A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information

Shiva Prasad Kasiviswanathan Adam Smith
Department of Computer Science and Engineering
Pennsylvania State University
e-mail: {kasivisw,asmith}@cse.psu.edu

Abstract

In this note we give a precise formulation of "resistance to arbitrary side information" and show that several relaxations of differential privacy imply it. The formulation follows the ideas originally due to Dwork and McSherry, stated implicitly in [4]. This is, to our knowledge, the first place such a formulation appears explicitly. The proof that relaxed definitions (and hence the schemes of [5, 10, 9]) satisfy the Bayesian formulation is new.

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 which 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 [6]) are believed to compose securely even in the presence of arbitrary side information. In this note we give a precise formulation of this statement. First, we provide a Bayesian formulation of differential privacy which makes its resistance to arbitrary side information explicit. Second, we prove that the relaxed definitions of [5, 9] still 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 the recent techniques mentioned above [5, 2, 10, 9] can be used modularly with the same sort of assurances as in the case of strictly differentially-private algorithms.

1.1 Differential Privacy

Databases are assumed to be vectors in $\mathcal{D}^n$ for some domain $\mathcal{D}$. The Hamming distance $d(\mathrm{x}, \mathrm{y})$ 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 $\mathbb{P}$ and $\mathbb{Q}$ are probability measure on a discrete space D, the *statistical difference* (a.k.a. *total variation distance*) between $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$\mathbf{SD}\,(\mathbb{P},\mathbb{Q}) = \max_{S\subset D} |\,\mathbb{P}\,[S] - \mathbb{Q}[S)|.$$

Definition 1.1 (ϵ-differential privacy [6])**.** *A randomized algorithm $\mathcal{A}$ is said to be ϵ-differentialy 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*

$$\Pr[\mathcal{A}(\mathrm{x})\in S] \le e^{\epsilon}\Pr[\mathcal{A}(\mathrm{y})\in S]. \tag{1}$$

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 total variation (also called 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 [3, 7, 1, 5, 2, 10, 9]. The schemes in all those papers, however, satisfy the following relaxation [5]:

Definition 1.2 ((ϵ,δ)-differential privacy)**.** *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,* $\Pr[\mathcal{A}(\mathrm{x})\in S]\le e^{\epsilon}\Pr[\mathcal{A}(\mathrm{y})\in S]+\delta$.

The relaxations used in [7, 1, 9] were in fact stronger (i.e., less relaxed) than Definition 1.1. One consequence of the results below is that all the definitions are equivalent up to polynomial changes in the parameters, and so given the space constraints we work only with the simplest notion.[1]

2 Semantics of Differential Privacy

There is a crisp, semantically-flavored interpretation of differential privacy, due to Dwork and McSherry, and explained in [4]: *Regardless of external knowledge, an adversary with access to the sanitized database draws the same conclusions whether or not my data is included in the original data.* (the use of the term "semantic" for such definitions dates back to semantic security of encryption [8]). In this section, we develop a formalization of this interpretation and show that the definition of differential privacy used in the line of work this paper follows ([3, 7, 1, 6]) is essential in order to satisfy the intuition.

We require a mathematical formulation of "arbitrary 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 about the database x using Bayes' rule to obtain a posterior $\bar{b}$:

$$\bar{b}[\mathrm{x}|t] = \frac{\Pr[\mathcal{A}(\mathrm{x})=t]b[\mathrm{x}]}{\sum_{\mathrm{y}}\Pr[\mathcal{A}(\mathrm{y})=t]b[\mathrm{y}]}\,. \tag{2}$$

Note that in an interactive scheme, 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.

[1] That said, some of the other relaxations, such as probabilistic differential privacy from [9], might lead to better parameters in Theorem 2.4.

For a database x, define x_{-i} to be the same vector where position i has been replaced by some fixed, default value in D. Any valid value in D will do for the default value. We can then imagine $n+1$ 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 $\mathcal{A}$. 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 included.

For a particular belief distribution b and transcript t, we can then define $n+1$ *a posteriori* distributions $\bar{b}_0, \ldots, \bar{b}_n$, where the $\bar{b}_0$ is the same as $\bar{b}$ (defined in 2) and, for larger i, the i-th belief distribution is defined with respect to Game i:

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

Given a particular transcript t, the 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. It turns out that the exact measure of "different" here does not matter much. We chose the weakest notion that applies, namely statistical difference. We say there is a problem for transcript t if 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 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 databases $\mathrm{x} \in \mathcal{D}^n$, for all possible transcripts t, and for all $i = 1, \ldots, n$:*

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

Dwork and McSherry proposed the notion of semantic privacy, informally, and observed that it is equivalent to differential privacy. We now formally show that the notions of ϵ-differential privacy (Definition 1.1) and ϵ-semantic privacy (Definition 2.1) are very closely related.

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

We extend the previous Bayesian formulation to capture situations where bad events can occur with some negligible probability (say, δ). We relax ϵ-semantic privacy to (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 pairs (x, t), where the database x is drawn according to b, and transcript t is drawn according to $\mathcal{A}(\mathrm{x})$, and for all $i = 1, \ldots, n$:*

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

This definition is only interesting when $\epsilon > \delta$; otherwise just use statistical difference 2δ and leave $\epsilon = 0$. Below, we assume $\epsilon > \delta$. In fact, in many of the proofs we will be assuming that δ is a negligible function (of $O(1/n^2)$). In Appendix A, we provide another related definition of (ϵ, δ)-semantic privacy.

Theorem 2.4 (Main Theorem)**.** *(ϵ, δ)-differential privacy implies (ϵ', δ')-semantic privacy for arbitrary (not necessarily informed) beliefs with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$. $(\bar{\epsilon}/2, \delta)$-semantic privacy implies $(2\epsilon, 2\delta)$-differential privacy with $\bar{\epsilon} = e^{\epsilon} - 1$.*

3 Some Properties of (ϵ, δ)-Differential Privacy

We now describe some properties of (ϵ, δ)-differential privacy that would be useful later on. This section could be of independent interest. Instead of restricting ourselves to outputs of randomized algorithms, we consider a more general definition of (ϵ, δ)-differential privacy.

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 $S \subseteq D$,*

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

We will also be using a simpler variant of (ϵ, δ)-indistinguishability, which we call *point-wise* (ϵ, δ)-indistinguishability. Claim 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:*

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

Claim 3.3. *The following are useful facts about indistinguishability.*[2]

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 $(2\epsilon, \frac{2\delta}{e^{\epsilon}\epsilon})$-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 $1 - \delta$ over $a \leftarrow X$ (a drawn from 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 $(\epsilon, 2\delta)$-indistinguishable.*
5. *If X, Y are (ϵ, δ)-indistinguishable and $\mathcal{G}$ is some randomized algorithm, then $\mathcal{G}(X)$ and $\mathcal{G}(Y)$ are (ϵ, δ)-indistinguishable.*
6. *If X, Y are (ϵ, δ)-indistinguishable, then $\mathbf{SD}(X, Y) \leq \bar{\epsilon} + \delta$, where $\bar{\epsilon} = e^{\epsilon} - 1$.*

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.

[2] A few similar properties relating to statistical difference were shown in [11]. Note that (ϵ, δ)-indistinguishability is not a metric, unlike statistical difference. But it does inherit some nice metric like properties.

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

$$\Pr[X \in S] > e^{2\epsilon}\Pr[Y \in S] > e^{\epsilon}(1+\epsilon)\Pr[Y \in S] \Rightarrow \Pr[X \in S] - e^{\epsilon}\Pr[Y \in S] > \epsilon e^{\epsilon}\Pr[Y \in S].$$

Since, $\Pr[X \in S] - e^{\epsilon}\Pr[Y \in S] \leq \delta$, we mush have $\epsilon e^{\epsilon}\Pr[Y \in S] < \delta$. A similar argument when considering the set $S' = \{a : \Pr[X = a] < e^{-2\epsilon}\Pr[Y = a]\}$ shows that $\epsilon e^{\epsilon}\Pr[Y \in S'] < \delta$. Putting both arguments together, $\Pr[Y \in S \cup S'] \leq 2\delta/(\epsilon e^{\epsilon})$. Therefore, with probability at least $1 - 2\delta/(e^{\epsilon}\epsilon)$ for any a drawn from 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 $S_a = \{b \in E : (a, b) \in S\}$.

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

By symmetry, we also have $\Pr[(X, \mathcal{A}'(X)) \in S] < \delta + \Pr[(X, \mathcal{A}(X)) \in S]$. Since S was arbitrary, $(X, \mathcal{A}(X))$ 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 $S_a = \{b \in E : (a, b) \in S\}$.

$$\begin{aligned}
\Pr[(X, \mathcal{A}(X)) \in S] &\leq \Pr[X \notin T] + \sum_{a \in T} \Pr[\mathcal{A}(X) \in S_a : X = a]\Pr[X = a] \\
&< \delta + \sum_{a \in T} (e^{\epsilon}\Pr[\mathcal{A}'(X) \in S_a : X = a] + \delta)\Pr[X = a] \\
&< 2\delta + e^{\epsilon}\Pr[(X, \mathcal{A}'(X)) \in S].
\end{aligned}$$

By symmetry, we also have $\Pr[(X, \mathcal{A}'(X)) \in S] < 2\delta + \Pr[(X, \mathcal{A}(X)) \in S]$. Since S was arbitrary, $(X, \mathcal{A}(X))$ and $(X, \mathcal{A}'(X))$ are $(\epsilon, 2\delta)$-indistinguishable.

Proof of Part 5. Let D be some domain. A randomized procedure $\mathcal{G}$ is a pair $\mathcal{G} = (g, R)$, where R is a random variable on some set E and g is a function from $D \times E$ to any set F. If X is a random variable on D, then $\mathcal{G}(X)$ denotes the random variable on F obtained by sampling $X \otimes R$ and applying g to the result, where the symbol $\otimes$ denotes the tensor product. Now for any set $S \subset F$,

$$\begin{aligned}
&\Pr[\mathcal{G}(X) \in S] - e^{\epsilon}\Pr[\mathcal{G}(Y) \in S] \\
&= \Pr[g(X \otimes R) \in S] - e^{\epsilon}\Pr[g(Y \otimes R) \in S] \\
&= \Pr[X \otimes R \in g^{-1}(S)] - e^{\epsilon}\Pr[Y \otimes R \in g^{-1}(S)] \\
&\leq \sum_{r \in E} \Pr[X \in S_r : R = r]\Pr[R = r] - e^{\epsilon}\sum_{r \in E} \Pr[Y \in S_r : R = r]\Pr[R = r] \\
&= \sum_{r \in E} (\Pr[X \in S_r : R = r] - e^{\epsilon}\Pr[Y \in S_r : R = r])\Pr[R = r] \\
&\leq \sum_{r \in E} \delta \Pr[R = r] = \delta.
\end{aligned}$$

By symmetry, we also have $\Pr[\mathcal{G}(Y) \in S] - e^{\epsilon} \Pr[\mathcal{G}(X) \in S] \leq \delta$. Since S was arbitrary, $\mathcal{G}(X)$ and $\mathcal{G}(Y)$ are (ϵ, δ)-indistinguishable.

Proof of Part 6. Let X and Y be random variables on D. By definition $\mathbf{SD}\left(X, Y\right) = \max_{S \subset D} |\Pr[X \in S] - \Pr[Y \in S]|$. For any set $S \subset D$,

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

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 $\bar{\epsilon} + \delta$. ☐

4 Proofs of Theorems 2.2 and 2.4

This section is devoted to proving Theorems 2.2 and 2.4. For convenience we restate the theorem statements.

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

Proof. Consider any database x. Consider belief distributions $\bar{b}_0[\mathrm{x}|t]$ and $\bar{b}_i[\mathrm{x}|t]$. differential privacy implies that the ratio of $\bar{b}_0[\mathrm{x}|t]$ and $\bar{b}_i[\mathrm{x}|t]$ is within $e^{\pm\epsilon}$ on every point, i.e., for every i and for every possible transcript t:

$$e^{-\epsilon}\bar{b}_i[\mathrm{x}|t] \leq \bar{b}_0[\mathrm{x}|t] \leq e^{\epsilon}\bar{b}_i[\mathrm{x}|t].$$

In the remainder of the proof we fix i and t. Substituting $\delta = 0$ in Claim 3.3 (part 6), implies that $\mathbf{SD}\left(\bar{b}_0[\mathrm{x}|t], \bar{b}_i[\mathrm{x}|t]\right) = \bar{\epsilon}$.

To see that $\bar{\epsilon}$-semantic privacy implies 2ϵ-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. 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 $1/2 \pm \bar{\epsilon}/2$ to each of the two databases (follows from ϵ-semantic privacy definition). Working through Bayes' rule shows that

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

This implies that $\mathcal{A}$ is point-wise 2ϵ-differentialy private. Using Claim 3.3 (part 1), implies that $\mathcal{A}$ is 2ϵ-differentialy private. $\square$

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 4.1 (Main Lemma)**.** *Suppose two pairs of random variables $(X,\mathcal{A}(X))$ and $(Y,\mathcal{A}'(Y))$ are (ϵ,δ)-differentialy private (for some randomized algorithms $\mathcal{A}$ and $\mathcal{A}'$). Then with probability at least $1-\delta''$ over $t\leftarrow\mathcal{A}(X)$ (equivalently $t\leftarrow\mathcal{A}'(Y)$), the random variables $X|_{\mathcal{A}(X)=t}$ and $Y|_{\mathcal{A}'(Y)=t}$ are $(\hat{\epsilon},\hat{\delta})$-differentialy private with $\hat{\epsilon}=3\epsilon$, $\hat{\delta}=2\sqrt{\delta}$, and $\delta''=\sqrt{\delta}+\frac{2\delta}{\epsilon e^{\epsilon}}=O(\sqrt{\delta})$.*

Proof. Let $(X,\mathcal{A}(X))$ and $(Y,\mathcal{A}'(Y))$ be random variables on $D\times E$. The first observation is that $\mathcal{A}(X)$ and $\mathcal{A}(Y)$ are (ϵ,δ)-differentialy private. To prove that consider any set $P\in E$,

$$\begin{aligned}\Pr[\mathcal{A}(X)\in P] &= \Pr[(X,\mathcal{A}(X))\in D\times P]\le e^{\epsilon}\Pr[(Y,\mathcal{A}'(Y))\in D\times P]+\delta\\ &= e^{\epsilon}\Pr[\mathcal{A}'(Y)\in P]+\delta.\end{aligned}$$

Since P was arbitrary, $\mathcal{A}(X)$ and $\mathcal{A}'(Y)$ are (ϵ,δ)-differentialy private. In the remainder of the proof, we will use the notation $X|_t$ for $X|_{\mathcal{A}(X)=t}$ and $Y|_t$ for $Y|_{\mathcal{A}'(Y)=t}$. Define,

$$\begin{gathered}Bad_0=\{a\,:\,e^{-2\epsilon}\Pr[\mathcal{A}'[Y]=a]>\Pr[\mathcal{A}(X)=a]>e^{2\epsilon}\Pr[\mathcal{A}'[Y]=a]\}\\ Bad_1=\{a\,:\,\exists S\subset D\text{ such that }\Pr[X|_a\in S]>e^{\hat{\epsilon}}\Pr[Y|_a\in S]+\hat{\delta}\}\\ Bad_2=\{a\,:\,\exists S\subset D\text{ such that }\Pr[Y|_a\in S]>e^{\hat{\epsilon}}\Pr[X|_a\in S]+\hat{\delta}\}.\end{gathered}$$

We need an upper bound for the probabilities $\Pr[\mathcal{A}(X)\in Bad_1\cup Bad_2]$ and $\Pr[\mathcal{A}'(Y)\in Bad_1\cup Bad_2]$. We know from Claim 3.3 (part 2), that

$$\Pr[\mathcal{A}(X)\in Bad_0]\le\frac{2\delta}{\epsilon e^{\epsilon}}\quad\text{and}\quad\Pr[\mathcal{A}'(Y)\in Bad_0]\le\frac{2\delta}{\epsilon e^{\epsilon}}.$$

Note that from the initial observation $\mathcal{A}(X)$ and $\mathcal{A}'(Y)$ are (ϵ,δ)-differentialy private, therefore the condition required for applying Claim 3.3 (part 2) holds. Now define,

$$Bad_1'=Bad_1\setminus Bad_0\quad\text{and}\quad Bad_2'=Bad_2\setminus Bad_0.$$

For each $a\in Bad_1'$ and $T\subset D\times E$, define $S_a=\{b\in D\,:\,(b,a)\in T\}$. Define $T_1=S_a\times\bigcup_{a\in Bad_1'}\{a\}$.

$$\begin{aligned}\Pr[(X,\mathcal{A}(X))\in T_1] &= \sum_{a\in Bad_1'}\Pr[X\in S_a\,:\,\mathcal{A}(X)=a]\Pr[\mathcal{A}(X)=a]\\ &> \sum_{a\in Bad_1'}(e^{\hat{\epsilon}}\Pr[Y\in S_a\,:\,\mathcal{A}'(Y)=a]+\hat{\delta})\Pr[\mathcal{A}(X)=a]\\ &= \sum_{a\in Bad_1'}e^{\hat{\epsilon}}\Pr[Y\in S_a\,:\,\mathcal{A}'(Y)=a]\Pr[\mathcal{A}(X)=a]+\hat{\delta}\sum_{a\in Bad_1'}\Pr[\mathcal{A}(X)=a]\\ &= \sum_{a\in Bad_1'}e^{3\epsilon}\Pr[Y\in S_a\,:\,\mathcal{A}'(Y)=a]e^{-2\epsilon}\Pr[\mathcal{A}'(Y)=a]+\hat{\delta}\Pr[\mathcal{A}(X)\in Bad_1']\\ &= e^{\epsilon}\Pr[(Y,\mathcal{A}'(Y))\in T_1]+\hat{\delta}\Pr[\mathcal{A}(X)\in Bad_1'].\end{aligned}$$

The inequality follows because of the definition of Bad_1'. By (ϵ,δ)-differential privacy, $\Pr[(X,\mathcal{A}(X)) \in T_1] \le e^{\epsilon}\Pr[(Y,\mathcal{A}(X)) \in T_1] + \delta$. Therefore,

$$\hat{\delta}\Pr[\mathcal{A}(X) \in Bad_1'] \le \delta \Rightarrow \Pr[\mathcal{A}(X) \in Bad_1'] \le \delta/\hat{\delta}.$$

Similarly, $\Pr[\mathcal{A}(X) \in Bad_2'] \le \delta/\hat{\delta}$. Finally,

$$\begin{aligned}\Pr[\mathcal{A}(X) \in Bad_1 \cup Bad_2] &\le \Pr[\mathcal{A}(X) \in Bad_0] + \Pr[\mathcal{A}(X) \in Bad_1'] + \Pr[\mathcal{A}(X) \in Bad_2'] \\ &= \frac{2\delta}{\epsilon e^{\epsilon}} + \frac{\delta}{\hat{\delta}} + \frac{\delta}{\hat{\delta}} = \frac{2\delta}{\epsilon e^{\epsilon}} + \sqrt{\delta}.\end{aligned}$$

By symmetry, we also have $\Pr[\mathcal{A}'(Y) \in Bad_1 \cup Bad_2] \le \frac{2\delta}{\epsilon e^{\epsilon}} + \sqrt{\delta}$. Therefore, with probability at least $1-\delta''$, $X|_t$ and $Y|_t$ are $(\hat{\epsilon},\hat{\delta})$-differentialy private. □

The following corollary follows by using the above proposition (with $Y = X$) in conjunction with Claim 3.3 (part 6).

Corollary 4.2. *Let $(X,\mathcal{A}(X))$ and $(X,\mathcal{A}'(X))$ be (ϵ,δ)-differentialy private. Then, with probability at least $1-\delta''$ over $t \leftarrow \mathcal{A}(X)$ (equivalently $t \leftarrow \mathcal{A}'(X)$), the statistical difference between $X|_{\mathcal{A}(X)=t}$ and $X|_{\mathcal{A}'(X)=t}$ is at most $e^{\hat{\epsilon}} - 1 + \hat{\delta}$ with $\hat{\epsilon} = 3\epsilon$, $\hat{\delta} = 2\sqrt{\delta}$, and $\delta'' = O(\sqrt{\delta})$.*

Theorem 2.4. *(ϵ,δ)-differential privacy implies (ϵ',δ')-semantic privacy for arbitrary (not necessarily informed) beliefs with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$. $(\bar{\epsilon}/2,\delta)$-semantic privacy implies $(2\epsilon, 2\delta)$-differential privacy with $\bar{\epsilon} = e^{\epsilon} - 1$.*

Proof. Let $\mathcal{A}$ be a (ϵ,δ)-differentialy private algorithm. Let b be any belief distribution. From Claim 3.3 (part 3), we know that $(b,\mathcal{A}(b))$ and $(b,\mathcal{A}_i(b))$ are (ϵ,δ)-differentialy private. Let $\delta'' = O(\sqrt{\delta})$. From Corollary 4.2, we get that with probability at least $1-\delta''$ over $t \leftarrow \mathcal{A}(b)$, the statistical difference between $b|_{\mathcal{A}(b)=t}$ and $b|_{\mathcal{A}_i(b)=t}$ is at most ϵ'. Therefore, for any $\mathrm{x} \leftarrow b$, with probability at least $(1-\delta'')$ over $t \leftarrow \mathcal{A}(\mathrm{x})$, $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \le \epsilon'$. Taking union bound over all coordinates i, implies that for any $\mathrm{x} \leftarrow b$ with probability at least $1-n\delta''$ over $t \leftarrow \mathcal{A}(b)$, for all $i = 1,\ldots,n$, we have $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \le \epsilon'$. Therefore, $\mathcal{A}$ satisfies (ϵ',δ')-semantic privacy for b. Since b was arbitrary, we get that (ϵ,δ)-differential privacy implies (ϵ',δ')-semantic privacy.

To see that $(\bar{\epsilon}/2,\delta)$-semantic privacy implies $(2\epsilon,2\delta)$-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 $\bar{\mathcal{A}}$ be an algorithm that with probability $1/2$ draws an output from $\mathcal{A}(\mathrm{x})$ and with probability $1/2$ draws an output from $\mathcal{A}(\mathrm{y})$. Consider a transcript t drawn from $\bar{\mathcal{A}}$. 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 with probability at least $1-\delta$ over $t \leftarrow \bar{\mathcal{A}}$, $\bar{b}_0[\cdot|t]$ will assign probabilities $1/2 \pm \bar{\epsilon}/2$ to each of the two databases. Working through Bayes' rule as in Theorem 2.2 shows that $\bar{\mathcal{A}}$ is point-wise $(2\epsilon,\delta)$-differentialy private (with probability at least at least $1-2\delta$ of $t \leftarrow \mathcal{A}(\mathrm{x})$, $e^{-2\epsilon}\Pr[\mathcal{A}(\mathrm{y}) = t] \le \Pr[\mathcal{A}(\mathrm{x}) = t] \le e^{2\epsilon}\Pr[\mathcal{A}(\mathrm{y}) = t]$). Therefore, with probability at least $1-\delta$ of $t \leftarrow \bar{\mathcal{A}}$, $e^{-2\epsilon}\Pr[\mathcal{A}(\mathrm{y}) = t] \le \Pr[\mathcal{A}(\mathrm{x}) = t] \le e^{2\epsilon}\Pr[\mathcal{A}(\mathrm{y}) = t]$. Similarly, for $t \leftarrow \mathcal{A}(\mathrm{y})$. This implies that $\mathcal{A}$ is point-wise $(2\epsilon,2\delta)$-differentialy private. Using Claim 3.3 (part 1), implies that $\mathcal{A}$ is $(2\epsilon,2\delta)$-differentialy private. □

5 Discussion and Consequences

Theorem 2.4 states that the relaxations notions of differential privacy used in some previous work still imply privacy in the face of arbitrary side information. This is *not* the case for *all* possible relaxations, even very natural ones. For example, if one replaced the multiplicative notion of distance used in differential privacy with total variation distance, then the following "sanitizer" would be deemed private: choose an index $i \in \{1, \ldots, n\}$ uniformly at random and publish the entire record of individual i together with his or her identity (example 2 in [6]). Such a "sanitizer" would not be meaningful at all, regardless of side information.

Theorems 2.4 and A.3 give some qualitative improvements over existing security statements. Theorem A.3 implies that the claims of [3, 7, 1] can be strengthened to hold for *all* predicates of the input simultaneously (a switch in the order of quantifiers). The strengthening does come at some loss in parameters since δ is increased. This incurs a factor of 2 in $\log\left(\frac{1}{\delta}\right)$, or a factor of $\sqrt{2}$ in the standard deviation. More significantly, Theorem 2.4 shows that noise processes with negligible probability of bad events have nice differential privacy guarantees even for adversaries who are not necessarily informed. There is a hitch however only adversaries whose beliefs somehow represent reality, i.e. for whom the real database is somehow "representative" of the adversary's view can be said to learn nothing.

Finally, the techniques used to prove Theorem 2.4 can also be used to analyze schemes which 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 "differential privacy" condition fails occur only with small probability.

Theorem 5.1. *Let $\mathcal{A}$ be a randomized algorithm. Let*

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

Then $\mathcal{A}$ satisfies (ϵ', δ')-semantic privacy for any belief distribution b such that $b[\mathcal{E}] = \Pr_{\mathrm{x}\leftarrow b}[\mathrm{x} \in \mathcal{E}] \geq 1-\delta$ with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$.

Proof. Let b be a belief distribution with $b[\mathcal{E}] \geq 1 - \delta$. Let $\delta'' = O(\sqrt{\delta})$. From Claim 3.3 (part 4), we know that $(b, \mathcal{A}(b))$ and $(b, \mathcal{A}_i(b))$ are $(\epsilon, 2\delta)$-differentialy private. From Corollary 4.2, we get that with probability at least $1 - \delta''$ over $t \leftarrow \mathcal{A}(b)$, the statistical difference between $b|_{\mathcal{A}(b)=t}$ and $b|_{\mathcal{A}_i(b)=t}$ is at most ϵ'. Therefore, with probability at least $(1 - \delta'')$ over pairs (x, t) where $\mathrm{x} \leftarrow b$ and $t \leftarrow \mathcal{A}(\mathrm{x})$, $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \leq \epsilon'$. Taking union bound over all coordinates i, implies that with probability at least $1 - n\delta''$ over pairs (x, t) where $\mathrm{x} \leftarrow b$ and $t \leftarrow \mathcal{A}(\mathrm{x})$, for all $i = 1, \ldots, n$, we have $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \leq \epsilon'$. Therefore, $\mathcal{A}$ satisfies (ϵ', δ')-semantic privacy for belief distribution b. $\square$

Let $LS_f(\cdot)$ denote the local sensitivity of function f (defined in [10]). 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 [6, 10], along with Theorem 5.1 we get,

Corollary 5.2. *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 $b[\mathcal{E}] = \Pr_{\mathrm{x}\leftarrow b}[\mathrm{x} \in \mathcal{E}] \geq 1 - \delta$. Then $\mathcal{A}$ satisfies (ϵ', δ')-semantic privacy for the belief distribution b with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$.*

Acknowledgements

We are grateful for helpful discussions with Cynthia Dwork, Frank McSherry, Moni Naor, Kobbi Nissim, and Sofya Raskhodnikova.

References

[1] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *PODS*, pages 128–138. ACM Press, 2005.

[2] K. Chaudhuri and N. Mishra. When random sampling preserves privacy. In *CRYPTO*, pages 198–213, 2006.

[3] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM Press, 2003.

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

[5] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503. Springer, 2006.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.

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

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

[9] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: From theory to practice on the map. In *ICDE*, 2008.

[10] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM Press, 2007.

[11] A. Sahai and S. Vadhan. Manipulating statistical difference. In P. Pardalos, S. Rajasekaran, and J. Rolim, editors, *Randomization Methods in Algorithm Design (DIMACS Workshop, December 1997)*, volume 43 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 251–270. American Mathematical Society, 1999.

Appendix A: 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 $i = 1, \ldots, n$:*

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

We first prove 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. [3] *(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 [6, 10] 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, if the adversary believes he is seeing $\mathcal{A}(\mathrm{x}_{-1})$, then no update will occur and the posterior distribution will remain uniform. 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. □

However, (ϵ, δ)-differential privacy does provide a strong reality-oblivious (ϵ', δ')-semantic privacy guarantee for *informed* belief distributions. Using terminology from [1, 6], we say that a belief distribution b is informed if b is constant on $n-1$ coordinates and agrees with the database in those coordinates. This corresponds to the adversary knowing some set of $n-1$ entries in the database before interacting with the algorithm, and then trying to learn the remaining one entry from the interaction. Let $\mathcal{A}_i$ be a randomized algorithm such that for all databases x, $\mathcal{A}_i(\mathrm{x}) = \mathcal{A}(\mathrm{x}_{-i})$.

Theorem A.3. *(ϵ, δ)-differential privacy implies reality-oblivious (ϵ', δ')-semantic privacy for informed beliefs with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$.*[4]

Proof. Let $\mathcal{A}$ be a (ϵ, δ)-differentialy private algorithm. Let x be any database. Let b be any informed belief distribution. This means that b is constant on all $n-1$ coordinates, and agrees with x in those $n-1$ coordinates. Let i be the coordinate which is not yet fixed in b. From Claim 3.3 (part 3), we know that $(b, \mathcal{A}(b))$ and $(b, \mathcal{A}_i(b))$ are (ϵ, δ)-differentialy private. Therefore, we can apply Lemma 4.1. Let $\delta'' = O(\sqrt{\delta})$. From Corollary 4.2, we get that with probability at least $1 - \delta''$ over $t \leftarrow \mathcal{A}(b)$, the statistical difference between $b|_{\mathcal{A}(b)=t}$ and $b|_{\mathcal{A}_i(b)=t}$ is at most ϵ'. Therefore, for x, with probability at least $(1 - \delta'')$ over $t \leftarrow \mathcal{A}(\mathrm{x})$, $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \leq \epsilon'$. Taking union bound over all coordinates i, implies that with probability at least $1 - n\delta''$ over $t \leftarrow \mathcal{A}(\mathrm{x})$, for all $i = 1, \ldots, n$, we have $\mathbf{SD}\left(b|_{\mathcal{A}(\mathrm{x})=t}, b|_{\mathcal{A}_i(\mathrm{x})=t}\right) \leq \epsilon'$. Therefore, $\mathcal{A}$ satisfies reality-oblivious (ϵ', δ')-semantic privacy for b. Since x was arbitrary, we get that (ϵ, δ)-differential privacy implies reality-oblivious (ϵ', δ')-semantic privacy for informed beliefs. □

[3]Note that adversaries whose belief distribution is very different from the real database (as in the counterexample of Theorem A.2 may think they have learned a lot. But does such "learning" represent a breach of privacy? We do not think so, but leave the final decision to the reader.

[4]Reality-oblivious $(\bar{\epsilon}/2, \delta)$-semantic privacy implies $(2\epsilon, 2\delta)$-differential privacy with $\bar{\epsilon} = e^{\epsilon} - 1$. For details see the proof of Theorem 2.4.