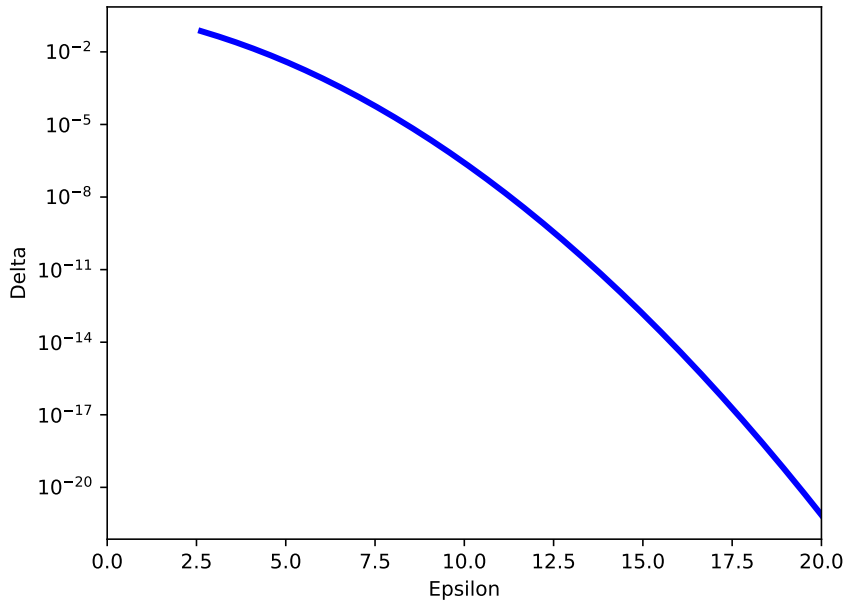

Figure 3: Bayesian (ϵ, δ) curve under the semantics of Theorem 7.3 for $\rho = 2.63$.

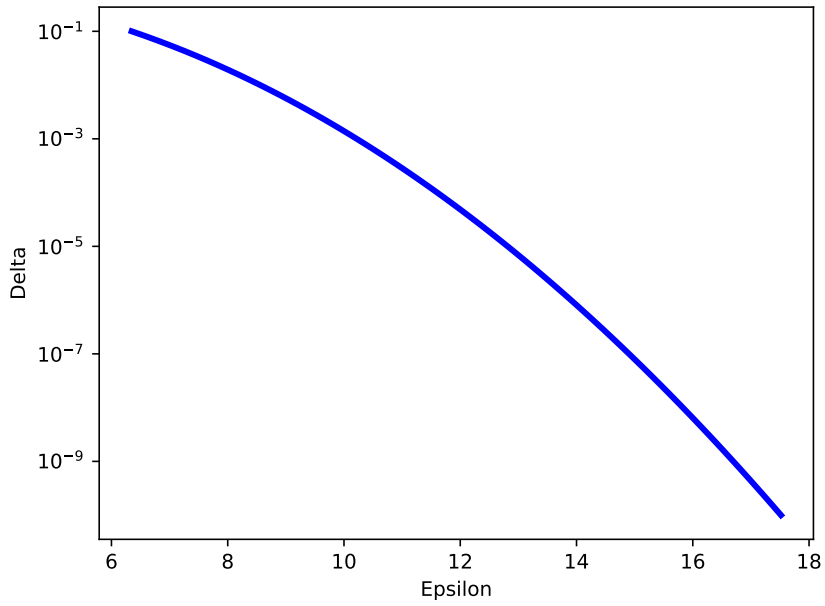

Figure 4: Bayesian (ϵ, δ) curve for the continuous Gaussian Mechanism satisfying ρ-zCDP under the semantics of Theorem 7.4, for $\rho = 2.63$.

The pbdp curve for the discrete Gaussian mechanism is difficult to compute, however, under the as-

[14]Recall, the conversion from f to the (ϵ, δ) curve is provided after Definition 5.9.

sumption that it would be similar to the curve for the continuous Gaussian mechanism (as in Section 6), we can instead analyze the curve for the Gaussian Mechanism that satisfies ρ-zCDP.[15] The (ϵ, δ)-curve for the continuous Gaussian mechanism provided by Theorem 7.4 for the production ρ settings is shown in Figure 4. While the Bayesian guarantees for such a ρ value do not appear to be strong, we show in Section 8 that more fine-grained guarantees can be stronger.

7.4.2 Bayesian semantics for arbitrary priors

When the prior is arbitrary, one can still provide semantic guarantees, but the bounds are weaker. In Section 7.4.1, we derive bounds for the posterior-to-posterior ratio when reasoning about an arbitrary record r. In this section, we provide bounds on the posterior-to-posterior ratio for reasoning about the *correct* record value.

For any given record r, we consider the distribution $P_{\mathfrak{A}}(\omega \mid r)$ and the probability that an ω from this distribution results in a large posterior-to-posterior ratio $\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)}$ for that same r. Thus the guarantee bounds the inference about the true record between the actual and counterfactual worlds.

Theorem 7.5. *Let $P_{\mathfrak{A}}$ be an attacker's prior. Let r be the target person's true record. Let M be a mechanism. Let $P_{\omega \mid r}$ be the distribution of ω conditioned on r. That is, a dataset $\mathcal{D}$ is generated from the distribution $P_{\mathfrak{A}}(\mathcal{X} \mid r)$ and $M(\mathcal{D})$ produces ω. Then M can provide the following guarantees on the posterior-to-posterior ratio $\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)}$ for the true record r.*

- *If M satisfies (α, γ)-RDP for $\alpha > 1$ then $P_{\omega \mid r}\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)} \geq e^{\epsilon}\right) \leq e^{(\alpha-1)(\gamma-\epsilon)}$.*
- *If M satisfies ρ-zCDP then for $\epsilon > \rho$, $P_{\omega \mid r}\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)} \geq e^{\epsilon}\right) \leq e^{-(\epsilon-\rho)^2/(4\rho)}$.*

Again note that these bounds are post-processing invariant in that M can be replaced with $A \circ M$ without increasing the upper bounds. The (ϵ, δ)-curve of Theorem 7.5 for the production settings for the 2020 Census redistricting data is shown in Figure 5.

By comparing Theorem 7.5 with Theorem 7.3, we see that the ρ-zCDP estimate of $\delta = e^{-(\epsilon-\rho^2)/4\rho}$ of Theorem 7.5 is larger than that of Theorem 7.3. However, the δ from Theorem 7.5 is exactly the same as the ρ-zCDP estimate of the δ parameter of *pbdp* (see Section 5.3).

8 Fine-grained, per-attribute Semantics

In this section, we explain how to use the structure of a mechanism to provide fine-grained privacy guarantees beyond those based on the overall privacy-loss parameters. This is an illustration of how a statistical agency could make additional choices about the privacy properties of the mechanisms it uses.

The structure of a mechanism allows an agency to consider scenarios with specialized privacy-violating inferences. For example, suppose Jessie has special concerns about inferences of detailed location such as a census block within a larger area such as a block group or tract but is unconcerned about the other attributes (and thus is also unconcerned about how the other attributes affect fine-grained location inference). In other words, more populous areas may not be as sensitive for Jessie as an area as small as a census block, which on average house about 50 persons. Or, Jessie may have special concerns only about inferences regarding race, ethnicity, marital status, or the composition of their household.

In these situations, the inferential focus is not on arbitrary pairs of neighbors, but rather a subset of those pairs of neighbors, and the fine-grained privacy guarantees are obtained by only examining the privacy-loss random variables associated with this subset of pairs of neighbors. In the first example, $\mathcal{D}_1$ and $\mathcal{D}_2$ would only differ on the block in Jessie's record, but would have the same tract. In another example, a pair of neighbors $\mathcal{D}_1, \mathcal{D}_2$ would only differ on the race and ethnicity in Jessie's record.

When focusing on such a more limited set of neighboring databases, additional privacy semantics can be provided. These semantics take similar forms to those described in Sections 5, 6, and 7 but the appropriate parameters and the resulting trade-off curves are different.

[15] The Gaussian mechanism that satisfies ρ-zCDP also satisfies f-DP with $f(x) = \Phi(\Phi^{-1}(1-x) - \sqrt{2\rho})$ [14, 24].

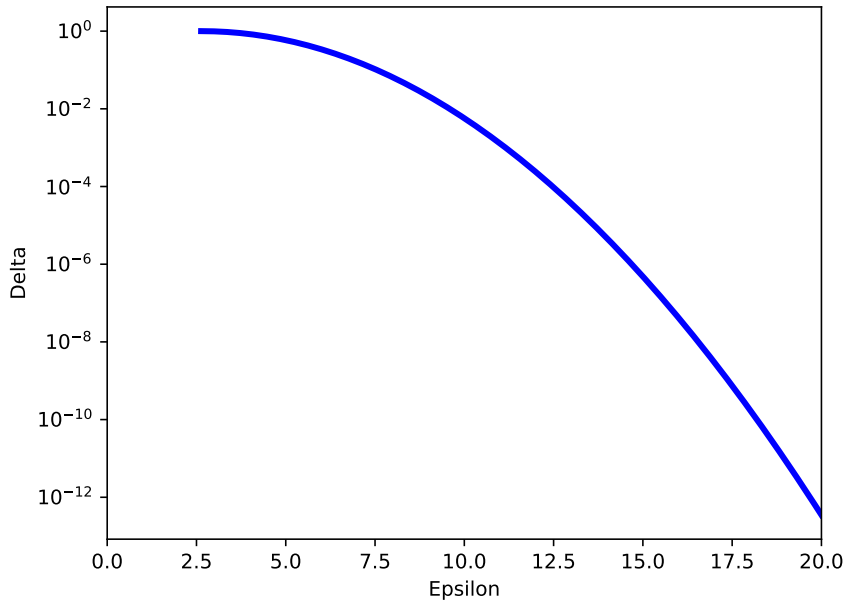

Figure 5: Bayesian (ϵ, δ) curve under the semantics of Theorem 7.5 for $\rho = 2.63$.

Rather than discuss these issues abstractly, in this section we use the disclosure avoidance system as implemented for the 2020 Census redistricting data as a case-study to illustrate how such fine-grained semantics are derived. To facilitate this discussion, we first provide a brief overview of the noise module of the TopDown Algorithm, which implements differential privacy, and then we examine the fine-grained privacy semantics it provides. More comprehensive details about the TopDown Algorithm and the data it uses can be found in [2].

8.1 An Overview of the Noise Module

The TopDown Algorithm, used to create privacy-protected microdata for the 2020 Census, consists of two stages [2]. The first stage computes query answers and adds noise to them. These noisy query answers are called *noisy measurements* and are designed to ensure differential privacy.[16] The second stage post-processes these noisy measurements to output privacy-protected microdata. The post-processing phase and its consequences are studied in other papers [2, 3]. Post-processing is largely irrelevant to this paper because all semantics we discuss are post-processing invariant. Thus, we only need to focus on the privacy semantics provided by the noisy measurements.

A query q is specified by a set of attributes, a unit of geography, and zCDP parameter ρ^*. One example is VOTINGAGE $\times$ CENRACE for Fairfax County with $\rho^* = 0.0175$. This query computes the number of people who are 18 and over for each of the 63 possible race combinations specified in Statistical Policy Directive 15 (add citation to OMB 1997) and the Census Act (add citation to 13 USC Section 191) in Fairfax County, and also the number who are 17 and under for each race combination in Fairfax County. Thus, the query answer is a vector of 126 numbers. Independent Discrete Gaussian noise with scale $\sigma = \sqrt{1/\rho^*}$ is

[16] There is a complication due to *invariants* – properties of the input data that are not known a prior but which must be enforced in the privacy-protected data. One example is that the state-level population counts, which are released as tabulated from the confidential microdata. A study of invariants and how they can alter the privacy guarantees is the subject of a future paper.

added to each of these numbers. This noise has probability mass function:

$$f(k) = \frac{e^{-k^2/(2\sigma^2)}}{\sum_{j=-\infty}^{\infty} e^{-j^2/(2\sigma^2)}}.$$

The types of queries are:

- TOTAL: the total number of people in the geographic unit.
- CENRACE: the number of people in each of the 63 OMB race combinations in the geographic unit.
- HISPANIC: the number of people who are Hispanic in the geographic unit and the number who are not Hispanic.
- VOTINGAGE: the number of people who are 18 and over in the geographic unit, and the number of people who are 17 and under.
- HHINSTLEVELS: the number of people who live in households, institutionalized group quarters, and non-institutionalized group quarters in the geographic unit.
- HHGQ: the number of people who live in a household and the number of people who live in each of the 7 major group quarters type, in the geographic region.
- HISPANIC × CENRACE: number of people who are Hispanic for each of the 63 OMB race combinations and similarly for non-Hispanic.
- VOTINGAGE × CENRACE: the number of people who are 18 and over for each of the 63 possible race combinations and similarly for people who are 17 and under.
- VOTINGAGE × HISPANIC: number of people for each combination of the VOTINGAGE and HISPANIC variables.
- VOTINGAGE × HISPANIC × CENRACE: number of people for each combination of the VOTINGAGE, HISPANIC, and CENRACE variables.
- HHGQ × VOTINGAGE × HISPANIC × CENRACE: number of people for each combination of person-level demographics for each HHGQ value.
- OCCUPANCY STATUS: Number of occupied housing units in the geographic region and the number of unoccupied housing units.

The ρ^* value associated with a query in a geographic unit can be computed from Tables 3, 4, 5, reproduced from [2]. To obtain the ρ^* value for Occupancy Status at a particular geographic location (e.g., Fairfax County), one multiplies the base ρ value for housing characteristics from Table 3 (i.e., 0.07) by the allocation for the geographic level of that location in Table 4 (e.g., 7/82 for counties, resulting in a $\rho^* = 0.07 * 7/82 \approx 0.0060$). For queries on person characteristics (e.g., VOTINGAGE × CENRACE in New York), one multiplies the base ρ value for person characteristics from Table 3 (i.e., 2.56) by the allocation for the geographic level of that location in Table 4 (e.g., 1440/4099 for states) by the allocation for that query in Table 5 (e.g., 12/4097, resulting in a $\rho^* = 2.56 * 1440/4099 * 12/4097 \approx 0.0026$).

Table 3: Base ρ value for Person Characteristics and Housing Unit Characteristic, Production Settings [2]

Person Characteristics	Housing Unit Characteristics
2.56	0.07

We note that in this geographic hierarchy, each state containing American Indian or Alaska Native areas is split into the AIAN and non-AIAN portions. Each of these latter portions is at the same level of the geographic hierarchy as states without AIAN areas. Thus, the "state" level of the hierarchy has more than

Table 4: Per Geographic Level ρ Allocation Proportions for Persons and Housing Units, Production Settings [2]

Geographic Level	Person ρ Proportions	Housing Units ρ Proportions
US	104/4099	1/205
State	1440/4099	1/205
County	447/4099	7/82
Tract	687/4099	364/1025
Custom Block Group*	1256/4099	1759/4100
Block	165/4099	99/820
*Custom block groups differ from tabulation block groups and are only used by the TDA.		

Table 5: Per Query ρ Allocation Proportions by Geographic Level for Persons, Production Settings [2]

Query	Per Query ρ Allocation Proportions by Geographic Level					
	US	State	County	Tract	CBG*	Block
TOTAL (1 cell)	0	3773/4097	3126/4097	1567/4102	1705/4099	5/4097
CENRACE (63 cells)	52/4097	6/4097	10/4097	4/2051	3/4099	9/4097
HISPANIC (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE (2 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHINSTLEVELS (3 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HHGQ (8 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
HISPANIC×CENRACE (126 cells)	130/4097	12/4097	28/4097	1933/4102	1055/4099	21/4097
VOTINGAGE×CENRACE (126 cells)	130/4097	12/4097	28/4097	10/2051	9/4099	21/4097
VOTINGAGE×HISPANIC (4 cells)	26/4097	6/4097	10/4097	5/4102	3/4099	5/4097
VOTINGAGE×HISPANIC ×CENRACE (252 cells)	26/241	2/241	101/4097	67/4102	24/4099	71/4097
HHGQ×VOTINGAGE ×HISPANIC×CENRACE (2,016 cells)	189/241	230/4097	754/4097	241/2051	1288/4099	3945/4097
*Custom block groups (CBG) differ from tabulation block groups and are only used by the TDA.						

51 entries [2]). Furthermore, the block groups for which noisy measurements are computed are custom block groups specific to the disclosure avoidance system and are different from the tabulation block groups.

The overall privacy accounting for zCDP is as follows [2, 14]. For a single block, we add up the ρ^* values for all queries in that block, and all queries in the block group, tract, county, and state containing that block, as well as the ρ^* values at the national level. The maximum of these values among all blocks is the overall ρ parameter for zCDP (these computed values happen to be the same for each block). In this case, $\rho = 2.56 + 0.07 = 2.63$.

8.2 The Fine-Grained Semantics Case Study

The guarantees for the overall privacy parameters of the algorithm were discussed in Sections 5, 6, and 7. Given the allocation of privacy-loss budget among different queries, one can analyze the additional fine-grained privacy protections that are available in a variety of situations. In this section, we consider the following scenarios.

8.2.1 Scenario A:

Jessie is concerned about protecting the exact location within a block group (so Jessie's custom block group would not be very sensitive, but the specific block within it would be). In this setting, the attacker possesses

statistical knowledge and combines it with Jessie's block group. The privacy guarantees are related to the causal effect of using Jessie's confidential block as part of the input to the TopDown Algorithm. In the frequentist case, the guarantee determines how well the attacker can distinguish between the setting when the confidential block was part of the input compared to using some other block within the same custom block group. In the Bayesian case, the guarantee determines how much better the attacker's inference about the confidential block is in the actual world compared to the inference the attacker would have in the counterfactual world.

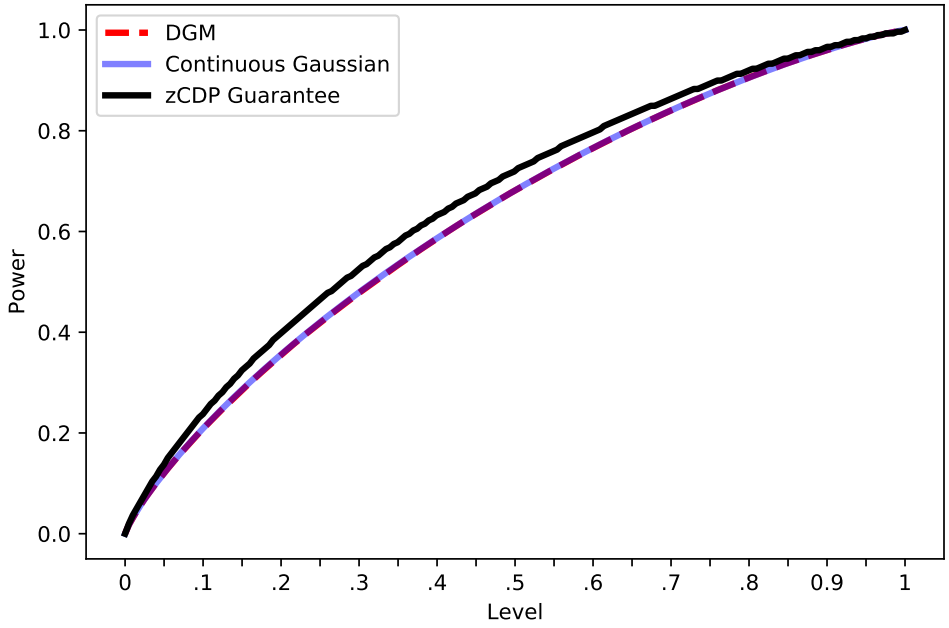

Figure 6: **Block within Custom Block Group:** Level (x-axis) vs. power (y-axis) curves for (1) the Gaussian mechanism over block-level queries at production settings for redistricting data ($\rho = 0.1115$), (2) the likelihood ratio test of the discrete Gaussian block-level noisy queries at production settings for redistricting data.

.

Significance Level	Power (Gaussian)	Power (DGM)	zCDP Upper Bound
0.01	0.03	0.03	0.04
0.05	0.12	0.12	0.14
0.10	0.21	0.21	0.24

Table 6: **Block within Custom Block Group:** Likelihood ratio test significance level/power trade-off for block-level queries (1) if Gaussian noise is used, (2) if discrete Gaussian noise is used, (3) guaranteed upper bound if an arbitrary ρ-zCDP mechanism with $\rho = 0.1115$ is used.

To explore this scenario, one considers only queries that would be affected by a change in block within the same custom block group. For instance, tract-level, county-level, and state-level queries would not be affected (their answers would not change), and thus we can apply the accounting rules using only to the affected queries (queries about person and housing unit characteristics at the block level). This results in $\rho \approx 0.1115$ for this scenario.

Figure 6 shows the significance level vs. power trade-off for this scenario. The black curve is the upper bound guarantee for any zCDP mechanism for using $\rho = 0.1115$. The blue curve is what the result would have been had the continuous Gaussian mechanism been used, and the dashed red line is a Monte Carlo simulation of the curve for the discrete Gaussian mechanism. Again, we see near-perfect alignment

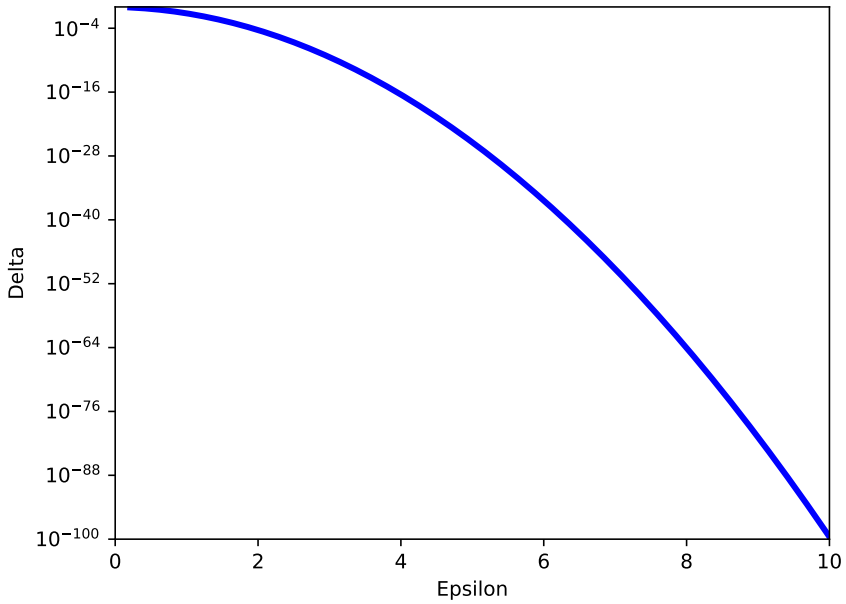

Figure 7: Bayesian (ϵ, δ) curve under the semantics of Theorem 7.3 (attacker knows all but one person) for block within custom block group inference under the privacy-loss budget allocated to block-level queries under the redistricting data production settings ($\rho = 0.1115$).

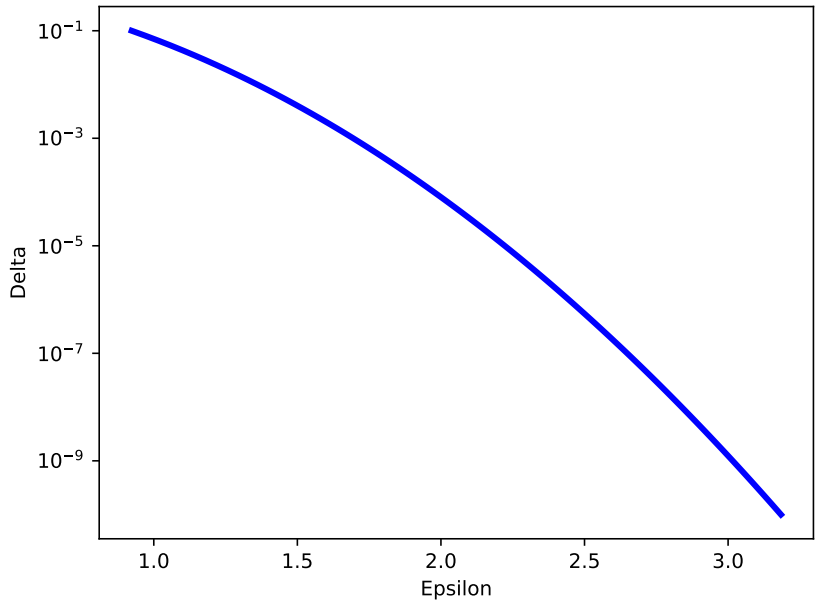

Figure 8: Bayesian (ϵ, δ) curve under the semantics of Theorem 7.4 for block within custom block group inference under the privacy budget allocated to block-level queries under the redistricting data production settings ($\rho = 0.1115$).

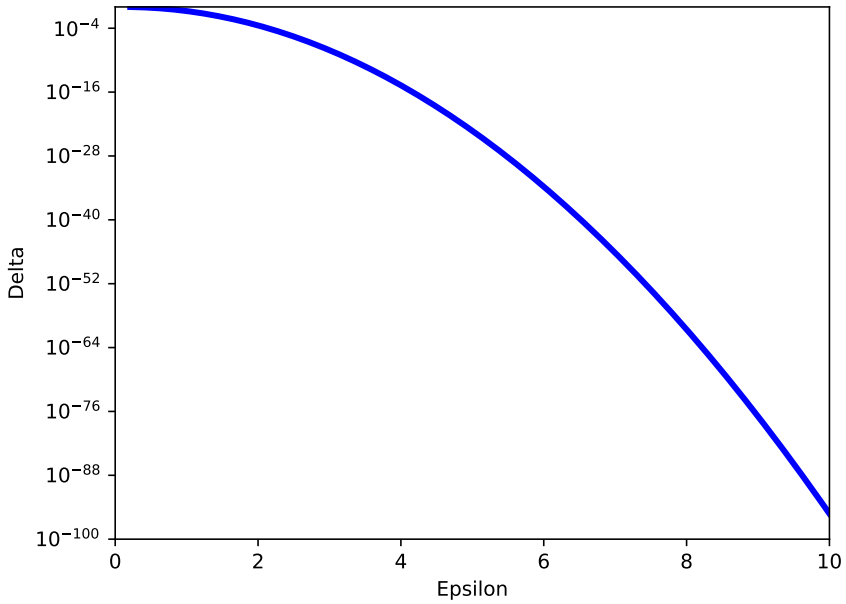

Figure 9: Bayesian (ϵ, δ) curve under the semantics of Theorem 7.5 (arbitrary priors) for block within custom block group inference under the privacy-loss budget allocated to block-level queries under the redistricting data production settings ($\rho = 0.1115$).

between the continuous and discrete Gaussian mechanisms. This suggests that calculations for the continuous Gaussian are a quick and accurate approximate to the discrete Gaussian mechanism used by the TopDown Algorithm.

All three curves in Figure 6 are near the ideal curve of "level=power" that would be achieved by a completely non-informative test indicating strong levels of privacy protection. The power at a selected set of significance levels is given in Table 6.

The corresponding Bayesian curves are shown in Figure 7 for the semantics of Theorem 7.3 (attacker knows all but one person), Figure 8 for the semantics of Theorem 7.4 for the continuous Gaussian mechanism – these semantics correspond to the (ϵ, δ)-pbdp curve which we believe approximate the curve for the discrete Gaussian Mechanism, and Figure 9 for the semantics of Theorem 7.5 (arbitrary adversarial priors). Both Theorems 7.3 and 7.5 were derived using Markov's inequality and hence are likely to overestimate δ for small values of ϵ.

Given the empirical similarity of the significance level vs. power curves between the continuous and discrete Gaussian mechanisms, and given the hypothesized similarity for the (ϵ, δ)-pbdp curves for the two mechanisms, in the remaining scenarios we will only present the frequentist semantics and Bayesian semantics of Theorem 7.4 for the continuous Gaussian (because of the computational cost associated with estimating the curves for the discrete Gaussian).

8.2.2 Scenario B:

Jessie is concerned with protecting the specific block within the tract that Jessie resides in. Applying the privacy accounting rules to queries at the block and custom block group level, we obtain $\rho \approx 0.926$.

The frequentist semantics are shown in Figure 10. For significance levels 0.01, 0.05 and 0.10, the corresponding power values are 0.17, 0.39, and 0.53, respectively. The Bayesian (ϵ, δ) curve of Theorem 7.4, which is also the (ϵ, δ)-pbdp curve is shown in Figure 11.

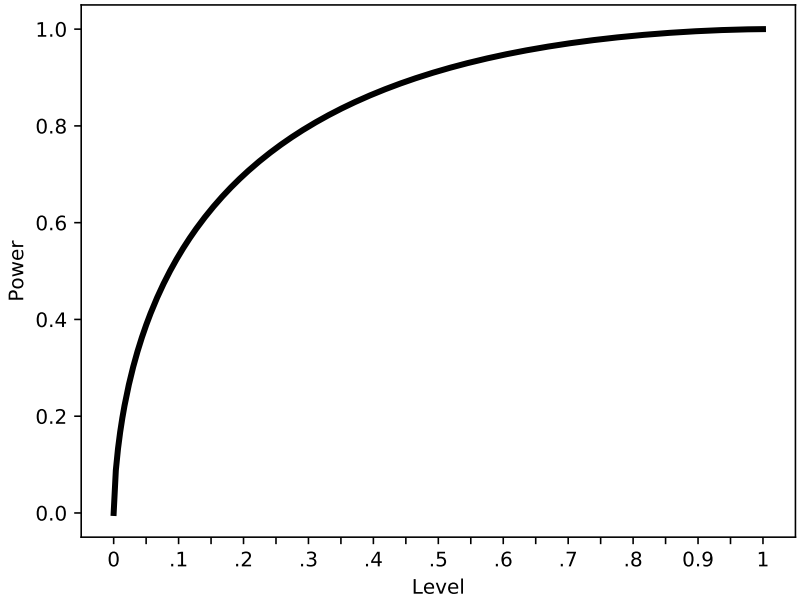

Figure 10: Significance level vs. power tradeoff for the Gaussian mechanism at $\rho \approx 0.926$.

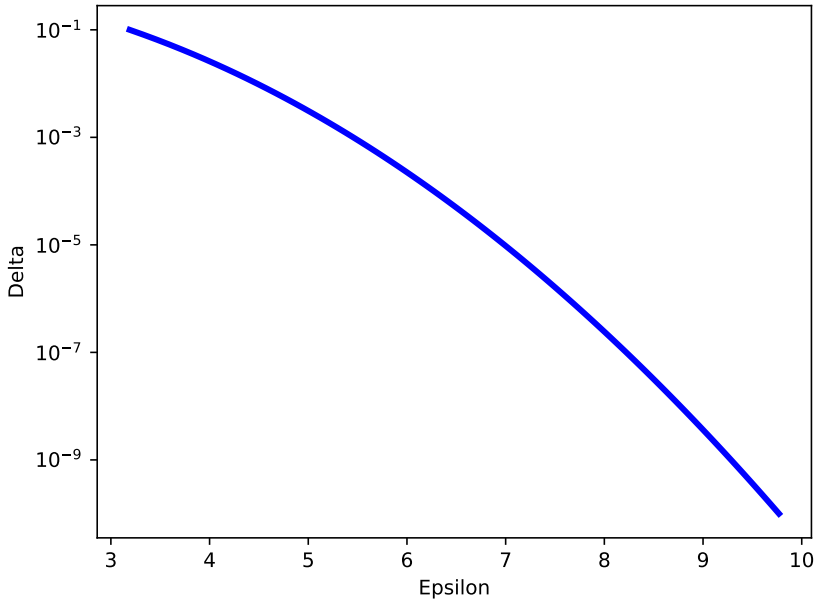

Figure 11: Bayesian (ϵ, δ) curve of Theorem 7.4 for inference about block within tract.

8.2.3 Scenario C:

Jessie is only concerned with protecting inference about race. The total ρ for queries that involve race is ≈ 0.952 and so the curves would be similar to the previous scenario and hence are not reproduced. The

powers for selected significance levels are also almost the same. For significance levels 0.01, 0.05 and 0.10, the corresponding power values are 0.17, 0.40, and 0.54, respectively.

8.2.4 Scenario D:

Jessie is only concerned about protecting inference about ethnicity. The total ρ for queries that involve ethnicity is ≈ 0.945 and so the curves would again be similar to the previous scenarios. For significance levels 0.01, 0.05 and 0.10, the corresponding power values are 0.17, 0.39, and 0.54, respectively.

8.2.5 Scenario E:

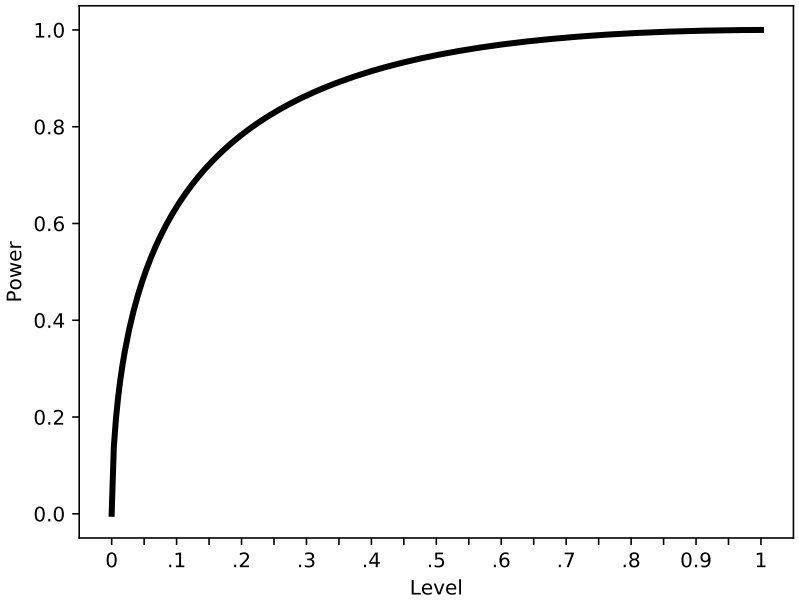

Figure 12: Significance level vs. power trade-off for the Gaussian mechanism for Scenario E.

Jessie is concerned about protecting inferences about race, but also believes that knowledge about the specific block within Jessie's tract would be too revealing about race. In this case, both race and block within tract are the attributes of concern, so all queries at the custom block group and block level are affected as well as all queries involving race at the tract, county, state, and national levels (individuals who do not think race queries at or above the tract level affect inference about them would disregard those queries in their ρ calculations). The total ρ in this setting is ≈ 1.32. At this higher level of ρ, the semantics get weaker. For example, for significance levels 0.01, 0.05 and 0.10, the corresponding power values are 0.24, 0.49, and 0.63, respectively. The full significance level vs. power tradeoff curve is shown in Figure 12 while the Bayesian (ϵ, δ)-curve is shown in Figure 13.

8.2.6 Scenario F:

Jessie is concerned about protecting inferences about voting age, but also believes that knowledge about specific block within Jessie's block group would be too revealing about voting age. In this case, both voting age and block within block group are attributes of concern. The total ρ of queries involving these attributes is ≈ 0.555. This is a fairly low setting of ρ. The significance level vs. power trade-offs for significance levels 0.01, 0.05 and 0.10 have the corresponding power values are 0.10, 0.28, and 0.41, respectively. The full significance level vs. power trade-off curve is shown in Figure 14, while the Bayesian (ϵ, δ)-curve is shown in Figure 15.

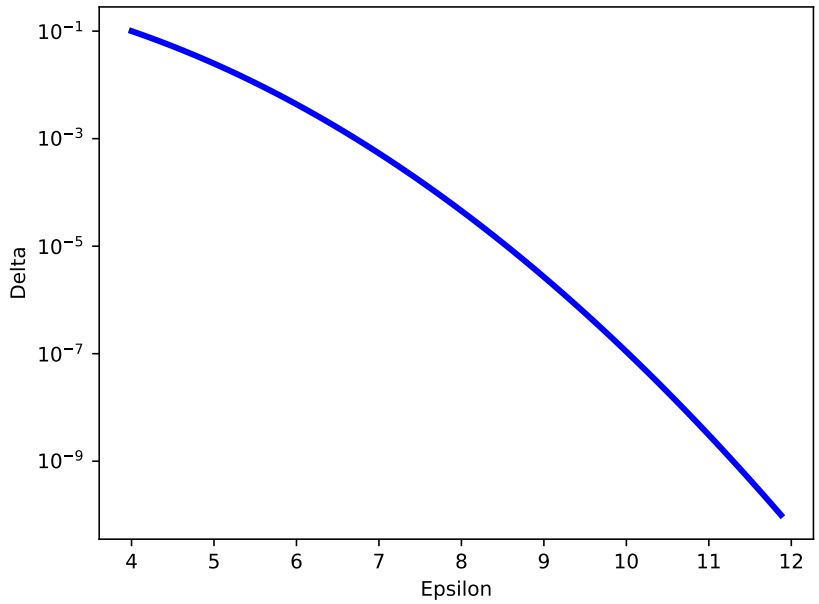

Figure 13: Bayesian (ϵ, δ) curve of Theorem 7.4 for Scenario E.

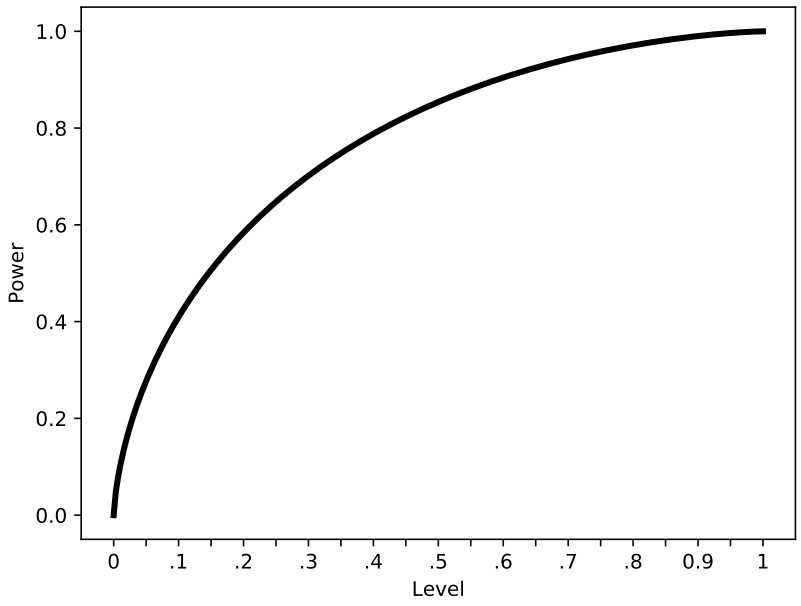

Figure 14: Significance level vs. power trade-off for the Gaussian mechanism for Scenario F.

8.2.7 Scenario G:

Jessie is concerned about protecting inferences about voting age and race, but also believes that knowledge about specific block within Jessie's block group would be too revealing about voting age and race. In this

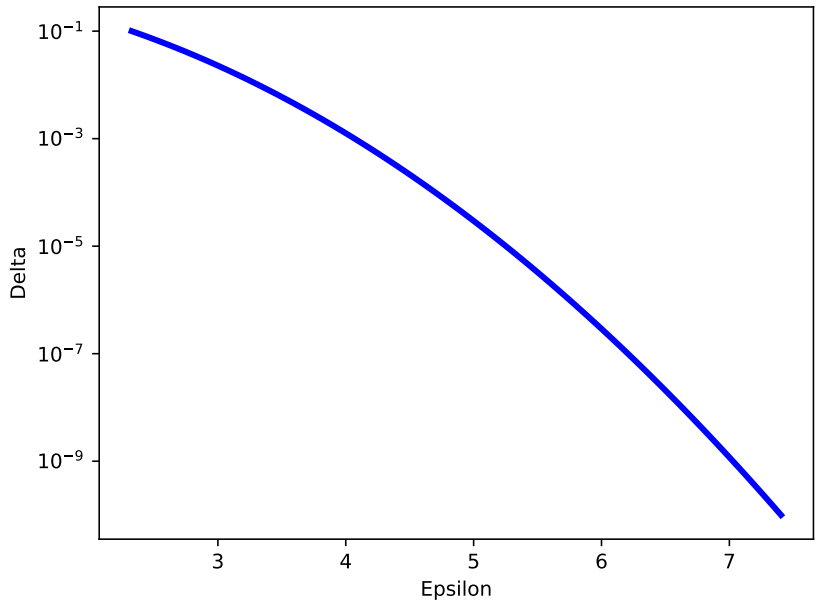

Figure 15: Bayesian (ϵ, δ) curve of Theorem 7.4 for Scenario F.

case, voting age, race, and block within block group are attributes of concern. The ρ for this scenario is ≈ 0.969 and so the semantics are very similar to Scenario C.

8.2.8 Scenario H:

Jessie is concerned about protecting inferences about voting age and ethnicity, but also believes that knowledge about specific block within Jessie's block group would be too revealing about voting age and ethnicity. In this case, voting age, ethnicity, and block within block group are attributes of concern. The ρ for this scenario is ≈ 0.968 and so the semantics are very similar to Scenario C.

8.3 Discussion

In this section, we considered details of the allocation of the overall production privacy-loss budget to different queries by the TopDown Algorithm and appropriate accounting rules to use to compute the ρ of a selected set of queries. These ρ values can be used to provide semantics to different scenarios, such as those studied in this section.

These scenarios are by no means complete, and different people may have different concerns. They can evaluate the privacy impact of their concerns by selecting the set of queries that may be potentially revealing in their opinion (for example, they may feel that race queries below the county level may be revealing). Once the ρ for the subset of queries is computed, the power at a given significance level ℓ can be computed as:

$$1 - \Phi_{\frac{1}{2\rho}}(\Phi^{-1}_{\frac{1}{2\rho}}(1-\ell) - 1),$$

where $\Phi_{\frac{1}{2\rho}}$ is the CDF of the zero-mean Gaussian distribution with variance $\sigma^2 = \frac{1}{2\rho}$. For the continuous Gaussian mechanism, this happens to mathematically correspond to the difficulty of distinguishing between a $N(0, \sigma^2)$ and $N(1, \sigma^2)$ random variable using a hypothesis test based on one sample.

Meanwhile, the (ϵ, δ)-curve of Theorem 7.4 for the continuous Gaussian can be computed from the formula:

$$\epsilon = \log\left(\frac{\delta}{\Phi_1(-\Phi_1^{-1}(1-\delta) - \sqrt{2\rho})}\right).$$

9 Related Work

Differential privacy has many variations, only some of which were discussed in this paper. See Desfontaines and Pejó [22] for a more comprehensive survey of variants of differential privacy.

There are several sources that explicitly treat the privacy guarantees as a causal property in which an intervention in the data (such as replacing a record) can occur before providing an input to M [43, 61, 12]. These papers all contain Bayesian elements. The Bayesian variations of differential privacy (e.g., [43, 61, 12, 60, 47, 19, 21]) also incorporate noise from the data into the constraints on a mechanisms output or privacy-loss random variable.

Among these, Kifer and Machanavajjhala [47] study the semantics of a proposed definition called *Pufferfish*, Bassily et al. [12] study the semantics of a proposed definition called *Coupled-worlds Privacy*, and Kasiviswanathan and Smith [43] study the semantics of approximate differential privacy under Bayesian inference. The Pufferfish definition comes with an absolute upper bound on the posterior-to-posterior ratio of a suitably chosen counterfactual world [47] under a restricted set of attacker priors. Bassily et al. [12] propose several variations of their privacy definition, one of which compares posteriors using inequalities similar to approximate differential privacy, also under a restricted set of priors. Similar studies were performed by Desfontaines et al. [21], studying Bayesian semantics of privacy definitions similar to pure differential privacy, and proposing definitions that compare posteriors similarly to approximate differential privacy. Kasiviswanathan and Smith study the Bayesian guarantees of approximate differential privacy under arbitrary priors, hence, similar in spirit to the Bayesian guarantees in Section 7. They compare posteriors using total variation distance and obtain a non-trivial bound (i.e., total variation distance less than 1) under a restricted set of parameters, such as $\epsilon < 1/4$ and δ much smaller than the inverse of the number of people in the data.

For Bayesian privacy definitions that require a limited set of priors for modeling the data, it is also important to study how privacy degrades under miss-specification of the priors, also known as *close adversaries*. Song et al. [59] provided such an analysis for Pufferfish.

An alternative approach, known as *restricted sensitivity* [13] flips the idea typically taken by Bayesian approaches. Instead of providing privacy protections only when certain assumptions are met, one can instead provide privacy protections with no assumptions about the data, but only guarantee statistical validity of the results under certain conditions on the data.

The idea of considering more fine-grained protection (i.e., protection of different parts of a record) has a very long history and is often achieved by redefining the choice of neighbors (e.g., [29, 27, 7, 12, 47, 41, 17, 49, 54, 40]).

Finally, the use of the Gaussian distribution is often justified with central limit theorems for privacy loss random variables [58, 24].

10 Conclusions

In this paper, we surveyed the major differential privacy frameworks that are important for understanding the privacy semantics of the redistricting data from the 2020 Census – namely, pure, approximate, Rényi, zero-concentrated, and f differential privacy. The semantics considered interpretability of the privacy parameters, the frequentist hypothesis testing guarantees, and the Bayesian posterior-to-posterior guarantees.

We interpreted the guarantees in the context of the production settings for the disclosure avoidance system implemented for the 2020 Census major data products including the redistricting data. The main purposes of this paper were (1) to explain the privacy protections used in the redistricting data, (2) to serve as a guide for how other organizations can study the privacy semantics of their data releases, and (3) to provide a guide for what peer-reviewers can look for when asked to evaluate new privacy proposals.

The paper also generated some novel results such as a new motivation for f-DP via pbdp and the Bayesian semantics for Rényi and zero-concentrated DP.

There are some topics relevant to the privacy guarantees of the 2020 Census redistricting data that are not covered. These include accounting for the policy-determined data *invariants* – properties of the confidential input data that were replicated without DP protection in the privacy-protected data. These include the exact as-enumerated state population totals, the number of housing units in each level of geography, and the number of occupied group quarters facilities in each major type for each level of geography. These more complex topics will be the subject of a separate paper.

Acknowledgments and Disclaimer

The views expressed in this technical paper are those of the authors and not those of the U.S. Census Bureau. We are grateful for the helpful comments provided by Mark Bun and Marco Gaboardi. This paper has also been released to arxiv.org as https://arxiv.org/abs/2209.03310. If there are revisions, the latest version may also be found there.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[2] John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, forthcoming. Preprint https://www.census.gov/library/working-papers/2022/adrm/CED-WP-2022-002.html.

[3] John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson L. Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, and Pavel Zhuravlev. An uncertainty principle is a price of privacy-preserving microdata. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11883–11895, 2021.

[4] John M. Abowd and Ian M. Schmutte. Economic analysis and statistical disclosure limitation. *Brookings Papers on Economic Activity*, pages 221–267, Spring 2015.

[5] J. Trent Alexander, Michael Davern, and Betsey Stevenson. Review – Inaccurate Age and Sex Data in the Census PUMS files: Evidence and Implications. *The Public Opinion Quarterly*, 74(3):551–569, 2010.

[6] American Cancer Society. History of the cancer prevention studies. https://www.cancer.org/research/population-science/cancer-prevention-and-survivorship-research-team/acs-cancer-prevention-studies/history-cancer-prevention-study.html, Retrieved June 23, 2021.

[7] Hilal Asi, John C. Duchi, and Omid Javidbakht. Element level differential privacy: The right granularity of privacy. *CoRR*, abs/1912.04042, 2019.

[8] Shahab Asoodeh, Jiachun Liao, Flavio P. Calmon, Oliver Kosut, and Lalitha Sankar. A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 920–925, 2020.

[9] B. Balle, G. Barthe, and M. Gaboardi. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1), 2020.

[10] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and Rényi differential privacy. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[11] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 403–412. PMLR, 2018.

[12] Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *FOCS*, pages 439–448, 2013.

[13] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, page 87–96, New York, NY, USA, 2013. Association for Computing Machinery.

[14] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings, Part I, of the 14th International Conference on Theory of Cryptography - Volume 9985*, 2016.

[15] Clément Canonne. What is δ, and what δifference does it make? (footnote). https://differentialprivacy.org/flavoursofdelta/, March 2021.

[16] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15676–15688. Curran Associates, Inc., 2020.

[17] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In Emiliano De Cristofaro and Matthew Wright, editors, *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[18] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429–444, 1977.

[19] Ashish Dandekar, Debabrota Basu, and Stéphane Bressan. Differential privacy at risk: Bridging randomness and privacy budget. *Proceedings on Privacy Enhancing Technologies*, 2021(1):64–84, 2021.

[20] Damien Desfontaines. https://desfontain.es/privacy/privacy-loss-random-variable.html. https://desfontain.es/privacy/privacy-loss-random-variable.html, July 2022.

[21] Damien Desfontaines, Elisabeth Krahmer, and Esfandiar Mohammadi. Passive and active attackers in noiseless privacy. *CoRR*, abs/1905.00650, 2019.

[22] Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020(2):288–313, 2020.

[23] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2003.

[24] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *JRSS-B*, 84:3–37, 2022.

[25] George T. Duncan, Mark Elliot, and Juan-José Salazar-Gonzàlez. *Statistical Confidentiality: Principles and Practice*. Springer, 2011.

[26] C. Dwork and R. Pottenger. Toward practicing privacy. *J Am Med Inform Assoc*, 20(1):102–8, Jan 2013.

[27] Cynthia Dwork. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1–12, 2006.

[28] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Serge Vaudenay, editor, *Advances in Cryptology (EUROCRYPT)*, 2006.

[29] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[30] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):93–107, 2010.

[31] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In Matthew K. Franklin, editor, *Advances in Cryptology - CRYPTO 2004, 24th Annual International CryptologyConference, Santa Barbara, California, USA, August 15-19, 2004, Proceedings*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 2004.

[32] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

[33] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016.

[34] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60. IEEE Computer Society, 2010.

[35] Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. In *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, 2015.

[36] Vitaly Feldman and Tijana Zrnic. Individual privacy accounting via a Rényi filter. In *Advances in Neural Information Processing Systems*, 2021.

[37] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition Attacks and Auxiliary Information in Data Privacy. In *KDD*, 2008.

[38] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing (STOC)*, page 351–360, 2009.

[39] E.C. Hammond and D. Horn. The relationship between human smoking habits and death rates: A follow-up study of 187,866 men. *Journal of the American Medical Association*, 155(15):1316–1328, 1954.

[40] Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *2009 Ninth IEEE International Conference on Data Mining*, pages 169–178, 2009.

[41] Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: tuning privacy-utility trade-offs using policies. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1447–1458, 2014.

[42] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[43] Shiva P Kasiviswanathan and Adam Smith. On the'Semantics' of Differential Privacy: A Bayesian Formulation. *Journal of Privacy and Confidentiality*, 6(1):1–16, 2014. See also: https://arxiv.org/abs/0803.3946.

[44] Daniel Kifer and Bing-Rong Lin. Towards an axiomatization of statistical privacy and utility. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, page 147–158, New York, NY, USA, 2010. Association for Computing Machinery.

[45] Daniel Kifer and Bing-Rong Lin. An axiomatic view of statistical privacy and utility. *Journal of Privacy and Confidentiality*, 4(1), Jul. 2012.

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

[47] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1):3, 2014.

[48] Ios Kotsogiannis, Stelios Doudalis, Sam Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. One-sided differential privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.

[49] Frank McSherry and Ratul Mahajan. Differentially-private network trace analysis. *SIGCOMM Comput. Commun. Rev.*, 40(4):123–134, aug 2010.

[50] Sebastian Meiser. Approximate and probabilistic differential privacy definitions. Cryptology ePrint Archive, Report 2018/277, 2018. https://ia.cr/2018/277.

[51] Ilya Mironov. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, CSF 2017, Santa Barbara, CA, USA, August 21-25, 2017*, pages 263–275, 2017.

[52] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. *THEORY OF COMPUTING*, 14(8), 2018.

[53] Jerzy Neyman and Egon Sharpe Pearson. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289—337, 1933.

[54] Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM, 2007.

[55] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[56] Ryan Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. Privacy odometers and filters: Pay-as-you-go composition. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[57] Pierangela Samarati and Latanya Sweeney. k-anonymity: A model for protecting privacy. In *IEEE Symposium on Research in Security and Privacy (S&P)*, 1998.

[58] David M. Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Proc. Priv. Enhancing Technol.*, 2019(2):245–269, 2019.

[59] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1291–1306. Association for Computing Machinery, 2017.

[60] Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9583–9592. PMLR, 13–18 Jul 2020.

[61] M. Tschantz, S. Sen, and A. Datta. SoK: Differential privacy as a causal property. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 354–371, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

[62] S. L. Warner. Randomised response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63—69, March 1965.

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

[64] Yingtai Xiao, Zeyu Ding, Yuxin Wang, Danfeng Zhang, and Daniel Kifer. Optimizing fitness-for-use of differentially private linear queries. *Proc. VLDB Endow.*, 14(10):1730–1742, 2021.

A Proofs from Section 5

Theorem 5.4. *Definitions 5.2 and 5.3 are equivalent for $\delta \in [0, 1]$.*

Proof of Theorem 5.4. First, we note that when $\delta = 0$, both definitions are equivalent to pure ϵ-differential privacy.

Let A be a post-processing algorithm as in Definition 5.2. Define A^* to be the algorithm such that $A^*(\omega) = 1$ if $P(A(M(\mathcal{D}_1)) = \omega) > e^\epsilon P(A(M(\mathcal{D}_2)) = \omega)$ and $A^*(\omega) = 0$ otherwise (for continuous distributions interpret these as Radon-Nikodym densities). Then, by definition of the privacy loss random variables, $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > e^\epsilon) = P(\mathcal{L}^{A^*\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > e^\epsilon)$. So we can restrict Definition 5.2 to binary-valued post-processing functions.

Also note that if A is a binary-valued algorithm, then considering both A and $1-A$, we see that Definition 5.3 does not change if we require both

$$\Big(P(A(M(\mathcal{D}_1)) = 1) > e^\epsilon P(A(M(\mathcal{D}_2)) = 1)\Big) \;\Rightarrow\; \Big(P(A(M(\mathcal{D}_1)) = 1) \le \delta\Big)$$
$$\Big(P(A(M(\mathcal{D}_1)) = 0) > e^\epsilon P(A(M(\mathcal{D}_2)) = 0)\Big) \;\Rightarrow\; \Big(P(A(M(\mathcal{D}_1)) = 0) \le \delta\Big).$$

Let A be a binary-valued post-processing algorithm, and let $\mathcal{D}_1$ and $\mathcal{D}_2$ be neighbors. Consider the following cases:

- If $P(A(M(\mathcal{D}_1)) = 1) \ge e^\epsilon P(A(M(\mathcal{D}_2)) = 1)$ and $P(A(M(\mathcal{D}_1)) = 0) \ge e^\epsilon P(A(M(\mathcal{D}_2)) = 0)$ or if $P(A(M(\mathcal{D}_1)) = 1) \le e^\epsilon P(A(M(\mathcal{D}_2)) = 1)$ and $P(A(M(\mathcal{D}_1)) = 0) \le e^\epsilon P(A(M(\mathcal{D}_2)) = 0)$ then, since A either outputs 0 or 1, the probabilities do not depend on whether the input was $\mathcal{D}_1$ or $\mathcal{D}_2$ (e.g., $P(A(M(\mathcal{D}_2)) = 1) = P(A(M(\mathcal{D}_1)) = 1)$), ϵ must therefore be 0, and the privacy loss is 0, so both definitions are vacuously satisfied (since they are active for $\epsilon > 0$).
- If $P(A(M(\mathcal{D}_1)) = 1) > e^\epsilon P(A(M(\mathcal{D}_2)) = 1)$ but $P(A(M(\mathcal{D}_1)) = 0) \le e^\epsilon P(A(M(\mathcal{D}_2)) = 0)$ then, by definition of the privacy-loss random variable, $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) = P(A(M(\mathcal{D}_1)) = 1)$ since 1 is the only output under $\mathcal{D}_1$ with sufficient privacy loss. Thus, either both $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) \le \delta$ and $P(A(M(\mathcal{D}_1)) = 1) \le \delta$ or both $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) > \delta$ and $P(A(M(\mathcal{D}_1)) = 1) > \delta$. Therefore, with this setting of $A, \mathcal{D}_1, \mathcal{D}_2$ the constraints of Definitions 5.3 and 5.2 are either both satisfied or both not satisfied.
- Similarly, if $P(A(M(\mathcal{D}_1)) = 1) \le e^\epsilon P(A(M(\mathcal{D}_2)) = 1)$ but $P(A(M(\mathcal{D}_1)) = 0) > e^\epsilon P(A(M(\mathcal{D}_2)) = 0)$ then, by definition of the privacy-loss random variable, $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) = P(A(M(\mathcal{D}_1)) = 0)$ since 0 is the only output under $\mathcal{D}_1$ with sufficient privacy loss. Thus either both $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) \le \delta$ and $P(A(M(\mathcal{D}_1)) = 0) \le \delta$ or both $P(\mathcal{L}^{A\circ M}_{\mathcal{D}_1,\mathcal{D}_2} > \epsilon) > \delta$ and $P(A(M(\mathcal{D}_1)) = 0) > \delta$. Therefore, with this setting of $A, \mathcal{D}_1, \mathcal{D}_2$ the constraints of Definitions 5.3 and 5.2 are either both satisfied or both not satisfied.

Thus the two definitions are equivalent. □

Theorem A.1. *Given an $\epsilon > 0$, a mechanism M satisfies pure ϵ-differential privacy if and only if it satisfies Definition 5.5 for all $\delta > 0$.*

Proof of Theorem A.1. If M satisfies pure differential privacy, then by postprocessing invariance, if $e^{-\epsilon}\delta \ge P(A(M(\mathcal{D}_1)) = 1)$, we have:

$$e^{-\epsilon}\delta \ge P(A(M(\mathcal{D}_1)) = 1) \ge e^{-\epsilon} P(A(M(\mathcal{D}_2)) = 1)$$
$$\text{and so } \delta \ge P(A(M(\mathcal{D}_2)) = 1)$$

for any $\delta \ge 0$.

For the other direction suppose M satisfies Definition 5.5 for the given ϵ and all $\delta > 0$. If M does not satisfy ϵ differential privacy then there exists an A and neighboring $\mathcal{D}_1, \mathcal{D}_2$ such that $P(A(M(\mathcal{D}_2)) = 1) > e^\epsilon P(A(M(\mathcal{D}_1)) = 1)$. Setting $\delta = e^\epsilon P(A(M(\mathcal{D}_1)) = 1)$, we get

$$\begin{aligned} P(A(M(\mathcal{D}_1)) = 1) &\leq e^{-\epsilon}\delta \quad \text{by choice of } \delta \\ P(A(M(\mathcal{D}_2)) = 1) &\leq \delta \quad \text{since } M \text{ satisfies Def 5.5 for this } \epsilon, \delta \text{ combination} \\ &= e^\epsilon P(A(M(\mathcal{D}_1)) = 1) \end{aligned}$$

contradicting the assumption that M does not satisfy pure ϵ-differential privacy. □

Theorem 5.7. *Definitions 5.3 and Definition 5.5 are equivalent for $\delta \in (0, 1]$.*

Proof of Theorem 5.7. We note that, due to symmetry of neighbors, if $\mathcal{D}_1$ is a neighbor of $\mathcal{D}_2$ then $\mathcal{D}_2$ is a neighbor of $\mathcal{D}_1$, thus, the condition in Definition 5.3 can be replaced with:

$$\Big(P(A(M(\mathcal{D}_2)) = 1) > e^\epsilon P(A(M(\mathcal{D}_1)) = 1)\Big) \;\Rightarrow\; \Big(P(A(M(\mathcal{D}_2)) = 1) \leq \delta\Big).$$

If there exists a M that does not satisfy Definition 5.5, then there exist a binary-valued post-processing algorithm A, neighbors $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $P(A(M(\mathcal{D}_1)) = 1) \leq e^{-\epsilon}\delta$ but $P(A(M(\mathcal{D}_2)) = 1) > \delta$. In this case, $P(A(M(\mathcal{D}_2)) = 1) > e^\epsilon P(A(M(\mathcal{D}_1)) = 1)$ but $P(A(M(\mathcal{D}_2)) = 1) > \delta$ and so M does not satisfy Definition 5.3. Hence Definition 5.3 implies Definition 5.5.

For the other direction, suppose M does not satisfy Definition 5.3. Then there exists a binary-valued post-processing algorithm A, neighbors $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $P(A(M(\mathcal{D}_2)) = 1) > e^\epsilon P(A(M(\mathcal{D}_1)) = 1)$ and $P(A(M(\mathcal{D}_2)) = 1) > \delta$. Suppose, for the sake of contradiction, that M satisfies Definition 5.5. Since $P(A(M(\mathcal{D}_2)) = 1) > \delta$ this would mean that $P(A(M(\mathcal{D}_1)) = 1) > e^{-\epsilon}\delta$. Define $\gamma = \frac{e^{-\epsilon}\delta}{P(A(M(\mathcal{D}_1))=1)}$ and note that $\gamma < 1$. Define the post-processing function A^* such that $A^*(\omega)$ returns $A(\omega)$ with probability γ and otherwise returns 0. Then $P(A^*(M(\mathcal{D}_1)) = 1) = \gamma P(A(M(\mathcal{D}_1)) = 1) = e^{-\epsilon}\delta$. Since we assumed (by way of contradiction) that M satisfies Definition 5.5, we must have:

$$\begin{aligned} \delta &\geq P(A^*(M(\mathcal{D}_2)) = 1) \\ &= \gamma P(A(M(\mathcal{D}_2)) = 1) \\ &= \frac{e^{-\epsilon}\delta P(A(M(\mathcal{D}_2)) = 1)}{P(A(M(\mathcal{D}_1)) = 1)} \\ &\Rightarrow \frac{P(A(M(\mathcal{D}_2)) = 1)}{P(A(M(\mathcal{D}_1)) = 1)} \leq e^\epsilon, \end{aligned}$$

which contradicts the starting requirements on M and A that $P(A(M(\mathcal{D}_2)) = 1) > e^\epsilon P(A(M(\mathcal{D}_1)) = 1)$. Thus, if M does not satisfy Definition 5.3 then it cannot satisfy Definition 5.5 either. Hence, Definition 5.5 implies Definition 5.3. □

B Proofs from Section 6

Theorem 6.4. *Given a transparent, symmetric, convex and postprocessing invariant privacy definition with privacy parameters θ, let $\mathfrak{M}$ denote the set of all mechanisms that satisfy the privacy definition at those privacy parameter settings. Define the significance level vs. power trade-off function g as follows. For any significance level ℓ, $g(\ell)$ is the supremum of achievable power in testing $H_0 : \mathcal{D}_1$ vs. $H_1 : \mathcal{D}_2$ for all neighboring pairs $\mathcal{D}_1, \mathcal{D}_2$, for all mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level is at most ℓ. Then g is a concave function.*

Proof of Theorem 6.4. For a mechanism M, neighboring datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, and significance level ℓ, let $\phi_{(M,\mathcal{D}_1,\mathcal{D}_2,\ell)}$ denote a hypothesis test on the output of M for distinguishing between $H_0 : \mathcal{D}_1$ and $H_1 : \mathcal{D}_2$ with significance level ℓ. Then we can write:

$$g(\ell) = \sup_{M \in \mathfrak{M},\, \mathcal{D}_1 \sim \mathcal{D}_2,\, \phi_{(M,\mathcal{D}_1,\mathcal{D}_2,\ell)}} power(\phi_{(M,\mathcal{D}_1,\mathcal{D}_2,\ell)})$$

We can define the significance level vs. power trade-off g_M for a specific mechanism M as follows:

$$g_M(\ell) = \sup_{\mathcal{D}_1 \sim \mathcal{D}_2,\, \phi_{(M,\mathcal{D}_1,\mathcal{D}_2,\ell)}} \sup power(\phi_{(M,\mathcal{D}_1,\mathcal{D}_2,\ell)})$$

Note that $g(\ell) = \sup_{M \in \mathfrak{M}} g_M(\ell)$.

We then note that postprocessing can only reduce the maximum power. That is, $g_M(\ell) \geq g_{A \circ M}(\ell)$. This is because any hypothesis test ϕ on the output of $A \circ M$ can be converted to a hypothesis test on the output of M by first running A and then ϕ.

Next, note that by transparency, M can also output its name without affecting privacy properties.

Now, for any $p \in [0,1]$ and $M_1, M_2 \in \mathfrak{M}$, Consider the mechanism $M_1 \oplus_p M_2$ such that $(M_1 \oplus_p M_2)(\mathcal{D})$ outputs $M_1(\mathcal{D})$ with probability p, and otherwise outputs $M_2(\mathcal{D})$ (in other words, it randomly chooses which mechanism to run). By convexity, $M_1 \oplus_p M_2 \in \mathfrak{M}$. Furthermore, the output of $M_1 \oplus_p M_2$ can include the name of the sub-mechanism that was run (M_1 or M_2, by transparency, since M_1 and M_2 are in $\mathfrak{M}$) and removing the names can only decrease the maximum power (by postprocessing invariance). Thus, since the identities of the chosen mechanisms are provided, any hypothesis test ϕ^* on the output of $M_1 \oplus_p M_2$ produces a pair of hypothesis tests – one hypothesis test ϕ_1 on the output of M_1 and another test ϕ_2 on the output of M_2 (and vice versa). Thus we can use the notation (ϕ_1, ϕ_2) to represent a hypothesis test on the output of $M_1 \oplus_p M_2$.

If the significance level of ϕ_1 is ℓ_1 with power $1 - \beta_1$, and the significance level of ϕ_2 is ℓ_2 with power $1 - \beta_2$, then the significance level of (ϕ_1, ϕ_2) is $p\ell_1 + (1-p)\ell_2$ and the power is a convex combination of the powers: $p(1-\beta_1) + (1-p)(1-\beta_2)$. Thus we have:

$$\begin{aligned}
&g(p\ell_1 + (1-p)\ell_2) \\
&\geq \sup_{\mathcal{D}_1,\mathcal{D}_2} \sup_{M_1} \sup_{M_2} \sup_{\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}} \sup_{\phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)}} power(\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}, \phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)}) \\
&= \sup_{\mathcal{D}_1,\mathcal{D}_2} \sup_{M_1} \sup_{M_2} \sup_{\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}} \sup_{\phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)}} p * power(\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}) + (1-p) power(\phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)}) \\
&= \sup_{M_1} \sup_{M_2} \sup_{\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}} \sup_{\phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)}} p * power(\phi_{(M_1,\mathcal{D}_1,\mathcal{D}_2,\ell_1)}) + (1-p) power(\phi_{(M_2,\mathcal{D}_1,\mathcal{D}_2,\ell_2)})
\end{aligned}$$

The symmetry property allows us to fix any neighboring pair $\mathcal{D}_1, \mathcal{D}_2$, since optimizing over mechanisms and neighbors becomes redundant. This is because power$(\phi_{(M',\mathcal{D}_3,\mathcal{D}_4,\ell_1)})$ is the same as power$(\phi_{(M' \circ \nu^{-1},\mathcal{D}_1,\mathcal{D}_2,\ell_1)}$ for the recoding ν that maps dataset $\mathcal{D}_3$ to $\mathcal{D}_1$ and $\mathcal{D}_4$ to $\mathcal{D}_2$

$$= pg(\ell_1) + (1-p)g(\ell_2)$$

□

Theorem 6.5. *Given a transparent, symmetric, convex and postprocessing invariant privacy definition with privacy parameters θ, let $\mathfrak{M}$ denote the set of all mechanisms that satisfy the privacy definition at those privacy parameter settings.*

- *Define the* pointwise *significance level vs. power tradeoff function g as follows. For any significance level ℓ, $g(\ell)$ is the supremum of achievable power in testing $H_0 : \mathcal{D}_1$ vs. $H_1 : \mathcal{D}_2$ for all neighboring pairs $\mathcal{D}_1, \mathcal{D}_2$, for all mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level as at most ℓ.*
- *Define the* compound *significance level vs. power tradeoff function g^* as follows. For any significance level ℓ, $g^*(\ell)$ is the supremum of achievable power in testing H_0^*: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_0\}$ is the input to M vs. H_1^*: $\mathcal{D}^-$ was sampled from F and $\mathcal{D}^- \cup \{r_1\}$ is the input to M, maximizing over all distributions F, pairs of records r_1, r_2, mechanisms $M \in \mathfrak{M}$, and for all hypothesis tests whose input is the output of the chosen mechanism M and whose significance level is at most ℓ.*

Then, $g = g^*$.

Proof of Theorem 6.5. First, we note that $g^*(\ell) \geq g(\ell)$ for all ℓ because testing between neighbors $\mathcal{D}_1$ and $\mathcal{D}_2$ is a special case where F puts probability 1 on the set of records on which $\mathcal{D}_1$ and $\mathcal{D}_2$ do not differ.

For the other direction, note that a hypothesis test ϕ_M for H_0^* vs H_1^* (defined by F, r_1, r_2) based on the output of M can be converted to a test $\phi_{M,\mathcal{D}_1,\mathcal{D}_2}$ for H_0 vs H_1 for neighboring pairs $\mathcal{D}_1, \mathcal{D}_2$ (in which the record that they differ by has value r_1 or r_2) just by taking its recommended action (reject or fail to reject the null hypothesis). F induces a distribution over pairs of neighbors (since it is a distribution over the records shared by $\mathcal{D}_1$ and $\mathcal{D}_2$ and the remaining record is either specified by the null or alternate hypothesis). Letting ℓ be the significance level of ϕ_M when testing H_0^* vs. H_1^* and $\ell_{\mathcal{D}_1,\mathcal{D}_2}$ be the significance level of $\phi_{M,\mathcal{D}_1,\mathcal{D}_2}$, we have that

$$\begin{aligned}
\ell &= E_{(\mathcal{D}_1,\mathcal{D}_2)\sim F}[\ell_{\mathcal{D}_1,\mathcal{D}_2}] \\
power(\phi_M) &= E_{(\mathcal{D}_1,\mathcal{D}_2)\sim F}[power(\phi_{M,\mathcal{D}_1,\mathcal{D}_2})] \\
&\leq E_{(\mathcal{D}_1,\mathcal{D}_2)\sim F}[g(\ell_{\mathcal{D}_1,\mathcal{D}_2})] \\
&\leq g(E_{(\mathcal{D}_1,\mathcal{D}_2)\sim F}[\ell_{\mathcal{D}_1,\mathcal{D}_2}]) \\
&= g(\ell)
\end{aligned}$$

Maximizing over all F, r_1,r_2, M and ϕ_M that have level ℓ we therefore get $g^*(\ell) \leq g(\ell)$.

Combining this with our first result and applying them for all ℓ, we have $g = g^*$.

□

C Proofs from Section 7

Theorem 7.1. *If M satisfies ϵ-differential privacy, then for any attacker prior $P_{\mathfrak{A}}$ and all ω and every r that has nonzero probability,*

$$e^{-\epsilon} \leq \frac{P_{\mathfrak{A}}(R = r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(R = r \mid M(\text{psample}(\mathcal{X})) = \omega)} \leq e^{\epsilon}.$$

Proof of Theorem 7.1.

$$\begin{aligned}
\frac{P_{\mathfrak{A}}(R = r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(R = r \mid M(\text{psample}(\mathcal{X})) = \omega)} &= \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big)} \\
&\leq \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) e^{-\epsilon} P(M(\mathcal{D}^- \cup \{r\}) = \omega) \Big)} \\
&= \frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r \mid \mathcal{D}^-) e^{-\epsilon} P(M(\mathcal{D}^- \cup \{r\}) = \omega)} \\
&= e^{\epsilon}.
\end{aligned}$$

The $\geq e^{-\epsilon}$ part of the theorem statement is proved analogously using the fact that $P(M(\mathcal{D}^- \cup \{r''\}) \leq e^{\epsilon} P(M(\mathcal{D}^- \cup \{r\})$. □

Theorem 7.3. *Let $P_{\mathfrak{A}}$ be an attacker's prior such that $P_{\mathfrak{A}}(\mathcal{X}^- = \mathcal{D}^-) = 1$ for some $\mathcal{D}^-$. Let M be a mechanism and let P_ω be the distribution of the output ω of M under the assumption that the attacker's prior $P_{\mathfrak{A}}$ is correct (i.e., a record for the target individual is sampled $R \sim P_{\mathfrak{A}}(R \mid \mathcal{D}^-)$ and then $\omega = M(\mathcal{D}^- \cup \{R\})$ is produced). Then for any r,*

- *If M satisfies (α, γ)-RDP for $\alpha > 1$, then $P_\omega \left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \geq e^{\epsilon} \right) \leq e^{-(\epsilon - \gamma)\alpha - \gamma}$ (i.e., the probability of seeing an ω for which the posterior-to-posterior ratio exceeds e^{ϵ} is at most $\delta = e^{-(\epsilon-\gamma)\alpha-\gamma}$).*
- *If M satisfies ρ-zCDP and $\epsilon > \rho$, then $P_\omega \left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{r'} P_{\mathfrak{A}}(R = r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \geq e^{\epsilon} \right) \leq e^{-(\epsilon+\rho)^2/(4\rho)}$.*

Proof of Theorem 7.3.

$$
\begin{aligned}
&P_\omega\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}\geq e^\epsilon\right)\\
&=P_\omega\left(\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}\right)^\alpha\geq e^{\epsilon\alpha}\right)\\
&\leq e^{-\epsilon\alpha}E_\omega\left[\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}\right)^\alpha\right]\quad\text{by Markov's inequality}\\
&=\sum_\omega e^{-\epsilon\alpha}\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}\right)^\alpha\left(\sum_{r''}P_{\mathfrak{A}}(R=r''\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r''\})=\omega)\right)\\
&=\sum_\omega e^{-\epsilon\alpha}\frac{\big(P(M(\mathcal{D}^-\cup\{r\})=\omega)\big)^\alpha}{\big(\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)\big)^{\alpha-1}}\\
&\leq\sum_\omega e^{-\epsilon\alpha}\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)\frac{\big(P(M(\mathcal{D}^-\cup\{r\})=\omega)\big)^\alpha}{\big(P(M(\mathcal{D}^-\cup\{r'\})=\omega)\big)^{\alpha-1}}\quad\text{by convexity of } f(z)=1/z^{\alpha-1}\text{ when }\alpha>1\\
&=e^{-\epsilon\alpha}\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)\sum_\omega\frac{\big(P(M(\mathcal{D}^-\cup\{r\})=\omega)\big)^\alpha}{\big(P(M(\mathcal{D}^-\cup\{r'\})=\omega)\big)^{\alpha-1}}
\end{aligned}
$$

Note that $\frac{1}{\alpha-1}$ times log of the inner summation over ω is a Rényi divergence.

$$
\begin{aligned}
&\leq e^{-\epsilon\alpha}\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)e^{(\alpha-1)\gamma}\quad\text{by definition of RDP}\\
&=e^{-\epsilon\alpha}e^{(\alpha-1)\gamma}=e^{-(\epsilon-\gamma)\alpha-\gamma}.
\end{aligned}
$$

Now we note that if M satisfies ρ-zCDP then it satisfies $(\alpha,\rho\alpha)$-RDP for every $\alpha>1$, and we can choose the α that minimizes the above expression when γ is replaced by $\rho\alpha$, which is $\alpha^*=\frac{\epsilon+\rho}{2\rho}$. Note that $\alpha^*>1$ when $\epsilon>\rho$. Plugging in $\rho\alpha^*$ for γ, we get:

$$
\begin{aligned}
P_\omega\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}\geq e^\epsilon\right)&\leq e^{-(\epsilon-\frac{\epsilon+\rho}{2})\frac{\epsilon+\rho}{2\rho}-\frac{\epsilon+\rho}{2}}\\
&=e^{-\frac{\epsilon-\rho}{2}\frac{\epsilon+\rho}{2\rho}-\frac{\epsilon+\rho}{2}}\\
&=e^{-\frac{\epsilon+\rho}{2}(1+\frac{\epsilon-\rho}{2\rho})}\\
&=e^{-\frac{\epsilon+\rho}{2}\frac{\epsilon+\rho}{2\rho}}=e^{-(\epsilon+\rho)^2/(4\rho)}.
\end{aligned}
$$

□

Theorem 7.4. *Let M be a mechanism that satisfies f-DP. Let $P_{\mathfrak{A}}$ be an attacker's prior such that $P_{\mathfrak{A}}(\mathcal{X}^-=\mathcal{D}^-)=1$ for some $\mathcal{D}^-$. Let $P_{\omega\mid r}$ be the distribution of the output ω of M under the assumption that the attacker's prior $P_{\mathfrak{A}}$ is correct and the true record is value r. Define $f^{-1}(z)=\inf\{y\in[0,1]\ :\ f(y)\leq z\}$. Then for any r and $\delta\in[0,1]$, if $\epsilon\geq\log\frac{\delta}{f^{-1}(1-\delta)}$, then*

$$
P_{\omega|r}\left(\frac{P(M(\mathcal{D}^-\cup\{r\})=\omega)}{\sum_{r'}P_{\mathfrak{A}}(R=r'\mid\mathcal{D}^-)P(M(\mathcal{D}^-\cup\{r'\})=\omega)}>e^\epsilon\right)\leq\delta.
$$

Furthermore, this same bound holds if M is replaced by $A\circ M$.

Proof of Theorem 7.4. First, note that the only uncertainty in $P_{\mathfrak{A}}$ is in the record r of the target person. Hence $P_{\mathfrak{A}}(M(\mathcal{X})=\omega\mid R=r)=P(M(\mathcal{D}^-\cup\{r\})=\omega)$.

Let r'' be a record different from r and for any r', let $M_{r'}$ be the algorithm that replaces r'' with r'. Then it is easy to see that $M_{r'}$ also satisfies f-DP. Now, let M^* be an algorithm that, on any input $\mathcal{D}^*$ (it does not have to be related to the attacker's $\mathcal{D}^-$ in any way), chooses among the $M_{r'}$ with probability $P(r' \mid \mathcal{D}^-)$ and runs the chosen algorithm on $\mathcal{D}^*$. By convexity of f-DP, M^* also satisfies f-DP.

Now, let $\mathcal{D}_1 = \mathcal{D}^- \cup \{r\}$, and let $\mathcal{D}_2 = \mathcal{D}^- \cup \{r''\}$. Then the distribution of $\log \frac{P(M(\mathcal{D}^- \cup \{r\})=\omega)}{\sum_{r'} P_{\mathfrak{A}}(R=r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\})=\omega)}$ for $\omega \sim P_{\omega \mid r}$ is the same as for the privacy loss random variable $\mathcal{L}^{M^*}_{\mathcal{D}_1,\mathcal{D}_2}$. By the equivalence between f-DP and the pbdp reparametrization, $P(\mathcal{L}^{\mathcal{D}_2}_{M^*,\mathcal{D}_1} > \epsilon) \leq \delta$ when $\epsilon \geq \log \frac{\delta}{f^{-1}(1-\delta)}$. The postprocessing invariance of pbdp finishes the proof. □

Theorem 7.5. *Let $P_{\mathfrak{A}}$ be an attacker's prior. Let r be the target person's true record. Let M be a mechanism. Let $P_{\omega \mid r}$ be the distribution of ω conditioned on r. That is, a dataset $\mathcal{D}$ is generated from the distribution $P_{\mathfrak{A}}(\mathcal{X} \mid r)$ and $M(\mathcal{D})$ produces ω. Then M can provide the following guarantees on the posterior-to-posterior ratio $\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)}$ for the true record r.*

- *If M satisfies (α, γ)-RDP for $\alpha > 1$ then $P_{\omega \mid r}\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)} \geq e^{\epsilon}\right) \leq e^{(\alpha-1)(\gamma-\epsilon)}$.*
- *If M satisfies ρ-zCDP then for $\epsilon > \rho$, $P_{\omega \mid r}\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X})=\omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X}))=\omega)} \geq e^{\epsilon}\right) \leq e^{-(\epsilon-\rho)^2/(4\rho)}$.*

Proof of Theorem 7.5. Recall that the target is known to be in the data, that $\mathcal{U}$ is the notation for set of possible datasets, and that $\mathcal{U}^-$ is the set of possible datasets after the target's data have been removed.

We first note that

$$\begin{aligned}
P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X})) = \omega) &= \sum_{\mathcal{D}^- \in \mathcal{U}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\text{psample}(\mathcal{D}^- \cup \{r'\})) = \omega) \\
&= \sum_{\mathcal{D}^- \in \mathcal{U}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^-) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) \Big(\sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \Big) \\
&= \sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^-) \sum_{r''} P_{\mathfrak{A}}(r'' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r''\}) = \omega) \\
&= P_{\mathfrak{A}}(M(\mathcal{X}) = \omega)
\end{aligned}$$

So the marginal probability of ω is the same in the actual and counterfactual world (but the joint distribution between Jessie's record and ω differs since in the actual world it is part of the input to M).

Next,

$$\begin{aligned}
&P_{\omega \mid r}\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X})) = \omega)} \geq e^{\epsilon}\right) \\
&= P_{\omega \mid r}\left(\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X})) = \omega)}\right)^{\alpha-1} \geq e^{\epsilon(\alpha-1)}\right) \quad \text{for any } \alpha > 1 \\
&\leq e^{-\epsilon(\alpha-1)} E_{\omega \mid r}\left[\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X})) = \omega)}\right)^{\alpha-1}\right] \\
&= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(r \mid M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r \mid M(\text{psample}(\mathcal{X})) = \omega)}\right)^{\alpha-1}\right] P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r) \\
&= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(r\ ,\ M(\mathcal{X}) = \omega)/P_{\mathfrak{A}}(M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r\ ,\ M(\text{psample}(\mathcal{X})) = \omega)/P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X})) = \omega)}\right)^{\alpha-1}\right] P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r) \\
&= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(r\ ,\ M(\mathcal{X}) = \omega)}{P_{\mathfrak{A}}(r\ ,\ M(\text{psample}(\mathcal{X})) = \omega)}\right)^{\alpha-1}\right] P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r) \\
&\quad \text{(Since } P_{\mathfrak{A}}(M(\mathcal{X}) = \omega) = P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X})) = \omega)\text{)}
\end{aligned}$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(M(\mathcal{X}) = \omega \mid r)}{P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X})) = \omega \mid r)} \right)^{\alpha-1} \right] P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r)}{P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)} \right)^{\alpha-1} \right] P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left[\left(\frac{P_{\mathfrak{A}}(M(\mathcal{X}^- \cup \{r\}) = \omega \mid r)}{P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)} \right)^{\alpha} \right] P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left(\frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) \sum_{r'} P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \right)^{\alpha} P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \left(\frac{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) (\sum_{r'} P_{\mathfrak{A}}(r' \mid \mathcal{D}^-)) P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{\sum_{\mathcal{D}^- \in \mathcal{U}^-} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) \sum_{r'} P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \right)^{\alpha} P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)$$

Now, use Jensen's inequality ($\alpha > 1$), note numerator/denominator only differ on record used by M.

$$\leq e^{-\epsilon(\alpha-1)} \sum_{\omega} \sum_{\mathcal{D}^-} \sum_{r'} \frac{P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}{\sum_{\mathcal{D}^-} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) \sum_{r'} P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \right)^{\alpha}$$
$$* P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\omega} \sum_{\mathcal{D}^-} \sum_{r'} \frac{P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) P(M(\mathcal{D}^- \cup \{r'\}) = \omega)}{P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)} \left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \right)^{\alpha}$$
$$* P_{\mathfrak{A}}(M(\text{psample}(\mathcal{X}^- \cup \{r\})) = \omega \mid r)$$

$$= e^{-\epsilon(\alpha-1)} \sum_{\mathcal{D}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) \left(\sum_{\omega} P(M(\mathcal{D}^- \cup \{r'\}) = \omega) \left(\frac{P(M(\mathcal{D}^- \cup \{r\}) = \omega)}{P(M(\mathcal{D}^- \cup \{r'\}) = \omega)} \right)^{\alpha} \right)$$

$$\leq e^{-\epsilon(\alpha-1)} \sum_{\mathcal{D}^-} \sum_{r'} P_{\mathfrak{A}}(\mathcal{D}^- \cup \{r\} \mid r) P_{\mathfrak{A}}(r' \mid \mathcal{D}^-) e^{(\alpha-1)\gamma} \quad \text{by } (\alpha, \gamma)\text{-RDP}$$

$$= e^{-\epsilon(\alpha-1)} e^{(\alpha-1)\gamma} = e^{(\alpha-1)(\gamma-\epsilon)}$$

In the case of ρ-zCDP, this is satisfied with $\gamma = \rho\alpha$ for all $\alpha > 1$ and thus we can choose the best α. The optimal choice is $\alpha = \frac{\epsilon+\rho}{2\rho}$, which is greater than 1 whenever $\epsilon > \rho$. Plugging in this value of α and using $\gamma = \rho\alpha$, we see that the probability that the posterior-to-posterior ratio exceeds e^{ϵ} is at most:

$$e^{(\frac{\epsilon+\rho}{2\rho}-1)(\frac{\epsilon+\rho}{2}-\epsilon)} = e^{\frac{\epsilon-\rho}{2\rho}\frac{-\epsilon+\rho}{2}}$$
$$= e^{-(\epsilon-\rho)^2/(4\rho)}.$$

□