On the 'Semantics' of Differential Privacy: A Bayesian Formulation[*]

Shiva Prasad Kasiviswanathan[†]
General Electric Global Research
`kasivisw@gmail.com`

Adam Smith[‡]
Department of Computer Science and Engineering
The Pennsylvania State University
`asmith@psu.edu`

Abstract

Differential privacy is a definition of "privacy" for algorithms that analyze and publish information about statistical databases. It is often claimed that differential privacy provides guarantees against adversaries with arbitrary side information. In this paper, we provide a precise formulation of these guarantees in terms of the inferences drawn by a Bayesian adversary. We show that this formulation is satisfied by both ϵ-differential privacy as well as a relaxation, (ϵ, δ)-differential privacy. Our formulation follows the ideas originally due to Dwork and McSherry, stated implicitly in [5]. This paper is, to our knowledge, the first place such a formulation appears explicitly. The analysis of the relaxed definition is new to this paper, and provides guidance for setting the δ parameter when using (ϵ, δ)-differential privacy.

1 Introduction

Privacy is an increasingly important aspect of data publishing. Reasoning about privacy, however, is fraught with pitfalls. One of the most significant is the auxiliary information (also called external knowledge, background knowledge, or side information) that an adversary gleans from other channels such as the web, public records, or domain knowledge. Schemes that retain privacy guarantees in the presence of independent releases are said to *compose securely*. The terminology, borrowed from cryptography (which borrowed, in turn, from software engineering), stems from the fact that schemes that compose securely can be designed in a stand-alone fashion without explicitly taking other releases into account. Thus, understanding independent releases is essential for enabling modular design. In fact, one would like schemes that compose securely not only with independent instances of themselves, but with *arbitrary external knowledge*.

Certain randomization-based notions of privacy (such as *differential privacy*, due to Dwork, McSherry, Nissim, and Smith [10]) are viewed as providing meaningful guarantees even in the presence of arbitrary side information. In this paper. we give a precise formulation of this statement. First, we provide a Bayesian formulation of "pure" differential privacy which explicitly models side information. Second, we prove that the relaxed definitions of Blum et al. [2], Dwork et al. [9] and Machanavajjhala et al. [16] imply the Bayesian formulation. The proof is non-trivial, and relies on the "continuity" of Bayes' rule with respect to certain

---

distance measures on probability distributions. Our result means that techniques satisfying the relaxed definitions can be used with the same sort of assurances as in the case of pure differentially-private algorithms, as long parameters are set appropriately. Specifically, (ϵ, δ)-differential privacy provides meaningful guarantees whenever δ, the additive error parameter, is smaller than about ϵ^2/n, where n is the size of the data set.

Organization. After introducing the basic definitions, we state and discuss our main results in Section 2. In Section 2.1, we relate our approach to other efforts—subsequent to the initial version of this work—that sought to pin down mathematical precise formulations of the "meaning" of differential privacy. Section 3 proves our main theorems. Along the way, we develop lemmas about (ϵ, δ)-*indistinguishability*—the notion of similarity that underlies (ϵ, δ)-differential privacy—-that we believe are of independent interest. The most useful of these, which we dub the Conditioning Lemma, is given in Section 3.3. Finally, we provide further discussion of our approach in Section 4.

1.1 Differential Privacy

Databases are assumed to be vectors in $\mathcal{D}^n$ for some domain $\mathcal{D}$. The Hamming distance $d_H(\mathrm{x}, \mathrm{y})$ on $\mathcal{D}^n$ is the number of positions in which the vectors x, y differ. We let $\Pr[\cdot]$ and $\mathbb{E}[\cdot]$ denote probability and expectation, respectively. Given a randomized algorithm $\mathcal{A}$, we let $\mathcal{A}(\mathrm{x})$ be the random variable (or, probability distribution on outputs) corresponding to input x. If X and Y are probability distributions (or random variables) on a discrete space D, the *statistical difference* (a.k.a. *total variation distance*) between X and Y is defined as:

$$\mathbf{SD}\left(X, Y\right) = \max_{S \subset D} \left| \Pr[X \in S] - \Pr[Y \in S] \right|.$$

Definition 1.1 (ϵ-differential privacy [10])**.** *A randomized algorithm $\mathcal{A}$ is said to be ϵ-differentially private if for all databases $\mathrm{x}, \mathrm{y} \in \mathcal{D}^n$ at Hamming distance at most 1, and for all subsets S of outputs,*

$$\Pr[\mathcal{A}(\mathrm{x}) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{A}(\mathrm{y}) \in S].$$

This definition states that changing a single individual's data in the database leads to a small change in the *distribution* on outputs. Unlike more standard measures of distance such as statistical difference or Kullback-Leibler divergence, the metric here is multiplicative and so even very unlikely events must have approximately the same probability under the distributions $\mathcal{A}(\mathrm{x})$ and $\mathcal{A}(\mathrm{y})$. This condition was relaxed somewhat in other papers [4, 8, 2, 9, 3, 17, 16]. The schemes in all those papers, however, satisfy the following relaxation (first formulated by Dwork, Kenthapadi, McSherry, Mironov, and Naor [9]):

Definition 1.2 ((ϵ, δ)-differential privacy [9])**.** *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all databases $\mathrm{x}, \mathrm{y} \in \mathcal{D}^n$ that differ in one entry, and for all subsets S of outputs,*

$$\Pr[\mathcal{A}(\mathrm{x}) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{A}(\mathrm{y}) \in S] + \delta.$$

2 Semantics of Differential Privacy

There is a crisp, semantically-flavored[1] interpretation of differential privacy, due to Dwork and McSherry, explained in [5]: *Regardless of external knowledge, an adversary with access to the sanitized database*

---

[1]The use of the term "semantic" for definitions that deal directly with adversarial knowledge dates back to *semantic security* of encryption [12].

draws the same conclusions whether or not my data is included in the original database. One might hope for a stronger statement, namely that the adversary draws the same conclusions whether or not the data is used at all. However, such a strong statement is impossible to provide in the presence of arbitrary external information (Dwork and Naor [6], Dwork [5]; see also Kifer and Machanavajjhala [14]), as illustrated by the following example.

Example 1. Consider a clinical study that explores the relationship between smoking and lung disease. A health insurance company who had no a priori understanding of that relationship might dramatically alter its "beliefs" (as encoded by insurance premiums) to account for the results of the study. The study would cause the company to raise premiums for smokers and lower them for nonsmokers, regardless of whether they participated in the study. In this case, the conclusions drawn by the company about the riskiness of any one individual (say Alice) are strongly affected by the results of the study. This occurs regardless of whether Alice's data are included in the study. ◇

In this section, we develop a formalization Dwork and McSherry's interpretation and explore its relation to standard definitions. To proceed, we require a mathematical formulation of "external knowledge", and of "drawing conclusions". The first is captured via a *prior* probability distribution b on $\mathcal{D}^n$ (b is a mnemonic for "beliefs"). Conclusions are modeled by the corresponding *posterior* distribution: given a transcript t, the adversary updates his belief b about the database x using Bayes' rule to obtain a posterior $\bar{b}$:

$$\bar{b}[\mathrm{x}|t] = \frac{\Pr[\mathcal{A}(\mathrm{x}) = t]b[\mathrm{x}]}{\sum_{\mathrm{z}} \Pr[\mathcal{A}(\mathrm{z}) = t]b[\mathrm{z}]} \ . \tag{1}$$

When the mechanism $\mathcal{A}$ is interactive, the definition of $\mathcal{A}$ depends on the adversary's choices; for legibility we omit the dependence on the adversary in the notation. Also, for simplicity, we discuss only discrete probability distributions. Our results extend directly to the interactive, continuous case.

For a database x, define x_{-i} to be the same vector except position i has been replaced by some fixed, default value $\mathcal{D}$. Any valid value in $\mathcal{D}$ will do for the default value. We define $n+1$ related games, numbered 0 through n. In Game 0, the adversary interacts with $\mathcal{A}(\mathrm{x})$. This is the interaction that actually takes place between the adversary and the randomized algorithm $\mathcal{A}$. The distribution $\bar{b}_0$ is just the distribution $\bar{b}$ as defined in (1).

In Game i (for $1 \leq i \leq n$), the adversary interacts with $\mathcal{A}(\mathrm{x}_{-i})$. Game i describes the hypothetical scenario where person i's data is not used.[2] In Game $i > 0$, given a transcript t, the adversary updates his belief b about database x again using Bayes' rule to obtain a posterior $\bar{b}_i$ as follows:

$$\bar{b}_i[\mathrm{x}|t] = \frac{\Pr[\mathcal{A}(\mathrm{x}_{-i}) = t]b[\mathrm{x}]}{\sum_{\mathrm{z}} \Pr[\mathcal{A}(\mathrm{z}_{-i}) = t]b[\mathrm{z}]}. \tag{2}$$

Through these $n + 1$ games, we get $n + 1$ *a posteriori* distributions $\bar{b}_0, \ldots, \bar{b}_n$, where $\bar{b}_0$ is same as $\bar{b}$ (defined in (1)), and $\bar{b}_i$ ($i > 0$) is the posterior distribution obtained when the adversary interacts with $\mathcal{A}(\mathrm{x}_{-i})$ and uses this interaction to update his belief distribution (defined in (2)).

Given a particular transcript t, we say privacy has been breached if the adversary would draw different conclusions about the world and, in particular, about a person i, depending on whether or not i's data was used. One could formally define "different" in many ways. In this paper, we choose a weak (but popular)

---

[2]It could happen by coincidence that person i's data equals the default value and hence that $\mathrm{x} = \mathrm{x}_{-i}$. This doesn't affect the meaning of the result since the default value is chosen independently of the data. Readers bothered by the possible coincidence may choose to think of the default value as a special value $\bot$ (e.g., "no data") that does not correspond to any real record.

3

measure of distance between probability distributions, namely statistical difference. We say the adversary has learned something, if for any transcript t the distributions $\bar{b}_0[\cdot|t]$ and $\bar{b}_i[\cdot|t]$ are far apart in statistical difference. We would like to avoid this from happening for any potential participant. This is captured by the following definition.

Definition 2.1 (ϵ-semantic privacy)**.** *A randomized algorithm $\mathcal{A}$ is said to be ϵ-semantically private if for all belief distributions b on $\mathcal{D}^n$, for all possible transcripts t, and for all $i = 1, \ldots, n$:*

$$\mathbf{SD}\left(\bar{b}_0[\cdot|t]\,,\,\bar{b}_i[\cdot|t]\,\right) \leq \epsilon.$$

Our formulation of semantic privacy is inspired by Dwork and McSherry's interpretation of differential privacy [5]. We now formally show that the notions of ϵ-differential privacy (Definition 1.1) and ϵ-semantic privacy (Definition 2.1) are essentially equivalent.

Theorem 2.2. *For all $\epsilon > 0$, ϵ-differential privacy implies $\bar{\epsilon}$-semantic privacy, where $\bar{\epsilon} = e^\epsilon - 1$. For $0 < \epsilon \leq 0.45$, $\epsilon/2$-semantic privacy implies 3ϵ-differential privacy.*

The proof of this and all other results in this section may be found in Section 3.

We can extend the previous Bayesian formulation to capture situations where bad events can occur with some negligible probability. Specifically, we formulate (ϵ, δ)-semantic privacy and show that it is closely related to (ϵ, δ)-differential privacy.

Definition 2.3 ((ϵ, δ)-semantic privacy)**.** *A randomized algorithm is (ϵ, δ)-semantically private if for all belief distributions b on $\mathcal{D}^n$, with probability at least $1 - \delta$ over $t \sim \mathcal{A}(\mathrm{x})$ (t drawn from $\mathcal{A}(\mathrm{x})$), where the database x is drawn according to b, and for all $i = 1, \ldots, n$:*

$$\mathbf{SD}\left(\bar{b}_0[\cdot|t]\,,\,\bar{b}_i[\cdot|t]\,\right) \leq \epsilon.$$

The (ϵ, δ)-privacy definition is most interesting when $\epsilon \gg \delta$, since every (ϵ, δ)-private algorithm is also $(0, \delta + (e^\epsilon - 1))$-differentially private. Below, we assume $\epsilon > \delta$. In fact, many of our results are meaningful only when δ is less than $1/n$, while ϵ must generally be much larger than $1/n$ to allow for useful algorithms.

Theorem 2.4 (Main Theorem)**.**

(1) If $\epsilon, \delta > 0$ and $\delta < (1 - e^{-\epsilon})^2/n$, then (ϵ, δ)-differential privacy implies (ϵ', δ')-semantic privacy on databases of size n with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{n\delta}$ and $\delta' = 4\sqrt{n\delta}$.

(2) If $\epsilon, \delta > 0$ and $\epsilon \leq 0.45$, then (ϵ, δ)-semantic privacy implies $(3\epsilon, 2\delta)$-differential privacy.

In Appendix B, we discuss a stronger notion of (ϵ, δ)-semantic privacy and show that (ϵ, δ)-differential privacy need not imply this stronger semantic privacy guarantee.

Remark 1. The implications in Theorems 2.2 and 2.4 would not hold if differential privacy were defined in terms of statistical difference (total variation distance) or mutual information instead of the multiplicative metric used in Definitions 1.1 and 1.2. For example, one could change the last line of the Definition 1.2 to

$$\Pr[\mathcal{A}(\mathrm{x}) \in S] \leq \Pr[\mathcal{A}(\mathrm{y}) \in S] + \epsilon_{SD}\,. \tag{3}$$

For this modified definition to allow publishing useful information, one would need $\epsilon_{SD} = \Omega(1/n)$ (otherwise, data sets that differ in all n elements would still be hard to distinguish). However, in that parameter

range there is a mechanism that satisfies the new definition but does not satisfy "semantic" privacy for any reasonable parameters. Namely, consider the mechanism which on input x $= (x_1, \ldots, x_n)$ samples a uniformly random index $i \in \{1, ..., n\}$ and outputs (i, x_i). This mechanism is intuitively unsatisfactory, since it always outputs some individual's data in the clear. It also does not satisfy semantic privacy for any pair (ϵ, δ) where $\epsilon < 1$ and $\delta < 1$. Nevertheless, it does satisfy the requirement of (3) with $\epsilon_{SD} = 1/n$. The same mechanism also satisfies the natural variant of differential privacy based on mutual information (for example, where the mutual information between $\mathcal{A}(\mathrm{x})$ and x_i is required to be small for all indices i and product distributions on x). $\diamondsuit$

2.1 Related Approaches

Prior to Posterior Comparisons. In the original paper on differential privacy, Dwork et al. [10] defined a notion of "semantic" privacy that involved comparing the prior and posterior distributions of the adversary. In the language of the preceding section, they require that $\mathbf{SD}\left(b[\cdot], \ \bar{b}_i[\cdot | t]\right) \leq \epsilon$ for a subclass of belief distributions, called "informed beliefs", in which all but one of the data set entries are fixed (constant). They show that this definition is equivalent to differential privacy. Kifer and Machanavajjhala [15] use this prior-to-posterior approach to generalize differential privacy to other settings.

However, the impossibility results of Dwork and Naor [5, 7] and Kifer and Machanavajjhala [14], exemplified by the smoking example in Example 1, imply that no mechanism that provides nontrivial information about the data set satisfies such a prior-to-posterior definition for *all* distributions.

This impossibility motivated the posterior-to-posterior comparison espoused in this paper, and subsequently generalized by Bassily et al. [1]. In contrast to the prior-to-posterior approach, the framework discussed in this paper does generalize to arbitrary distributions on the data (and, hence, to arbitrary side information). Bassily et al. [1] suggest the term "inference-based" for definitions which explicitly discuss the posterior distributions constructed by Bayesian adversaries.

Hypothesis Testing. Wasserman and Zhou [18] relate differential privacy to the type I and II errors of a hypothesis test. Specifically, fix an ϵ-differentially private mechanism $\mathcal{A}$, an i.i.d. distribution on the data x, an index i, and disjoint sets S and T of possible values for the i-th entry x_i of x. Wasserman and Zhou [18] show that any hypothesis test (given $\mathcal{A}(\mathrm{x})$, and full knowledge of the input product distribution on x and the differentially private mechanism $\mathcal{A}$) for the hypothesis $H_1 : x_i \in S$ versus the alternative $H_1 : x_i \in T$ must satisfy

$$1 - \beta \leq e^{\epsilon} \alpha, \tag{4}$$

where α is the significance level (maximum type-I error) and $1 - \beta$ is the power (maximum type-II error) of the test. In other words, the test rejects the hypothesis with approximately the same probability regardless of whether the hypothesis is true. This perspective was extended to (ϵ, δ)-differential privacy by Hall et al. [13].

This is a reasonable requirement. Note, however, that it holds only for product distributions, which limits its applicability. More importantly, a very similar statement can be proven for the statistical difference-based definition discussed in Remark 1. Specifically, one can show that

$$1 - \beta \leq \alpha + \epsilon_{SD}. \tag{5}$$

when the mechanism satisfies the definition of Remark 1. Equation (4) has the same natural language interpretation as equation (5), namely, "the test rejects the hypothesis with approximately the same probability regardless of whether the hypothesis is true". However, as mentioned in the Remark 1, the statistical

difference-based definition allows mechanisms that publish detailed personal data in the clear. This makes the meaning of a hypothesis-testing-based definition hard to evaluate intuitively. We hope the definitions provided here are easier to interpret.

3   Proofs of Main Results

We begin this section by defining (ϵ, δ)-*indistinguishability* and stating a few of its basic properties (Section 3.1, with proofs in Appendix A). Section 3.2 gives the proof of our main result for ϵ-differential privacy. In Section 3.3 we state and prove the Conditioning Lemma, the main tool which allows us to prove our results about (ϵ, δ)-differential privacy (Section 3.4).

3.1 (ϵ, δ)-Indistinguishability and its Basic Properties

The relaxed notions of (ϵ, δ)-differential privacy implicitly uses a two-parameter distance measure on probability distributions (or random variables) which we call (ϵ, δ)-indistinguishability. In this section, we develop a few basic properties of this measure. These properties listed in Lemma 3.3 will play an important role in establishing the proofs of Theorems 2.2 and 2.4

Definition 3.1 ((ϵ, δ)-indistinguishability)**.** *Two random variables X, Y taking values in a set D are (ϵ, δ)-indistinguishable if for all sets $S \subseteq D$,*

$$\Pr[X \in S] \leq e^{\epsilon} \Pr[Y \in S] + \delta \quad and \quad \Pr[Y \in S] \leq e^{\epsilon} \Pr[X \in S] + \delta.$$

We will also be using a variant of (ϵ, δ)-indistinguishability, which we call *point-wise (ϵ, δ)-indistinguishability*. Lemma 3.3 (Parts 1 and 2) shows that (ϵ, δ)-indistinguishability and point-wise (ϵ, δ)-indistinguishability are almost equivalent.

Definition 3.2 (Point-wise (ϵ, δ)-indistinguishability)**.** *Two random variables X and Y are point-wise (ϵ, δ)-indistinguishable if with probability at least $1 - \delta$ over a drawn from either X or Y, we have:*

$$e^{-\epsilon} \Pr[Y = a] \leq \Pr[X = a] \leq e^{\epsilon} \Pr[Y = a].$$

Lemma 3.3. *Indistinguishability satisfies the following properties:*

1. *If X, Y are point-wise (ϵ, δ)-indistinguishable then they are (ϵ, δ)-indistinguishable.*

2. *If X, Y are (ϵ, δ)-indistinguishable then they are point-wise $\left(\epsilon, \ \delta \cdot \frac{2}{1 - e^{-\epsilon}} \right)$-indistinguishable.*

3. *Let X be a random variable on D. Suppose that for every $a \in D$, $\mathcal{A}(a)$ and $\mathcal{A}'(a)$ are (ϵ, δ)-indistinguishable (for some randomized algorithms $\mathcal{A}$ and $\mathcal{A}'$). Then the pairs $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ are (ϵ, δ)-indistinguishable.*

4. *Let X be a random variable. Suppose with probability at least $1 - \delta_1$ over $a \sim X$, $\mathcal{A}(a)$ and $\mathcal{A}'(a)$ are (ϵ, δ)-indistinguishable (for some randomized algorithms $\mathcal{A}$ and $\mathcal{A}'$). Then the pairs $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ are $(\epsilon, \delta + \delta_1)$-indistinguishable.*

5. *If X, Y are (ϵ, δ)-indistinguishable (or X, Y are point-wise (ϵ, δ)-indistinguishable), then $\mathbf{SD}(X, Y) \leq \bar{\epsilon} + \delta$, where $\bar{\epsilon} = e^{\epsilon} - 1$.*

The lemma is proved in Appendix A.

3.2 Case of ϵ-Differential Privacy: Proof of Theorem 2.2

Theorem 2.2 (restated) (Dwork-McSherry). *$\epsilon/2$-differential privacy implies $\bar{\epsilon}$-semantic privacy, where $\bar{\epsilon} = e^\epsilon - 1$. $\epsilon/2$-semantic privacy implies 3ϵ-differential privacy as long as $\epsilon \leq 0.45$.*

Proof. Consider any database $x \in \mathcal{D}^n$. Let $\mathcal{A}$ be an $\epsilon/2$-differentially private algorithm. Consider any belief distribution b. Let the posterior distributions $\bar{b}_0[x|t]$ and $\bar{b}_i[x|t]$ for some fixed i and t be as defined in (1) and (2). $\epsilon/2$-differential privacy implies that for every database $z \in \mathcal{D}^n$

$$e^{-\epsilon/2} \Pr[\mathcal{A}(z_{-i}) = t] \leq \Pr[\mathcal{A}(z) = t] \leq e^{\epsilon/2} \Pr[\mathcal{A}(z_{-i}) = t].$$

These inequalities imply that the ratio of $\bar{b}_0[x|t]$ and $\bar{b}_i[x|t]$ (defined in (1) and (2)) is within $e^{\pm\epsilon}$. Since these inequalities holds for every x, we get:

$$\forall x \in \mathcal{D}^n, \ e^{-\epsilon}\bar{b}_i[x|t] \leq \bar{b}_0[x|t] \leq e^{\epsilon}\bar{b}_i[x|t].$$

This implies that the random variables (distributions) $\bar{b}_0[\cdot|t]$ and $\bar{b}_i[\cdot|t]$ are point-wise $(\epsilon, 0)$-indistinguishable. Applying Lemma 3.3 (Part 5) with $\delta = 0$, gives $\mathbf{SD}\left(\bar{b}_0[\cdot|t], \bar{b}_i[\cdot|t]\right) \leq \bar{\epsilon}$. Repeating the above arguments for every belief distribution, for every i, and for every t, shows that $\mathcal{A}$ is $\bar{\epsilon}$-semantically private.

To see that ϵ-semantic privacy implies 3ϵ-differential privacy, consider a belief distribution b which is uniform over two databases x, y which are at Hamming distance of one. Let i be the position in which x and y differ. Fix a transcript t. The distribution $\bar{b}_i[\cdot|t]$ will be uniform over x and y since they induce the same distribution on transcripts in Game i. This means that $\bar{b}_0[\cdot|t]$ will assign probabilities $1/2 \pm \epsilon$ to each of the two databases (by Definition 2.1). Working through Bayes' rule shows that (note that $b[x] = b[y]$)

$$\frac{\Pr[\mathcal{A}(x) = t]}{\Pr[\mathcal{A}(y) = t]} = \frac{\bar{b}_0[x|t]}{\bar{b}_0[y|t]} \leq \frac{\frac{1}{2} + \epsilon}{\frac{1}{2} - \epsilon} \leq e^{3\epsilon} (\text{since } \epsilon \leq 0.45). \tag{6}$$

Since the bound in (6) holds for every t, $\mathcal{A}(x)$ and $\mathcal{A}(y)$ are point-wise $(3\epsilon, 0)$-indistinguishable. Using Lemma 3.3 (Part 1), implies that $\mathcal{A}(x)$ and $\mathcal{A}(y)$ are $(3\epsilon, 0)$-indistinguishable. Since this relationship holds for every pair of neighboring databases x and y, means that $\mathcal{A}$ is 3ϵ-differentially private. $\square$

3.3 A Useful Tool: The Conditioning Lemma

We will use the following lemma to establish connections between (ϵ, δ)-differential privacy and (ϵ, δ)-semantic privacy. Let $B|_{A=a}$ denote the conditional distribution of B given that $A = a$ for jointly distributed random variables A and B.

Lemma 3.4 (Conditioning Lemma). *Suppose the pair of random variables (A, B) is (ϵ, δ)-indistinguishable from the pair (A', B'). Then, for $\hat{\epsilon} = 3\epsilon$ and for every $\hat{\delta} > 0$, the following holds: with probability at least $1 - \delta''$ over $t \sim B$ (or, alternatively, over $t \sim B'$), the random variables $A|_{B=t}$ and $A'|_{B'=t}$ are $(\hat{\epsilon}, \hat{\delta})$-indistinguishable, where $\delta'' = \frac{2\delta}{\hat{\delta}} + \frac{2\delta}{1 - e^{-\epsilon}}$.*

We can satisfy the conditions of the preceding lemma by setting $\hat{\delta} = \delta'' = O(\sqrt{\delta})$ for any constant ϵ. However, the proof of our main theorem will use a slightly different setting (with δ'' smaller than $\hat{\delta}$).

Proof. Let (A, B) and (A', B') take values in the set $D \times E$. In the remainder of the proof, we will use the notation $A|_t$ for $A|_{B=t}$ and $A'|_t$ for $A'|_{B'=t}$. Define,

$$Bad_1 = \left\{ t \in E : \exists S_t \subset D \text{ such that } \Pr[A|_t \in S_t] > e^{\hat{\epsilon}} \Pr[A'|_t \in S_t] + \hat{\delta} \right\}$$

$$Bad_2 = \left\{ t \in E : \exists S_t \subset D \text{ such that } \Pr[A'|_t \in S_t] > e^{\hat{\epsilon}} \Pr[A|_t \in S_t] + \hat{\delta} \right\}.$$

To prove the lemma, it suffices to show that the probabilities $\Pr[B \in Bad_1 \cup Bad_2]$ and $\Pr[B' \in Bad_1 \cup Bad_2]$ are each at most δ''. To do so, we first consider the set

$$Bad_0 = \left\{ t \in E \ : \ \Pr[B = t] < e^{-2\epsilon}\Pr[B' = t] \text{ or } \Pr[B = t] > e^{2\epsilon}\Pr[B' = t] \right\}.$$

We will separately bound the probabilities of Bad_0, $Bad_1' = Bad_1 \setminus Bad_0$ and $Bad_2' = Bad_2 \setminus Bad_0$.

To bound the mass of Bad_0, note that B and B' are (ϵ, δ)-indistinguishable (since they are functions of (A, B) and (A', B'))[3]. Since (ϵ, δ)-indistinguishability implies point-wise $(2\epsilon, \frac{2\delta}{1-e^{-\epsilon}})$-indistinguishability (Lemma 3.3, Part 2), we have

$$\Pr[B \in Bad_0] \leq \frac{2\delta}{1 - e^{-\epsilon}}.$$

We now turn to $Bad_1' = Bad_1 \setminus Bad_0$. For each $t \in Bad_1'$, let S_t be any set that witnesses t's membership in Bad_1 (that is, for which $\Pr[A|_t \in S_t]$ exceeds $e^{\hat{\epsilon}}\Pr[A'|_t \in S_t] + \hat{\delta}$). Consider the critical set

$$T_1 = \bigcup_{t \in Bad_1'} (S_t \times \{t\}).$$

Intuitively, this set will have large mass if Bad_1' does. Specifically, by the definition of S_t, we get a lower bound on the probability of T_1:

$$
\begin{aligned}
\Pr[(A, B) \in T_1] &= \sum_{t \in Bad_1'} \Pr[A|_t \in S_t]\Pr[B = t] \\
&> \sum_{t \in Bad_1'} (e^{\hat{\epsilon}}\Pr[A'|_t \in S_t] + \hat{\delta})\Pr[B = t] \\
&= \left(\sum_{t \in Bad_1'} e^{\hat{\epsilon}}\Pr[A'|_t \in S_t]\Pr[B = t] \right) + \hat{\delta}\Pr[B \in Bad_1'].
\end{aligned}
$$

Because Bad_1' does not contain points in Bad_0, we know that $\Pr[B = t] \geq e^{-2e}\Pr[B' = t]$. Substituting this into the bound above and using the fact that $\hat{\epsilon} = 3\epsilon$ and $\Pr[A'|_t \in S_t] = \Pr[A' \in S_t \mid B' = t]$, we get

$$
\begin{aligned}
\Pr[(A, B) \in T_1] &\geq \sum_{t \in Bad_1'} e^{\hat{\epsilon}}\Pr[A' \in S_t \mid B' = t]e^{-2\epsilon}\Pr[B' = t] + \hat{\delta}\Pr[B \in Bad_1'] \\
&= e^{\epsilon}\Pr[(A', B') \in T_1] + \hat{\delta}\Pr[B \in Bad_1'].
\end{aligned}
$$

By (ϵ, δ)-indistinguishability, $\Pr[(A, B) \in T_1] \leq e^{\epsilon}\Pr[(A', B') \in T_1] + \delta$. Combining the upper and lower bounds on the probability that $(A, B) \in T_1$, we have $\hat{\delta}\Pr[B \in Bad_1'] \leq \delta$, which implies that

$$\Pr[B \in Bad_1'] \leq \delta/\hat{\delta}.$$

By a similar argument, one gets that $\Pr[B \in Bad_2'] \leq \delta/\hat{\delta}$. Finally,

$$
\begin{aligned}
\Pr[B \in Bad_1 \cup Bad_2] &\leq \Pr[B \in Bad_0] + \Pr[B \in Bad_1'] + \Pr[B \in Bad_2'] \\
&= \frac{2\delta}{1 - e^{-\epsilon}} + \frac{\delta}{\hat{\delta}} + \frac{\delta}{\hat{\delta}} = \frac{2\delta}{1 - e^{-\epsilon}} + \frac{2\delta}{\hat{\delta}} = \delta''.
\end{aligned}
$$

By symmetry, we also have $\Pr[B' \in Bad_1 \cup Bad_2] \leq \frac{2\delta}{1-e^{-\epsilon}} + \frac{2\delta}{\hat{\delta}}$. Therefore, with probability at least $1 - \delta''$, $A|_t$ and $A'|_t$ are $(\hat{\epsilon}, \hat{\delta})$-indistinguishable, as claimed. $\qquad\square$

---

[3]*Note:* Even if we started with the stronger assumption that the pairs (A, B) and (A', B') are point-wise indistinguishable, we would still have to make a nontrivial argument to bound Bad_0, since point-wise indistinguishability is *not*, in general, closed under postprocessing.

3.4 The General Case: Proof of Theorem 2.4

Theorem 2.4 (restated).

(1) If $\epsilon, \delta > 0$ and $\delta < (1 - e^{-\epsilon})^2/n$, then (ϵ, δ)-differential privacy implies (ϵ', δ')-semantic privacy on databases of size n with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{n\delta}$ and $\delta' = 4\sqrt{n\delta}$.

(2) If $\epsilon, \delta > 0$ and $\epsilon \leq 0.45$, then (ϵ, δ)-semantic privacy implies $(3\epsilon, 2\delta)$-differential privacy.

Proof. (1) Let $\mathcal{A}$ be an (ϵ, δ)-differentially private algorithm. Let b be any belief distribution and let $x \sim b$. Let $\mathcal{A}_i(x) = \mathcal{A}(x_{-i})$, i.e., $\mathcal{A}_i$ on input x constructs x_{-i} and then applies $\mathcal{A}$ on it. From Lemma 3.3 (Part 3), we know that $(x, \mathcal{A}(x))$ and $(x, \mathcal{A}_i(x))$ are (ϵ, δ)-indistinguishable for every index $i = 1, \ldots, n$.

Apply Lemma 3.4 with $\mathcal{A}(X) = \mathcal{A}(x)$, $\mathcal{A}'(X) = \mathcal{A}_i(x)$, $\hat{\epsilon} = 3\epsilon$, and $\hat{\delta} = \sqrt{n\delta}$. We get that with probability at least $1 - \delta''$ over $t \sim \mathcal{A}(x)$, the random variables $x|_{\mathcal{A}(x)=t}$ and $x|_{\mathcal{A}_i(x)=t}$ are $(\hat{\epsilon}, \hat{\delta})$-indistinguishable, where $\delta'' \leq 2\delta/\hat{\delta} + 2\delta/(1 - e^{-\epsilon}) \leq 4\sqrt{\delta/n}$. Note that $1 - e^{-\epsilon} > \hat{\delta} = \sqrt{n\delta}$ (a condition assumed in the theorem).

Let $\delta' = n\delta''$; note that $\delta' \leq 4\sqrt{n\delta}$. Taking a union bound over all n choices for the index i, we get that with probability at least $1 - \delta'$ over the choice of $t \sim \mathcal{A}(x)$, all n variables $x|_{\mathcal{A}_i(x)=t}$ (for different i's) are $(\hat{\epsilon}, \hat{\delta})$-indistinguishable from $x|_{\mathcal{A}(x)=t}$.

To complete the proof of (1), recall that $(\hat{\epsilon}, \hat{\delta})$-indistinguishability implies statistical distance at most $e^{3\hat{\epsilon}} - 1 + \hat{\delta} = \epsilon'$.

(2) To see that (ϵ, δ)-semantic privacy implies $(3\epsilon, 2\delta)$-differential privacy, consider a belief distribution b which is uniform over two databases x, y which are at Hamming distance of one. The proof idea is same as in Theorem 2.2. Let i be the position in which x and y differ. Let $\mathcal{A}$ be an algorithm that satisfies (ϵ, δ)-semantic privacy.

In Game i, x and y induce the same distribution on transcripts, so the distribution $\bar{b}_i[\cdot|t]$ will be uniform over x and y (for all transcripts t). We now turn to Game 0 (the real world). Let E denote the set of transcripts t such that $\bar{b}_0[\cdot|t]$ assigns probabilities in $1/2 \pm \epsilon$ to each of the two databases x and y. Let $\bar{\mathcal{A}}$ denote the (random) output of $\mathcal{A}$ when run on a database sampled from distribution b. The semantic privacy of $\mathcal{A}$ implies E occurs with probability at least $1 - \delta$ over $t \sim \bar{\mathcal{A}}$. Working through Bayes' rule as in Theorem 2.2 shows that

$$e^{-3\epsilon} \Pr[\mathcal{A}(y) = t] \leq \Pr[\mathcal{A}(x) = t] \leq e^{3\epsilon} \Pr[\mathcal{A}(y) = t]$$

for all $t \in E$. (This last step uses the assumption that $\epsilon \leq 0.45$). Moreover, since $\bar{\mathcal{A}}$ is an equal mixture of $\mathcal{A}(x)$ and $\mathcal{A}(y)$, the event E must occur with probability at least $1 - 2\delta$ under both $t \sim \mathcal{A}(x)$ and $t \sim \mathcal{A}(y)$ Hence, $\mathcal{A}(x)$ and $\mathcal{A}(y)$ are $(3\epsilon, 2\delta)$-indistinguishable. Since this relationship holds for every pair of neighboring databases x and y, means that $\mathcal{A}$ is $(3\epsilon, 2\delta)$-differentially private. □

4 Further Discussion

Theorem 2.4 states that the relaxations of differential privacy in some previous work still provide meaningful guarantees in the face of arbitrary side information. This is *not* the case for all possible relaxations, even very natural ones, as noted in Remark 1.

Calibrating Noise to a High-Probability Bound Local Sensitivity. In a different vein, the techniques used to prove Theorem 2.4 can also be used to analyze schemes that do not provide privacy for *all* pairs of neighboring databases x and y, but rather only for *most* such pairs (remember that neighboring databases are the ones that differ in one entry). Specifically, it is sufficient that those databases where the indistinguishability condition fails occur only with small probability.

We first define a weakening of Definition 2.3 so that it only holds for specific belief distributions.

Definition 4.1 $((\epsilon, \delta)$-local semantic privacy**)**. *A randomized algorithm is (ϵ, δ)-local semantically private for a belief distribution b on $\mathcal{D}^n$ if with probability at least $1 - \delta$ over $t \sim \mathcal{A}(\mathrm{x})$ (t drawn from $\mathcal{A}(\mathrm{x})$), where the database x is drawn according to b, and for all $i = 1, \ldots, n$:*

$$\mathbf{SD} \left(\bar{b}_0[\cdot|t] \,,\, \bar{b}_i[\cdot|t] \right) \leq \epsilon.$$

Theorem 4.2. *Let $\mathcal{A}$ be a randomized algorithm. Let*

$$\mathcal{E} = \{\mathrm{x} : \forall \text{ neighbors } \mathrm{y} \text{ of } \mathrm{x}, \mathcal{A}(\mathrm{x}) \text{ and } \mathcal{A}(\mathrm{y}) \text{ are } (\epsilon, \delta)\text{-indistinguishable}\}.$$

Then $\mathcal{A}$ satisfies (ϵ', δ')-local semantic privacy for any belief distribution b such that $b[\mathcal{E}] = \Pr_{\mathrm{x} \sim b}[\mathrm{x} \in \mathcal{E}] \geq 1 - \delta_1$ with $\epsilon' = e^{3\epsilon} - 1 + \sqrt{n\delta_2}$ and $\delta' \leq 4\sqrt{n\delta_2}$ as long as $\epsilon > \sqrt{n\delta_2}$, where $\delta_2 = \delta + \delta_1$.

Proof. The proof is similar to Theorem 2.4 (1). Let b be a belief distribution with $b[\mathcal{E}] \geq 1 - \delta_1$ and let $\mathrm{x} \sim b$. From Lemma 3.3 (Part 4), we know that $(\mathrm{x}, \mathcal{A}(\mathrm{x}))$ and $(\mathrm{x}, \mathcal{A}_i(\mathrm{x}))$ are $(\epsilon, \delta + \delta_1)$-indistinguishable, where $\mathcal{A}_i(\mathrm{x}) = \mathcal{A}(\mathrm{x}_{-i})$. The remaining proof follows exactly as in Theorem 2.4 (1). $\qquad\square$

We now discuss a simple consequence of the above theorem to the technique of adding noise according to *local sensitivity* of a function.

Definition 4.3 (Local Sensitivity, [17])**.** *For a function $f : \mathcal{D}^n \to \mathbb{R}$, and $\mathrm{x} \in \mathcal{D}^n$, the local sensitivity of f at x is:*

$$LS_f(\mathrm{x}) = \max_{\mathrm{y} : d_H(\mathrm{x},\mathrm{y})=1} |f(\mathrm{x}) - f(\mathrm{y})|.$$

Let $Lap(\lambda)$ denote the Laplacian distribution. This distribution has density function $h(y) \propto \exp(-|y|/\lambda)$, mean 0, and standard deviation λ. Using the Laplacian noise addition procedure of [10, 17], along with Theorem 4.2 we get[4],

Corollary 4.4. *Let $\mathcal{E} = \{\mathrm{x} : LS_f(\mathrm{x}) \leq s\}$. Let $\mathcal{A}(\mathrm{x}) = f(\mathrm{x}) + Lap\left(\frac{s}{\epsilon}\right)$. Let b be a belief distribution such that $b[\mathcal{E}] = \Pr_{\mathrm{x} \sim b}[\mathrm{x} \in \mathcal{E}] \geq 1 - \delta_1$. Then $\mathcal{A}$ satisfies (ϵ', δ')-local semantic privacy for belief distribution b with $\epsilon' = e^{3\epsilon} - 1 + \sqrt{n\delta_1}$ and $\delta' \leq 4\sqrt{n\delta_1}$ as long as $\epsilon > \sqrt{n\delta_1}$.*

Proof. Let $\mathrm{x} \sim b$. If $\mathrm{x} \in \mathcal{E}$, then it follows from [10, 17], that $\mathcal{A}(\mathrm{x})$ and $\mathcal{A}(\mathrm{x}_{-i})$ are $(\epsilon, 0)$-indistinguishable for every index $i = 1, \ldots, n$. We can apply Theorem 4.2 to complete the proof. $\qquad\square$

The approach discussed here was generalized significantly by Bassily et al. [1]; we refer to their work for a detailed discussion.

---

[4]Similar corollaries could be derived for other differential privacy mechanisms like those that add Gaussian noise instead of Laplacian noise.

Acknowledgements

References

[1] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in private data analysis. In *Foundations of Computer Science (FOCS)*, 2013.

[2] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *PODS*, pages 128–138. ACM Press, 2005.

[3] Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In *CRYPTO*, pages 198–213. Springer, 2006.

[4] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM Press, 2003.

[5] Cynthia Dwork. Differential privacy. In *ICALP*, pages 1–12. Springer, 2006.

[6] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *J. Privacy and Confidentiality*, 2(1), 2010.

[7] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention, or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

[8] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *CRYPTO*, pages 528–544. Springer, 2004.

[9] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503. Springer, 2006.

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

[11] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273. ACM, 2008.

[12] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270–299, 1984.

[13] Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727, 2013.

[14] Daniel Kifer and Ashwin Machanavajjhala. No Free Lunch in Data Privacy. In *SIGMOD*, pages 193–204, 2011.

[15] Daniel Kifer and Ashwin Machanavajjhala. A rigorous and customizable framework for privacy. In *PODS*, pages 77–88, 2012.

[16] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: From theory to practice on the map. In *ICDE*, pages 277–286. IEEE Computer Society, 2008.

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

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

Appendix

A Proof of Lemma 3.3

Proof of Part 1. Let Bad be the set of *bad* values of a, that is

$$Bad = \{a \,:\, \Pr[X = a] < e^{-\epsilon} \Pr[Y = a] \text{ or } \Pr[X = a] > e^{\epsilon} \Pr[Y = a]\}.$$

By definition, $\Pr[X \in Bad] \leq \delta$. Now consider any set S of outcomes.

$$\Pr[X \in S] \leq \Pr[X \in S \setminus Bad] + \Pr[X \in Bad].$$

The first term is at most $e^{\epsilon} \Pr[Y \in S \setminus Bad] \leq e^{\epsilon} \Pr[Y \in S]$. Hence, $\Pr[X \in S] \leq e^{\epsilon} \Pr[Y \in S] + \delta$, as required. The case of $\Pr[Y \in S]$ is symmetric. Therefore, X and Y are (ϵ, δ)-indistinguishable.

Proof of Part 2. Let $S = \{a \,:\, \Pr[X = a] > e^{2\epsilon} \Pr[Y = a]\}$. Then

$$\Pr[X \in S] > e^{2\epsilon} \Pr[Y \in S].$$

By (ϵ, δ) indistringuishability, we have $\delta \geq \Pr[X \in S] - e^{\epsilon} \Pr[Y \in S] > (e^{2\epsilon} - e^{\epsilon}) \Pr[Y \in S]$. Equivalently,

$$\Pr[Y \in S] < \frac{\delta}{e^{2\epsilon} - e^{\epsilon}} = \frac{\delta}{e^{2\epsilon}(1 - e^{-\epsilon})} \,. \tag{7}$$

Now consider the set $S' = \{a \,:\, \Pr[X = a] < e^{-2\epsilon} \Pr[Y = a]\}$. A symmetric argument to the one above shows that $\Pr[X \in S'] < \delta/(e^{2\epsilon} - e^{\epsilon})$. Again using indistinguishability, we get

$$\Pr[Y \in S'] \leq e^{\epsilon} \Pr[X \in S'] + \delta < e^{\epsilon} \frac{\delta}{e^{2\epsilon} - e^{\epsilon}} + \delta = \frac{\delta}{1 - e^{-\epsilon}} \,. \tag{8}$$

The bound of (8) is always larger than that of (7), so we have $\Pr[Y \in S \cup S'] \leq \delta \cdot \frac{2e^{\epsilon}}{e^{\epsilon}-1}$. We can get the same bound on $\Pr[X \in S \cup S']$ by symmetry. Therefore, with probability at least $1 - \delta \cdot \frac{2e^{\epsilon}}{e^{\epsilon}-1}$ for a drawn from the distribution of either X or Y we have: $e^{-2\epsilon} \Pr[Y = a] \leq \Pr[X = a] \leq e^{2\epsilon} \Pr[Y = a]$.

Proof of Part 3. Let $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ be random variables on $D \times E$. Let S be an arbitrary subset of $D \times E$ and, for every $a \in D$, define $S_a = \{b \in E \,:\, (a, b) \in S\}$.

$$\begin{aligned}
\Pr[(X, \mathcal{A}(X)) \in S] &\leq \sum_{a \in D} \Pr[\mathcal{A}(X) \in S_a \,|\, X = a] \Pr[X = a] \\
&\leq \sum_{a \in D} (e^{\epsilon} \Pr[\mathcal{A}'(X) \in S_a \,|\, X = a] + \delta) \Pr[X = a] \\
&\leq \delta + e^{\epsilon} \Pr[(X, \mathcal{A}'(X)) \in S].
\end{aligned}$$

By symmetry, we also have $\Pr[(X, \mathcal{A}'(X)) \in S] < \delta + e^\epsilon \Pr[(X, \mathcal{A}(X)) \in S]$. Since the above inequalities hold for every selection of S, implies that $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ are (ϵ, δ)-indistinguishable.

Proof of Part 4. Let $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ be random variables on $D \times E$. Let $T \subset D$ be the set of a's for which $\mathcal{A}(a) \leq e^\epsilon \mathcal{A}'(a)$. Now, let S be an arbitrary subset of $D \times E$ and, for every $a \in D$, define $S_a = \{b \in E : (a, b) \in S\}$.

$$
\begin{aligned}
\Pr[(X, \mathcal{A}(X)) \in S] &= \sum_{a \notin T} \Pr[\mathcal{A}(X) \in S_a \mid X = a] \Pr[X = a] + \sum_{a \in T} \Pr[\mathcal{A}(X) \in S_a \mid X = a] \Pr[X = a] \\
&\leq \sum_{a \notin T} \Pr[X = a] + \sum_{a \in T} \Pr[\mathcal{A}(X) \in S_a \mid X = a] \Pr[X = a] \\
&= \Pr[X \notin T] + \sum_{a \in T} \Pr[\mathcal{A}(X) \in S_a \mid X = a] \Pr[X = a] \\
&\leq \delta_1 + \sum_{a \in T} (e^\epsilon \Pr[\mathcal{A}'(X) \in S_a \mid X = a] + \delta) \Pr[X = a] \\
&\leq \delta + \delta_1 + e^\epsilon \Pr[(X, \mathcal{A}'(X)) \in S].
\end{aligned}
$$

By symmetry, we also have $\Pr[(X, \mathcal{A}'(X)) \in S] < \delta + \delta_1 + e^\epsilon \Pr[(X, \mathcal{A}(X)) \in S]$. Since the above inequalities hold for every selection of S, implies that $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ are $(\epsilon, \delta + \delta_1)$-indistinguishable.

Proof of Part 5. Let X and Y be random variables on D. By definition $\mathbf{SD}(X, Y) = \max_{S \subset D} |\Pr[X \in S] - \Pr[Y \in S]|$. For any set $S \subset D$,

$$
\begin{aligned}
2|\Pr[X \in S] &- \Pr[Y \in S]| \\
&= |\Pr[X \in S] - \Pr[Y \in S]| + |\Pr[X \notin S] - \Pr[Y \notin S]| \\
&= \left| \sum_{c \in S} (\Pr[X = c] - \Pr[Y = c]) \right| + \left| \sum_{c \notin S} (\Pr[X = c] - \Pr[Y = c]) \right| \\
&\leq \sum_{c \in S} |\Pr[X = c] - \Pr[Y = c]| + \sum_{c \notin S} |\Pr[X = c] - \Pr[Y = c]| \\
&= \sum_{c \in D} |\Pr[X = c] - \Pr[Y = c]| \\
&\leq \sum_{c \in D} (e^\epsilon \Pr[Y = c] + \delta - \Pr[Y = c]) + \sum_{c \in D} (e^\epsilon \Pr[X = c] + \delta - \Pr[X = c]) \\
&= 2\delta + (e^\epsilon - 1) \sum_{c \in D} \Pr[Y = c] + (e^\epsilon - 1) \sum_{c \in D} \Pr[X = c] \\
&\leq 2(e^\epsilon - 1) + 2\delta = 2\bar{\epsilon} + 2\delta.
\end{aligned}
$$

This implies that $|\Pr[X \in S] - \Pr[Y \in S]| \leq \bar{\epsilon} + \delta$. Since the above inequality holds for every $S \subset D$, it immediately follows that the statistical difference between X and Y is at most $\bar{\epsilon} + \delta$. $\qquad\square$

B   Another View of Semantic Privacy

In this section, we discuss another possible definition of (ϵ, δ)-semantic privacy. Even though this definition seems to be the more desirable one, it also seems hard to achieve.

Definition A.1 (reality-oblivious (ϵ, δ)-semantic privacy). *A randomized algorithm is reality-oblivious (ϵ, δ)-semantically private if for all belief distributions b on $\mathcal{D}^n$, for all databases $\mathrm{x} \in \mathcal{D}^n$, with probability at least $1 - \delta$ over transcripts t drawn from $\mathcal{A}(\mathrm{x})$, and for all $i = 1, \ldots, n$:*

$$\mathbf{SD}\left(\bar{b}_0[\cdot | t] ,\ \bar{b}_i[\cdot | t]\right) \leq \epsilon.$$

We prove that if the adversary has arbitrary beliefs, then (ϵ, δ)-differential privacy doesn't provide any reasonable reality-oblivious (ϵ', δ')-semantic privacy guarantee.

Theorem A.2. *(ϵ, δ)-differential privacy does not imply reality-oblivious (ϵ', δ')-semantic privacy for any reasonable values of ϵ' and δ'.*

Proof. This counterexample is due to Dwork and McSherry: suppose that the belief distribution is uniform over $\{(0^n), (1, 0^{n-1})\}$, but that real database is (1^n). Let the database $\mathrm{x} = (x_1, \ldots, x_n)$. Say we want to reveal $f(\mathrm{x}) = \sum_i x_i$. Adding Gaussian noise with variance $\sigma^2 = \log\left(\frac{1}{\delta}\right) / \epsilon^2$ satisfies (ϵ, δ)-differential privacy (refer [10, 17] for details). However, with overwhelming probability the output will be close to n, and this will in turn induce a very non-uniform distribution over $\{(0^n), (1, 0^{n-1})\}$ since $(1, 0^{n-1})$ is exponentially (in n) more likely to generate a value near n than (0^n). More precisely, due to the Gaussian noise added,

$$\frac{\Pr[\mathcal{A}(\mathrm{x}) = n \,|\, \mathrm{x} = (0^n)]}{\Pr[\mathcal{A}(\mathrm{x}) = n \,|\, \mathrm{x} = (1, 0^{n-1})]} = \frac{\exp\left(\frac{-n^2}{2\sigma}\right)}{\exp\left(\frac{-(n-1)^2}{2\sigma}\right)} = \exp\left(\frac{-2n+1}{2\sigma}\right).$$

Therefore, given that the output is close to n, the posterior distribution of the adversary would be exponentially more biased toward $(1, 0^{n-1})$ than (0^n). Hence, it is exponentially far away from the prior distribution which was uniform. On the other hand, on x_{-1}, no update will occur and the posterior distribution will remain uniform over $\{(0^n), (1, 0^{n-1})\}$ (same as the prior). Since the posterior distributions in these two situations are exponentially far apart (one exponentially far from uniform, other uniform), it shows that (ϵ, δ)-differential privacy does not imply any reasonable guarantee on reality-oblivious semantic privacy. $\square$

The counterexample of Theorem B implies that adversaries whose belief distribution is very different from the real database may observe a large change in their posterior distributions. We do not consider this a 'violation of "privacy", since the issue lies in the incorrect beliefs, not the mechanism per se.