# A Practical Application of Differential Privacy to Personalized Online Advertising[*]

Yehuda Lindell Eran Omri

Department of Computer Science
Bar-Ilan University, Israel.
lindell@cs.biu.ac.il,omrier@gmail.com

## Abstract

Online advertising plays an important role in supporting many Internet services. Personalized online advertising offers marketers a way to direct ads at very specific audiences. The vast body of Internet users combined with the ease of creating and monitoring personalized advertising campaigns make online advertising an extremely strong tool for marketers. However, many concerns arise regarding the implications of online advertising for the privacy of web users. Specifically, recent works show how the privacy of Internet users may be breached by attacks utilizing personalized advertising campaigns such as those provided by Facebook. Such attacks succeed even without the user ever noticing the attack or being able to avoid it (unless refraining from going on the Internet).

In this work, we suggest practical and concrete measures for preventing the feasibility of such attacks on online advertising systems, taking Facebook as our case study. We present a mechanism for releasing statistics on advertising campaigns in a way that preserves the privacy of web users. The notion of privacy that we adopt is a mathematically rigorous definition of privacy called *differential privacy*. In addition, we show that the seemingly overly restrictive notion of differential privacy is in fact the one necessary here, and that weaker notions would not suffice.

**Keywords.** Sum queries, Differential privacy, Online advertising, Facebook.

[*]This work was supported by a starting grant from the European Research Council.

# 1 Introduction

Online advertising plays an important role in supporting many Internet services. From a marketer's point of view, it serves as an extremely powerful tool, offering a way to direct ads at very specific audiences that the marketer can specify out of the vast body of Internet users. Being extremely easy to use and allowing continuous monitoring and updating, online advertising campaigns prove to be highly cost effective. However, many concerns arise regarding the implications of online advertising for the privacy of web users. Some concerns have to do with cases where a user clicks on an ad that was posted by a malicious company, allowing this company to extract information about the user, e.g., by keeping track of the user's browsing. However, as recent works show, Internet users may also be concerned with their privacy being breached even without ever noticing the presented ads, let alone clicking on them. In this work, we suggest practical measures for preventing the feasibility of such attacks on online advertising mechanisms. We use Facebook as our main case study, but our analyses and mechanisms are applicable to other online advertising systems, such as, Google, Yahoo, and others.

According to Facebook, no personal information of users is shared with advertisers, which are only allowed to select targeting criteria when creating a marketing campaign (e.g., specify the age, gender, and interests of people fitting the campaign). The actual posting of ads is then carried out automatically by Facebook's system, hence not (directly) revealing to the advertising company which users comply with their criteria. However, a recent paper by Korolova [12] considering the type of personalized advertising offered by Facebook shows that this may already allow a breach in the privacy of Facebook users. Korolova's main findings is that it is possible to exploit the campaign designing system offered by Facebook to reveal private information about specific Facebook users (that is, to reveal information about these users that is not in the public domain). Such private information may include sensitive data, such as, sexual preferences, political views, and religious affiliations. The key point of the work of [12] is that this blatant violation of the privacy of Facebook users is not a result of any malicious behavior on the side of Facebook, but rather that attackers can utilize the targeting capabilities given by the Facebook ad creation system combined with the seemingly benign statistics that Facebook conveys to the advertising company in order to evaluate and price their campaign. Korolova [12] explains that "Facebook provides detailed performance reports specifying the total number of impressions and clicks the ad has received, as well as the number of unique impressions and clicks, broken up by day; as well as rudimentary responder demographics. The performance report data is reported close to real time."

The general (high level) structure of the attacks proposed by [12] proceed as follows. In order to reveal some private information from the profile of a user U (that is, a certain feature $F$ that can be specified as a criterion in the targeting process, e.g., U's sexual preferences) – (i) collect auxiliary information regarding U's profile to uniquely identify U's profile from all other Facebook profiles (such auxiliary information may include: age range, location, and even specific known interests, such as "Likes Skiing"); (ii) form an ad campaign for each possible value of the targeted feature $F$, selecting targeting criteria that identify U (using the auxiliary information) (iii) observe impression and click reports received by the campaign; if all impressions belong to a unique user and appear only in a single campaign (where a criterion was given for the feature $F$ to equal some value $f$), then assume it is U and conclude that $f$ is the value appearing in U's profile; if impressions are not unique, refine the auxiliary information and rerun the attack. A specific attack scheme is described in Appendix B, the success of this attack does not rely on the user U ever noticing or clicking on the ad. It suffices that for the ad to be displayed on U's account. Hence, a person's privacy may be breached without her ever noticing it, and she is also powerless to prevent it.

Korolova [12] discusses two possible solutions for dealing with the type of attacks proposed in her paper. The first solution seems to be the one that Facebook implemented after being notified of Korolova's results. Their solution is to introduce some minimum lower bound on the "estimated reach" of any proposed campaign (the threshold they use is of about 20 users). Korolova suggests to circumvent this restriction by creating 20 fake Facebook accounts that will all agree with the criteria that the subject of the attack (i.e., the user U) currently uniquely satisfies. In [11], Facebook are quoted to argue that this strategy would not work since their system will detect 20 similar accounts being created in close proximity. This restriction may make the attacker's task somewhat harder, however, since this is a deterministic ad-hoc solution it may very well leave the path open for new variants of Korolova's attacks (even without resorting to the construction of fictitious accounts).[1]

The second solution that Korolova discusses is the one she suggests Facebook should adopt, namely, to only use information that is already public (i.e., user profile information designated as visible to "Everyone"). While this would evidently render all the attacks discussed above completely useless, it seems to be a solution that Facebook would have a hard time complying with since it would severely limit the utility of targeted advertising campaigns. We suggest a third solution that allows Facebook to use private data in its advertising system, while preserving privacy via mechanisms that provide rigorous and provable privacy guarantees.

**Differential Privacy is a most relevant privacy notion.** A criterion for mechanisms that provide analyses over a data set, while preserving privacy of individual entries has evolved in a sequence of recent works [4, 9, 8, 2, 7, 5, 6]. This definition is called *differential privacy* and it relies on a rigorous mathematical treatment capturing the intuition of what privacy means in this setting. Informally, differential privacy guarantees that hardly any information is gained on individual records upon seeing the output of a computation. As we explain in Section 2, differential privacy is a very restrictive definition of privacy, allowing a possible attacker extensive power. Specifically, the definition assumes that an attacker is able to determine all data-base entries except for the entry under attack. Such a notion may seem much too stringent at first sight as it is hard to imagine an attacker capable of manipulating a substantial fraction of a huge data base such as the Facebook user network (estimated to consist of some 500,000,000 users). However, in this work we argue that differential privacy is actually exactly the right notion of privacy for the task at hand and that a weaker notion would not suffice. This is shown in Section 2.1 by observing that the attacker of [12] effectively has the type of power an attacker is assumed to have according to the definition of differential privacy. That is, while the attacker is evidently unable to manipulate the entries of existing Facebook users, it is able to pick the value of all the entries of some related Boolean abstract database, implied by the real Facebook network and the advertising campaign held by the attacker.

**A differentially-private mechanism for releasing online advertising statistics.** Despite the restrictive definition, it has been demonstrated that differentially private analyses exist for a variety of tasks including the approximation of numerical functions (by adding carefully chosen random noise that conceals any single individual's contribution) [7, 2, 19, 10]. Still, differential

[1] E.g., If the attacker is able to identify a set of, say, 30 users, such that, the user U is one of them and all other 29 users agree on the feature of interest $F$, then it can run Korolova's attack only trying to distinguish whether the number unique impressions is 29 or 30 (rather than whether it is 0 or 1, in the original attack).

privacy remained almost exclusively the subject of theoretical research. A few recent works have made some progress in showing that differentially private constructions can be applicable to practical problems, see e.g., [13, 16, 15]. The work of [14] introduces the Privacy Integrated Queries (PINQ) language, as a platform for conducting private data analysis. The goal of [14] was to make much of the successful privacy research available to privacy non-experts. The work of [16] considers the problem of producing recommendations from collective user behavior while simultaneously providing guarantees of privacy for these users. Specifically, they show how leading recommendation algorithms for the Netflix Prize data set can be adapted to the framework of differential privacy, without significantly degrading their accuracy. Nevertheless, with the exception of [16], we are not aware of concrete uses of differential privacy with real-world parameters for a real-world problems.

In this paper we construct a differentially private mechanism for releasing statistics used for monitoring and pricing of online Facebook advertising campaigns. Our mechanism is quite straightforward and builds on known techniques from the area of differential privacy. In Section 3.3 we suggest concrete privacy parameters for our mechanism so that privacy of Facebook users will be preserved, while the accuracy of the released statistics remains reasonable even for very small campaigns. We believe that the simplicity of this mechanism together with the low cost it inflicts on the accuracy of the released statistics will serve as an incentive for Facebook (and other online advertising providers that are concerned with preserving the privacy of users) to use it in practice. We believe it is of high value to bring the notion differential privacy to the knowledge and awareness of such large Internet agencies and that by employing differentially private mechanisms, these agencies will make a grand step towards preserving the privacy of web users.

In Section 4, we demonstrate how our mechanism can be adjusted to release different statistics from the ones discussed in our main construction in Section 3. This leads to a very flexible mechanism allowing online advertising providers control over the trade off between the number of different statistics released and the degree of accuracy of these statistics. This control will allow for different mechanisms to be used for large advertising campaigns from those used for small advertising campaigns. Specifically, large campaigns may receive more statistics allowing larger additive error, as opposed to small campaigns that cannot allow too much error, but require less statistical measurements.

In Section 3.4 we give some experimental evidence that our mechanism indeed causes only a reasonable decrease in the accuracy of the released data.

## 2 Differential Privacy and its Relevance to Online Advertising

We use the definition of *differential privacy* suggested in [7] to capture the notion of individual privacy. The privacy is defined to be a property of the database mechanism (rather than, say, the output of the computation or the knowledge of the adversary). Informally, we require that a change of any single entry in the database may only slightly change the distribution of the responses of the database seen by the user (i.e., the view of a possible adversary). Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a database, where each entry $x_i$ is taken from some domain $\mathcal{D}$. Define the *Hamming distance* between two databases $\mathbf{x}, \mathbf{x}'$ to be the number of entries on which $\mathbf{x}$ and $\mathbf{x}'$ do not agree. That is,

$$\mathrm{d}_H(\mathbf{x}, \mathbf{x}') = \left|\left\{i : x_i \neq x'_i\right\}\right|.$$

We say that two databases $\mathbf{x}, \mathbf{x}'$ are a *neighboring pair* if they differ in exactly one entry, i.e., $\mathrm{d}_H(\mathbf{x}, \mathbf{x}') = 1$.

The above refers to a mechanism, which is just a computational process applied to a database $\mathbf{x}$ and a query $q$. The output of a mechanism is some (possibly) randomized function of the database and the query, i.e., $\hat{f}(\mathbf{x}, q)$. By way of example, to view the scenario of evaluating the performance of online advertising campaigns, consider the set of user profiles as a database $\mathbf{x}$ and a campaign as a query $q$, mapping each user U (i.e., database entry) to a pair of values $(I_U, C_U)$, where $I_U$ denotes the number of times that the ad was presented to U, and $C_U$ denotes the number of times that U clicked on the presented ad. more precisely, an aggregated sum of impressions and clicks throughout a set period of time, e.g., a day, defines a query. Note that each campaign (and time period) defines a new query. However there is usually a single mechanism for answering all possible queries. The naive mechanism used initially by Facebook is simply to return the actual campaign statistics. This mechanism is deterministic, a fact that is used crucially in the attack of Korolova [12]. As we will see, under the formulation of differential privacy, any privacy-preserving mechanism must be randomized.

We next define what it means for a function to preserve differential privacy. For simplicity, we give a definition for function with respect to a fixed query $q$. However, it is possible to also define a general function where the query type is also given as input.

**Definition 2.1 ($\varepsilon$-differential privacy [7])** *Let $\hat{f} : \mathcal{D}^n \to R$ be a randomized function. We say that $\hat{f}$ is $\varepsilon$-*differentially private *if for all neighboring vectors $\mathbf{x}, \mathbf{x}'$, and for all possible sets of outcomes $\mathcal{V} \subseteq R$ it holds that*

$$\Pr[\hat{f}(\mathbf{x}) \in \mathcal{V}] \leq e^{\varepsilon} \cdot \Pr[\hat{f}(\mathbf{x}') \in \mathcal{V}], \tag{1}$$

*where the probability is taken over the randomness of $\hat{f}$. We say that a mechanism $\mathcal{S}$ is $\varepsilon$-*differentially private *if the randomized function it computes is $\varepsilon$-differentially private.*

One way to understand the definition is as a mental game, in which we let an adversary pick $i$ and pick all entries in the database except for $x_i$; in addition, we let the adversary pick (any) two possible values $x_i, x_i'$ for the $i$-th entry; we fix the $i$-th entry to either $x_i$ or $x_i'$ and apply the $\varepsilon$-differentially private mechanism ($\varepsilon$ is the privacy parameter and we think of it a small constant) to the database; finally, we let the adversary try to distinguish which of the two values of $x_i$ we chose. Consider for example, the case where $x_i$ is taken with probability $1/2$ and let $p$ be the probability that the adversary succeeds in guessing the $i$-th entry. We say that $p - \frac{1}{2}$ is the advantage of the adversary (over an adversary that simply guesses by tossing a fair coin). Then, the definition above says that the advantage of the adversary will be at most $\varepsilon/2$. To see this, consider the set $\mathcal{V}$ of outputs, such that, for all $y \in \mathcal{V}$ the probability of outputting $y$ is higher with $x_i$ than $x_i'$. Hence, the best possible strategy of the adversary is to guess $x_i$ whenever the output is $y \in \mathcal{V}$ and $x_i'$ whenever the output is $y \notin \mathcal{V}$. Denote by $X$ the $i$-th entry, denote by $Y$ the output of the mechanism, and denote by WIN the event that the adversary guesses the $i$-th entry correctly in this experiment. Hence, the probability of the event WIN is

$$\begin{aligned}
\Pr[\mathrm{WIN}] &= \Pr[Y \in \mathcal{V} \mid X = x_i] \cdot \Pr[X = x_i] + \Pr\left[Y \notin \mathcal{V} \mid X = x_i'\right] \cdot \Pr\left[X = x_i'\right] \\
&= \frac{1}{2} \cdot \Big( \Pr[Y \in \mathcal{V} \mid X = x_i] + \Pr\left[Y \notin \mathcal{V} \mid X = x_i'\right] \Big) \\
&= \frac{1}{2} \cdot \Big( \Pr[Y \in \mathcal{V} \mid X = x_i] + 1 - \Pr\left[Y \in \mathcal{V} \mid X = x_i'\right] \Big) \\
&= \frac{1}{2} + \frac{1}{2} \cdot \Big( \Pr[Y \in \mathcal{V} \mid X = x_i] - \Pr\left[Y \in \mathcal{V} \mid X = x_i'\right] \Big) \\
&= \frac{1}{2} + \frac{1}{2} \cdot \Big( \Pr\left[Y \in \mathcal{V} \mid X = x_i'\right] \cdot e^{\varepsilon} - \Pr\left[Y \in \mathcal{V} \mid X = x_i'\right] \Big) \\
&= \frac{1}{2} + \frac{1}{2} \cdot \Pr\left[Y \in \mathcal{V} \mid X = x_i'\right] \cdot \Big( e^{\varepsilon} - 1 \Big) \\
&\leq \frac{1}{2} + \frac{e^{\varepsilon} - 1}{2}
\end{aligned}$$

where the fifth equality follows from Definition 2.1. For small values of $\varepsilon$, it holds that $e^{\varepsilon} \approx 1 + \varepsilon$, hence obtaining that the probability of the adversary guessing correctly is

$$\Pr[\mathrm{WIN}] \leq \frac{1}{2} + \frac{e^{\varepsilon} - 1}{2} \approx \frac{1}{2} + \frac{\varepsilon}{2}.$$

## 2.1 Differential Privacy is Necessary in Practice

Differential privacy seems to be a overly stringent notion of privacy, since, as explained above, it allows the adversary to select *all* entries in the database other than the entry under attack, and in the addition to select two possible values for this last entry. One may dismiss the necessity of such a strict notion arguing that it is hardly ever the case that the adversary has auxiliary information about most entries in a huge database such as that of Facebook, let alone, select all the entries but the one it is attacking. However, we show that the seemingly overly strong definition of differential privacy is actually exactly what is needed here and nothing weaker. We use the attack of Korolova [12] to illustrate our point.

**The attacker of [12].** We first outline the two main steps in the general scheme of the attacks described in [12]. In Appendix B, we give a more detailed description of one specific attack. Let U be a Facebook user profile for which we are interested in inferring information classified as restricted to a small subset of Facebook users, e.g., marked with "Only me" or "Friends only" visibility mode (that is, a certain feature $F$ that can be specified as a criterion in the targeting process, e.g., U's sexual preferences).

1. Collect auxiliary information on the user U, on its Facebook profile, and on what distinguishes U's profile from the rest of the Facebook user network. Form a set of criteria (excluding any criterion on the feature of interest $F$) in order to uniquely identify the user U.

2. Run a campaign with the above criteria for every possible value $f$ for the feature $F$. If there is exactly one campaign (invoked with some value $f'$ for $F$) for which impressions (or clicks) are reported and in this campaign they are unique (i.e., attributed to a single user), then conclude that $f'$ is the value appearing in U's profile for the feature $F$. Otherwise, update auxiliary information and run the attack again.

For the sake of simplicity, consider the case where the above attacker only observes impression reports on the campaigns (see Figure 2 for an example of such attack). Each campaign that the attacker invokes defines a sum-query on the database of Facebook users. More precisely, it defines a count query $q$, i.e., $q$ maps each entry either to 0 (if the ad was never presented to the user, hence, no impressions are attributed to this user) or to 1 (if the ad was presented to the user, hence, impressions are attributed to this user). Each such query $q$ defines an abstract Boolean database $D_q$ such that $D_q$ contains an entry for every Facebook user consisting of a single bit, where this bit is set to the value assigned to the user by $q$. Now, the first part of all the attacks of [12] is to come up with a campaign that uniquely identifies a user U. In other words, the attacker first tries to find a query $q$, such that, the (abstract) Boolean database $D_q$ contains all zeros in all entries but the entry of the subject of the attack U. The attacker is able to follow through with the attack whenever it is able to find such a query $q$ (i.e., a criteria that uniquely identifies the user U).

In other words, the first step in the attack of [12] is to bring the attacker to a point that it can define an abstract database of the size of the Facebook network (i.e., containing some 500,000,000 entries) for which the attacker is able to select the value of all entries but the entry of U. In the second step of the attack, the attacker succeeds if it wins a very similar game to the mental gave defined by the definition of differential privacy, i.e., the attacker needs to distinguish between the two possible (either 0 or 1) values of the entry of U in the abstract database. Note that while the attacker is evidently unable to select the actual profile of every Facebook user, all that is required for the success of the attack is for the attacker to be able to select the entries of the abstract database. Finally, the success of the experiments of Korolova [12] demonstrate that it is indeed possible to construct such an attacker, and hence, that any definition of privacy should indeed assume that the attacker is able to determine all entries in the database.

Surprisingly, powerful techniques exist for constructing mechanisms that yield useful outcome, and yet preserve differential privacy. In Section 2.2 we present one basic (and simple) technique, which belongs to a class of techniques for constructing analyses via output perturbation. In Section 3 we show that this technique can be used to construct a mechanism for releasing online advertising statistics while preserving the privacy of Facebook users against attacks such as those of [12]. For more on this class and on other techniques, see, e.g., [8, 2, 7, 19, 1, 17, 3]. In Appendix A we mention a few valuable properties of Definition 2.1 (for a deeper consideration of this definition the reader is referred to [18]).

### 2.2 Differential Privacy via Output Perturbation and Global Sensitivity

We are generally interested in constructing mechanisms that take a deterministic query $q$ and a database $\mathbf{x}$ and apply some randomized function $\hat{f}$ to obtain an output. In a world where privacy is not an issue the mechanism would just return $q(\mathbf{x})$ as its output, i.e., it would let $\hat{f}(\mathbf{x}, q) = q(\mathbf{x})$, hence yielding maximum utility. However, Definition 2.1 does not allow the function $\hat{f}$ to be deterministic. In this section we consider one way to construct mechanisms that yield valuable utility, while preserving differential privacy. For the sake of simplicity, we consider mechanisms that deal with a fixed query $q$, i.e., mechanisms that compute a function $\hat{f}_q(\mathbf{x}) = \hat{f}(\mathbf{x}, q)$ from some predetermined $q$.

Given a (deterministic) query $q : \mathcal{D}^n \to \mathbb{R}$, it is natural to ask whether there exists some randomized approximation $\hat{f}_q$ of $q$ that is differentially private. Clearly, this depends on our definition of approximation, but staying on the intuitive level, the answer to this question is correlated with the *sensitivity* of $q$, namely, the magnitude of change in the output of $q$, caused by a change in a single entry of the input. We next show that for queries that have low sensitivity (in a very strong sense), it is enough to mask the value of $q(\mathbf{x})$ by some carefully selected random variable.

**Query sensitivity [7].** Given a query $q : \mathcal{D}^n \to \mathbb{R}$, the *local sensitivity* is a function of both $q$ and a given database $\mathbf{x}$, defined by

$$\mathrm{LS}_q(\mathbf{x}) = \max_{\{\mathbf{x}':\mathrm{d}_H(\mathbf{x},\mathbf{x}')=1\}} \left|q(\mathbf{x}) - q(\mathbf{x}')\right|.$$

The *global sensitivity* is a function of $q$ taken to be the maximum local sensitivity over all databases $\mathbf{x}$, i.e.,

$$\mathrm{GS}_q = \max_{\mathbf{x}'\in\mathcal{D}^n} \left(\mathrm{LS}_q(\mathbf{x})\right).$$

In the case of the Facebook impressions statistic, the global sensitivity is the maximum number of times that a single person could conceivably view the advertisement within the given time period. In contrast, the global sensitivity of the unique impression statistic is exactly 1 (because each person either saw the advertisement or did not and this is what is counted).

**Differential privacy for sum queries from the Laplace distribution.** The *Laplace distribution*, $\mathrm{Lap}(\lambda)$, is the continuous probability distribution with probability density function

$$h(y) = \frac{\exp(-|y|/\lambda)}{2\lambda}.$$

For $Y \sim \mathrm{Lap}(\lambda)$ we have that $\mathbb{E}[Y] = 0$, $\mathrm{Var}[Y] = 2\lambda^2$, and $\Pr[|Y| > k\lambda] = e^{-k}$.
The following holds for all $y, y'$:

$$\frac{h(y)}{h(y')} = \frac{e^{\frac{-|y|}{\lambda}}}{e^{\frac{-|y'|}{\lambda}}} = e^{\frac{|y'|-|y|}{\lambda}} \le e^{\frac{|y-y'|}{\lambda}}\,, \tag{2}$$

where the inequality follows from the triangle inequality.

The framework of output perturbation via *global sensitivity* was suggested in [7]. In this framework we consider queries of the form $q : \mathcal{D}^n \to \mathbb{R}$. The outcome is obtained by adding to $q(\mathbf{x})$ noise sampled from the Laplace distribution, calibrated to $\mathrm{GS}_q$. Formally, $\hat{f}_q$ is defined as

$$\hat{f}_q(\mathbf{x}) = q(\mathbf{x}) + Y, \text{ where } Y \sim \mathrm{Lap}(\mathrm{GS}_q/\varepsilon). \tag{3}$$

This results in an $\varepsilon$-differentially private mechanism. To verify this, for a database $\mathbf{y}$, denote by $h_{\mathbf{y}}(\cdot)$ the probability density function of the distribution on the output of $\hat{f}_q(\mathbf{y})$. For every $v \in \mathbb{R}$ representing an outcome of $\hat{f}_q(\mathbf{x})$, it holds that $h_{\mathbf{y}}(v) = h(v - q(\mathbf{y}))$. Thus, for every pair of neighboring databases $\mathbf{x}, \mathbf{x}'$ and for every possible outcome $v \in \mathbb{R}$, we have that,

$$\begin{aligned}
\frac{h_{\mathbf{x}}(v)}{h_{\mathbf{x}'}(v)} &= \frac{h(v - q(\mathbf{x}))}{h(v - q(\mathbf{x}'))} \\
&\le e^{\frac{\varepsilon\left|(v-q(\mathbf{x}))-(v-q(\mathbf{x}'))\right|}{\mathrm{GS}_q}} && \text{(by Equation (2))} \\
&= e^{\frac{\varepsilon\left|q(\mathbf{x}')-q(\mathbf{x})\right|}{\mathrm{GS}_q}} \\
&\le e^{\varepsilon} && \left(\text{since } \mathrm{GS}_q \ge \left|q(\mathbf{x}') - q(\mathbf{x})\right|\right).
\end{aligned}$$

**Example 2.2** *A sum query* SUM : $\mathcal{D}^n \to \mathbb{R}$ *(where $\mathcal{D} = \{0, 1, \ldots, \gamma\}$, for some fixed value of $\gamma$) is defined as*

$$\mathrm{SUM}(\mathbf{x}) = \sum_{i=1}^{n} x_i.$$

*For every two neighboring $\mathbf{x}, \mathbf{x}' \in \mathcal{D}^n$ we have that* $|\,\mathrm{SUM}(\mathbf{x}) - \mathrm{SUM}(\mathbf{x}')| \le \gamma$ *and hence* $\mathrm{GS}_{\mathrm{SUM}} = \gamma$. *Applying Equation (3), we have that* $\hat{f}_{\mathrm{SUM}}(\mathbf{x}) = \mathrm{SUM}(\mathbf{x}) + Y$, with $Y \sim \mathrm{Lap}(\gamma/\varepsilon)$ *is an $\varepsilon$-differentially private; that is, we get a differentially private approximation of* SUM *with $O(\gamma)$ additive error.*

# 3 A Differentially-Private Mechanism for Releasing Facebook Advertising Statistics

## 3.1 The Facebook Database and Campaign Statistics

Before proceeding to describe our privacy-preserving solution, we describe the format of the database containing the advertising campaign statistics that are presented to the advertising company. We stress that this is an abstract view of the statistics for the purposed of defining the mechanism; the actual way that Facebook holds these statistics is completely irrelevant.

As we have mentioned, there are four main statistics provided currently by Facebook. These are:

1. *Impressions:* The number of times that the advertisement was shown within the defined time period.

2. *Unique impressions:* The number of different users that the advertisement was shown to within the defined time period.

3. *Clicks:* The number of times that the advertisement was clicked upon within the defined time period.

4. *Unique clicks:* The number of different users that clicked upon the advertisement within the defined time period.

We now define the relevant database and queries for the above. Each entry in the database is a triple $(U, I, C)$ where $U$ is a user's identity, $I \in \mathbb{N}$ is the number of times that this user was shown the advertisement, and $C \in \mathbb{N}$ is the number of times that this user clicked on the advertisement. Note that each advertisement/campaign defines a new database. In addition, we remark that the real Facebook database contains much more information, and this is actually what an attacker wishes to learn. However, it is not of relevance for fully defining the queries, and thus is not of relevance for defining the privacy-preserving mechanism. The four query types associated with the aforementioned statistics over a database $\mathbf{x} = \{(U_j, I_j, C_j)\}_{j=1}^{n}$ are as follows:

1. *Impressions:* $q_1(\mathbf{x}) = \sum_{j=1}^{n} I_j$

2. *Clicks:* $q_2(\mathbf{x}) = \sum_{j=1}^{n} C_j$

3. *Unique impressions:* $q_3(\mathbf{x}) = |\{j \mid I_j > 0\}|$

4. *Unique clicks:* $q_4(\mathbf{x}) = |\{j \mid C_j > 0\}|$

We note that the unique impressions and clicks statistics are actually count queries, which are special cases of sum queries where all values are Boolean.

### 3.2 The Proposed Privacy-Preserving Mechanism

In this section we present our concrete proposal for releasing the impression and click statistics in a privacy-preserving manner. The general mechanism we suggest follows by composing four differentially private mechanisms (each denoted $\mathcal{S}_i$ and used with a privacy parameter $\varepsilon_i$, where $i \in \{1, \ldots, 4\}$); mechanism $\mathcal{S}_i$ relates to query $q_i$ defined above. Each of the four mechanisms works by calibrating noise to the sensitivity of one of the four query types, using the method described in Example 2.2 for sum queries. As will be explained below, we select the privacy parameter $\varepsilon_i$ for each sub-mechanism $\mathcal{S}_i$ according to the sensitivity of $i$'th sum-query $q_i$ and to the level of accuracy that we consider relevant to this query. By Lemma A.5, a mechanism releasing all four (perturbed) counts is $\varepsilon$-differentially private for $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$.

**The sensitivity of the Facebook statistics.** In order to use the mechanism of Example 2.2 (for each of the four sum-queries), we need to know the sensitivity of each sum-query. That is, we need to bound the change in each sum that may result in a change of the data of a single entry (i.e., the personal information of a single Facebook user). Regarding unique impressions and clicks, each database entry can change the result by only 1. Thus, the global sensitivity of the functions $q_3(\cdot)$ and $q_4(\cdot)$ is 1 (i.e., we have $\gamma_3 = \gamma_4 = 1$, where $\gamma_i$ denotes the global sensitivity of the $i$'th sum-query).

The case of non-unique impressions and non-unique clicks is less clear since we do not have an a priori bound on these counts. However, we argue that it quite easy to give some bound that is very unlikely to be crossed. This is especially true when giving the statistics per day, as does Facebook. For this case, we use $\gamma_1 = 20$ (the global sensitivity of $q_1(\cdot)$) as the bound on the number of impressions attributed to a single person in a given day, and we use $\gamma_2 = 3$ (the global sensitivity of $q_2(\cdot)$) as the bound on the number of clicks attributed to a single person in a given day. We note that such a bound can be artificially imposed (by Facebook in its statistics) by omitting counts of any additional impressions or clicks by the same user (alternatively, all counts may be taken into account, and the privacy guarantees rely on the low probability that some user surpasses the bounds).

Let $\varepsilon_1, \ldots, \varepsilon_4$ be privacy parameters that can be modified to tradeoff the privacy and utility, as desired (this will be discussed in more detail below). The mechanisms are then defined as follows:

**Mechanism $\mathcal{S}_1$ – output the perturbed number of impressions $Z_1$:** To obtain $Z_1$, we sample Laplacian noise $Y_1 \sim \mathrm{Lap}(\gamma_1/\varepsilon_1)$ and add it to the actual number of impressions. That is, the first sub-mechanism $\mathcal{S}_1$ outputs $Z_1 \leftarrow q_1(\mathbf{x}) + Y_1$, where $Y_1 \sim \mathrm{Lap}(20/\varepsilon_1)$.

**Mechanism $\mathcal{S}_2$ – outputting perturbed number of clicks $Z_2$:** To obtain $Z_2$, we sample Laplacian noise $Y_2 \sim \mathrm{Lap}(\gamma_2/\varepsilon_2)$ and add it to the actual number of clicks. That is, the second sub-mechanism $\mathcal{S}_2$ outputs $Z_2 \leftarrow q_2(\mathbf{x}) + Y_2$, where $Y_2 \sim \mathrm{Lap}(3/\varepsilon_2)$.

**Mechanism $\mathcal{S}_3$ – outputting perturbed number of unique impressions $Z_3$:** The mechanism outputs $Z_3 \leftarrow q_3(\mathbf{x}) + Y_3$, where $Y_3 \sim \mathrm{Lap}(1/\varepsilon_3)$.

**Mechanism $\boldsymbol{\mathcal{S}_4}$ – outputting perturbed number of unique clicks $\boldsymbol{Z_4}$:** The mechanism outputs $Z_4 \leftarrow q_4(\mathbf{x}) + Y_4$, where $Y_4 \sim \text{Lap}(1/\varepsilon_4)$.

In all of the above, it does not seem natural to ever release negative values. To prevent this from happening, whenever the outcome is negative (i.e., to ensure $Z_i \geq 0$), we simply round the $Z_i$ to 0 whenever $Y_i < -T_i$. It is easy to verify that this has no effect on the privacy of the mechanism (this is because the rounding operation is applied to the output of the mechanism, and does not depend on the original database and query; indeed, this operation can only reduce the amount of information that is released on the database).

### 3.3 Practical Choice of Privacy Parameters for Mechanism $\boldsymbol{\mathcal{S}}$

In this section we suggest how to instantiate the privacy parameters $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ of the mechanism $\mathcal{S}$ described in the previous section. We stress that this is just one suggestion. The principles described below for making this choice can be used to tailor the actual privacy and utility requirements, as desired by Facebook or any other advertising agency.

**The choice of privacy parameters.** Before presenting the concrete privacy parameters for our mechanism, we discuss how to choose these parameters. The tension between privacy and utility is inherent; hence, our goal is to find privacy parameters that optimize this trade off. Specifically, a smaller value of $\varepsilon$ yields a stronger privacy guarantee but also reduces the accuracy of the published statistics (since more Laplacian noise is added). Conversely, a larger $\varepsilon$ yields more accurate results, but decreases the level of privacy. We choose the privacy parameters by first fixing the privacy parameter for the overall mechanism $\mathcal{S}$ and then adjusting the privacy parameters of the sub-mechanisms to comply with this parameter. That is, our first requirement is that the mechanism $\mathcal{S}$ will indeed preserve privacy in a meaningful sense. We, hence, fix $\varepsilon = 0.2$ to be the privacy parameter for the overall mechanism and we seek privacy parameters for the sub-mechanisms $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ such that $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$.

To choose $\varepsilon_i$, we take into account the sensitivity of the appropriate query and the magnitude of error that we deem reasonable for this query. Specifically, in choosing $\varepsilon_1$ (i.e., the privacy parameter for releasing the number of impressions), we consider an error of a few thousands of impressions to be reasonable, even for small campaigns (mainly since impressions are counted in thousands). Since the global sensitivity of this sum-query $\gamma_1 = 20$ implies that the add noise $Y_1$ will be sampled from the Laplace distribution with scale parameter $20/\varepsilon_1$, we take $\varepsilon_1 = 0.03$ to obtain the desired guarantees. In choosing $\varepsilon_2$ (i.e., the privacy parameter for releasing the number of clicks), we consider an error of a bit over a hundred clicks to be reasonable. Since the global sensitivity of this sum-query $\gamma_2 = 3$ implies that the add noise $Y_2$ will be sampled with scale parameter $3/\varepsilon_2$, we take $\varepsilon_2 = 0.11$ to obtain the desired guarantees. In choosing $\varepsilon_3$ and $\varepsilon_4$ we follow similar considerations, where in the case of unique impressions, we consider an error magnitude of 500 to be highly acceptable, and in the case of unique clicks, we consider an error magnitude of 70 to be acceptable. We, hence, take $\varepsilon_3 = 0.01$, and $\varepsilon_4 = 0.05$ to be the privacy parameters of our mechanisms. Hence, by Lemma A.5, the privacy parameter we obtain for the overall mechanism is indeed $\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = 0.2$. We remark that it is possible to allow a user (advertiser) to request any setting of the $\varepsilon_i$ parameters as long as the sum of all four does not exceed 0.2. (This makes sense if the accuracy of one of the statistics is much more important to a given advertiser than the others.)

**Probability of error.** Adding random noise to the released statistics reduces their correctness. Clearly, too much noise may reduce the accuracy of the evaluation of the effectiveness of the campaign. Furthermore, it may cause the marketer to pay more than is reasonable for such a campaign. We argue that the error will only be meaningful for very small campaigns (i.e., campaigns with a very small expected number of impressions and clicks). This is true since noise is added irrespective of the count results, and depends only on the privacy parameter $\varepsilon$. Hence, the probability of large error relative to a large count is exponentially small. We now analyze the size campaigns should have in order for the error to be meaningless, according to the parameters we selected.

**Error in impression count ($Z_1$):** We analyze the probability that the reported number of impressions is with distance at least $k$ from the actual impression count. Since this distance is exactly the amount of noise we added, we need to bound the probability that we added noise of size at least $k$, i.e., $|Y_1| \geq k$. By the properties of the Laplace distribution, this probability is $e^{-k\varepsilon_1/20} = e^{-0.0015k}$. Hence, for an error of size $k = 2000$, this probability is 0.05, and for an error of size $k = 5000$, this probability is 0.0005. Recall that impressions are counted in thousands, and hence this should not be considered to be a large error.

**Error in click count ($Z_2$):** We analyze the probability that the reported number of clicks is with distance at least $k$ from the actual click count. I.e., we analyze the probability of $|Y_2| > k$. By the properties of the Laplace distribution, this probability is $e^{-k\varepsilon_2/3} = e^{-0.0367k}$. Hence, for an error of size $k = 50$, this probability is 0.16 and for error size $k = 100$, this probability 0.025, however, for an error of size $k = 200$, this probability decreases to 0.0006. We remark that this statistic seems to incorporate the most noise of all four counts.

**Error in unique impression count ($Z_3$):** We analyze the probability of an error of size $k$ in reporting the number of unique impressions, i.e., the probability that $|Y_3| > k$. By the properties of the Laplace distribution, this probability is $e^{-k\varepsilon_3} = e^{-0.01k}$. Hence, for an error of size $k = 250$, this probability is 0.082, and for an error of size 500, this probability is 0.0067.

**Error in unique click count ($Z_4$):** We analyze the probability of an error of size $k$ in reporting the number of unique impressions, i.e., the probability that $|Y_4| > k$. By the properties of the Laplace distribution, this probability is $e^{-k\varepsilon_4} = e^{-0.05k}$. Hence, for an error of size 50, this probability is 0.082, and for an error of size 100, this probability is 0.0067.

**Remark 3.1** *Evidently, using the suggested mechanism may cause small advertising campaigns to be somewhat less cost effective than they currently are. However, we believe that most genuine advertisers aim for large enough campaigns and may only suffer small percentage of additional cost. Furthermore, a marketer that invokes many (large enough) campaigns, will eventually suffer almost zero additional cost since, on average, a zero amount of noise will be added to the reported statistics (for small campaign this may not be true, due to the one-sided rounding of negative outputs).*

**The meaning of $\varepsilon$-differential privacy with $\varepsilon = 0.2$.** We now try to shed some light on the privacy guarantee that is yielded by our mechanism. Consider an attacker that executes an attack similar to the one suggested in [12] aimed at revealing some private information of a specific Facebook user. For the sake of being specific, consider an employer trying to categorize its employees by their sexual orientation. That is, for each employee the employer would like to know whether the employee is heterosexual or homosexual. For those employees that the employer does not know the answer, it runs an attack similar to that of [12] trying to obtain this information.

First, consider the employer running a single attack. This employer runs a campaign and gets the report of the statistics. The employer then applies some predicate $P$ to the resulting statistics. An employee U is categorized as homosexual if and only if the predicate is true (that is, if $P = 1$). By the $\varepsilon$-differential privacy, the ratio between the probability that U is categorized as being homosexual when this is indeed stated in his Facebook user profile and the probability that U is categorized as being homosexual when this is not stated (even if the opposite is stated) in his Facebook user profile is bounded by $e^{\varepsilon} = e^{0.2}$. Put differently, this means that for any network of users, if the probability that a user U is categorized as homosexual when this *is* stated in U's Facebook profile is $\alpha$, then the probability that a user U is categorized as homosexual when this *is not* stated in U's Facebook profile is at least $0.82\alpha$.

By Lemma A.5, the employer needs to run the attack at least 5 times in order to decrease the privacy parameter to 1, that is, to allow the ratio between the above probabilities be $e$. Even in such a case, it would only mean that for any network of users, if the probability that a user U is categorized as homosexual when this is stated in U's Facebook profile is $\alpha$, then the probability that a user U is categorized as homosexual when this is not stated in U's Facebook profile is at least $0.37\alpha$. To be even more concrete, consider an employer running the aforementioned attack on a mechanism that is 1-differentially private (e.g., by running 5 attacks as described above) with the aim of categorizing 30 employees, where 10 are homosexual and 20 are heterosexual. Assume that the employer concludes that 5 of the 10 homosexual employees are indeed homosexual (i.e., is correct for half of them), then by differential privacy the employer will also conclude that 4 of the heterosexual employees are also homosexual. Thus, the employer's certainty regarding any given employee is low.

We stress that an adversary can increase its certainty significantly by running many campaigns. For example, if it runs the campaign 20 times then the differential privacy guarantee will only give $\varepsilon = 4$. Continuing the above concrete example, if the employer correctly categorizes all of the homosexual employees, then the differential privacy guarantee merely states that it will categorize a heterosexual employee as homosexual with probability $20/e^4 \approx 0.37$. Thus, the definition no longer guarantees that the employer will not correctly categorize all employees with high probability.

However, we argue that the time and cost efficiency of such attacks should be evaluated in a broader view. Recall that for running attacks similar to that of [12], even on a system giving unperturbed data, requires running many different campaigns before obtaining sufficient auxiliary information on the relation between the user U and the rest of the entries in the database. Furthermore, even when the attack of [12] is conclude (i.e., when there is only one entry in the database that satisfies the campaign criteria), there is still substantial probability that this is not the entry that the attacker was aiming at. To overcome this, the attacker will yet again need to run a few campaigns with different sets of criteria. We thus argue that if revealing the sexual preferences of a known person (e.g., a friend or an employee, i.e., one on which we already obtain valuable auxiliary information) requires a month or two of dedicated work and analysis, it might be easier to try and obtain this new information in the real-world. Furthermore, the cost of a successful attack seems to become higher than a reasonable attacker may be willing to spend. We note that if the attacker is interested in revealing some information about a large group of users (e.g., all employees in the company) using methods similar to those of [12], then this will require many campaigns for each of the users separately, in order to uniquely identify each of users by the criteria of the campaign.

In addition to the above, combining the above mechanisms with some effort in detecting attacks on the privacy of users, it may be possible to ensure much better privacy for users. Such efforts are already taken by Facebook in trying to detect the construction of a number of extendedly similar profiles. It may also be helpful to try to detect the invocation of, say, a few relatively small campaigns having similar sets of criteria.

It is always possible to make it harder for an attacker by taking smaller privacy parameters $\varepsilon_i$ for each mechanism. However, considering the trade off between the additional cost it would inflict on (legitimate) small campaigns, together with the high penalty the attacker already needs to pay (as explained in the foregoing discussion), we argue that our parameters are close to optimal for the task at hand. Still, we encourage online advertising service providers to consider the smallest security parameters that the system can work with. It may be the case that a small change in the pricing of small campaigns, may allow the use of smaller parameters (we stress, however, that the pricing should not depend on the amount of noise that was added, since this may disclose private information). It is also possible for Facebook to decide that less campaign statistics are already sufficient, allowing choosing a smaller epsilon with no additional cost to small advertising campaigns.

### 3.4 Experimental Results

It is inherent to the definition that any differentially private mechanism for releasing campaign statistics must degrade the accuracy of the data released. In Section 3.3 we gave mathematical arguments as to why this decrease in accuracy can still be considered reasonable, even with respect to small advertising campaigns. In this section, we describe the result of experimenting with measurements taken from four real-life campaigns. We considered this experiment as a kind of sanity check, and ran it repeatedly with the same data of the four campaigns, where in each time we selected fresh Laplacian random variables $Y_1, Y_2, Y_3$, and $Y_4$ (as described in Section 3). Indeed, our finding is that the decrease in accuracy is tolerable. Below, we describe the experiment in more detail. In Figure 1, we describe the results of a single experiment, i.e., a selection of the appropriate noise for each of the measurements for each of the four campaigns.

We took the raw data available in Christian Thurston's tutorial on Facebook online advertising reports [20]. We considered the four campaigns that are the used as the example in this tutorial. The raw data from these reports appear in the upper part of the table in Figure 1. We applied Mechanism $\mathcal{S}$ to each of the campaigns, by selecting Laplace random variables $Y_1, Y_2, Y_3$, and $Y_4$ for each of the four campaigns, where all random variables were selected independently of each other (with scale parameter as prescribed by the mechanism, e.g., $Y_1 \sim \text{Lap}(20/\varepsilon_1)$). The resulting (perturbed) data appears in the lower part of the table in Figure 1.

One measurement that may indicate the magnitude of change in the evaluation of the raw data versus that of the perturbed data is the click through rate (CTR), which denotes the percentage of impressions (i.e., times that an ad was shown) that ended up in a user clicking the ad. This is a very important measurement as it is used by Facebook to determine whether to present an ad to the users matching the criteria of the campaign (together with the bid of the marketer). As the table shows, the change from the CTR values in the original data to the noisy CTR is only a minor one.

| Impressions | Y1 | Clicks | Y2 | CTR | Unique impressions | Y3 | Unique clicks | Y4 | Unique CTR |
|---|---|---|---|---|---|---|---|---|---|
| 177028 | | 171 | | 0.10 | 10709 | | 161 | | 1.50 |
| 10252.00 | | 2.00 | | 0.02 | 3055.00 | | 2.00 | | 0.07 |
| 36222.00 | | 120.00 | | 0.33 | 11735.00 | | 19.00 | | 0.16 |
| 212659.00 | | 97.00 | | 0.05 | 34263.00 | | 97.00 | | 0.28 |
| Noisy Impressions | | Noisy Clicks | | Noisy CTR | Noisy Unique impressions | | Noisy Unique clicks | | Noisy Unique CTR |
| 176334.35 | -693.65 | 156.17 | -14.83 | 0.09 | 10624.50 | -84.50 | 195.05 | 34.05 | 1.84 |
| 10608.23 | 356.23 | 0.00 | -4.80 | 0.00 | 3075.29 | 20.29 | 2.76 | 0.76 | 0.09 |
| 36301.08 | 79.08 | 127.48 | 7.48 | 0.35 | 11731.11 | -3.89 | 18.44 | -0.56 | 0.16 |
| 211845.50 | -813.50 | 94.09 | -2.91 | 0.04 | 34198.63 | -64.37 | 102.26 | 5.26 | 0.30 |

Figure 1: A table demonstrating the effect of applying Mechanism $\mathcal{S}$ to the raw data of four real-world advertising campaigns.

## 4 Releasing Additional Statistics

Recently Facebook began providing its advertisers with additional count statistics on their campaigns. These counts are called social impressions and social clicks, both of which count an action of a user if it first happened for a friend of that user, e.g., the social impression count is incremented whenever a user U was presented with an ad after a friend of U was presented with the same ad. This raises the question of how to release additional statistics on a given campaign, while preserving similar privacy and utility guarantees.

It is possible to apply the technique discussed in Example 2.2 to derive a mechanism for each such new sum query (provided that an upper bound $\gamma$ can be imposed on the count of each user). However, releasing the result of each such mechanism would degrade the level of privacy of the overall mechanism. We argue that this should not be allowed. We suggest an alternative approach, which yields much flexibility to the advertising system and the marketers in addition to maintaining a fixed level of privacy. We suggest that the service provider (e.g., Facebook) predetermines the privacy parameter $\varepsilon$ for the overall mechanism. To construct a data releasing mechanism for a specific campaign, the provider will allow the marketer to choose a subset of a set of statistics (sum queries). Given a subset of queries $\{q_i\}$, the service provider will select the privacy parameter $\varepsilon_i$ for each sub-mechanism $\mathcal{S}_i$ (for releasing the result of $q_i$) depending on the sensitivity of $q_i$ (i.e., on $\gamma$) and on the assumed level of accuracy that is relevant to this query.[2]

Using the approach we suggest, the marketer is given control over the trade off between the number of different sums released and the accuracy of each of these sums, while the overall level of privacy of the campaign is preserved (by the fixed parameter $\varepsilon$). The main idea behind this approach is that large campaigns should be treated differently from small campaigns. Specifically, a large campaign may allow larger additive error in each sum query, hence, allowing the use of better privacy parameters (i.e., smaller $\varepsilon_i$) for the sub-mechanisms. Thus, it becomes possible to release the answers of more sum queries while preserving the same overall privacy. In contrast, in smaller campaigns, the accuracy of each query may be more crucial to the evaluation of the effectiveness of the campaign (since a large additive error to a small sum may yield a meaningless

[2] See the way we chose the concrete parameters in the suggested mechanism we present in Section 3.3 for more intuition on this approach. Specifically, in Section 3.3 we explain the considerations we use in choosing the privacy parameters of our sub-mechanisms.

result). Thus the marketer holding the campaign may prefer to receive less (but more accurate) counts. Furthermore, it may well be the case that many of the counts that are of interest in large campaigns are quite irrelevant in smaller ones. Consider for example the social impressions and social clicks discussed above. These counts seem to be meaningless for campaigns with only a few hundred impressions and a few tens of clicks.

# References

[1] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proc. of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, New York, NY, USA, 2007. ACM.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *Proc. of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

[3] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proc. of the 40th ACM Symp. on the Theory of Computing*, pages 609–618, New York, NY, USA, 2008. ACM.

[4] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.

[5] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Proc. of the 33rd International Colloquium on Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer-Verlag, 2006.

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

[7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Proc. of the Third Theory of Cryptography Conference – TCC 2006*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer-Verlag, 2006.

[8] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In M. Franklin, editor, *Advances in Cryptology – CRYPTO 2004*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer-Verlag, 2004.

[9] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 211–222, 2003.

[10] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proc. of the 41st ACM Symp. on the Theory of Computing*, 2009.

[11] M. Helft. Marketers can glean private data on facebook. The New York Times, 2010. See: http://www.nytimes.com/2010/10/23/technology/23facebook.html.

[12] A. Korolova. Privacy violations using microtargeted ads: A case study. IEEE International Workshop on Privacy Aspects of Data Mining (PADM 2010), 2010. See: http://theory.stanford.edu/~korolova/Privacy_violations_using_microtargeted_ads.pdf.

[13] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *WWW*, pages 171–180. ACM, 2009.

[14] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Commun. ACM*, 53(9):89–97, 2010.

[15] F. McSherry and R. Mahajan. Differentially-private network trace analysis. In *SIGCOMM*, pages 123–134, 2010.

[16] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *KDD*, pages 627–636, 2009.

[17] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proc. of the 48th IEEE Symp. on Foundations of Computer Science*, pages 94–103, 2007.

[18] K. Nissim. Private data analysis via output perturbation a rigorous approach to constructing sanitizers and privacy preserving algorithms. In Charu C. Aggarwal and Philip S. Yu, editors, *Privacy-Preserving Data Mining: Models and Algorithms*, volume 34 of *Advances in Database Systems*, pages 383–414. Springer Publishing Company, Incorporated, 2008.

[19] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proc. of the 39th STOC*, pages 75–84, 2007.

[20] C. Thurston. An explanation of facebook advertising reports. http://www.internetmarketingsolution.com.au/facebook-ad-performance-reports.html, 2010.

# A Differential Privacy – A Few Important Properties

In this section we make a few remarks about Definition 2.1 and mention some nice properties it entails.

## A.1 A Relaxed Privacy Definition

A natural relaxation of Definition 2.1 allows for events occurring with negligible probability, for which the definition does not hold (i.e., the ratio between probabilities of these events occurring with some neighboring inputs is not bounded by $e^{\varepsilon}$). The next two examples give some motivation for this relaxation.

**Example A.1** *Consider a mechanism that given $x \in \{0,1\}$ outputs $x + Y$ where $Y$ is sampled according to $\mathrm{Lap}(1/\varepsilon)$. For every $x, x' \in \{0,1\}$ it holds by Equation (2) that $\frac{h_x(v)}{h_{x'}(v)} \leq e^{\varepsilon}$. It is easy* *to see that this implies the requirement of Definition 2.1. Hence, this mechanism is $\varepsilon$-differentially private.*

*We remark that this mechanism yields almost no usefulness, however as shown above, generalizations of these ideas prove highly useful in constructing differentially private analyses when dealing with larger databases.*

**Example A.2** *Consider a very similar mechanism to that of Example A.1, which given $x \in \{0, 1\}$ outputs $x + Y'$ where $Y'$ is obtained by limiting the random variable $Y$, sampled as before (i.e., $Y$ is sampled according to* $\mathrm{Lap}(1/\varepsilon)$*), to be within the interval $[-k/\varepsilon, k/\varepsilon]$, for some large $k$ (that is, if $Y > k/\varepsilon$ we set $Y' = k/\varepsilon$, similarly, if $Y < -k/\varepsilon$ we set $Y' = -k/\varepsilon$, and otherwise we set $Y' = Y$).*

*Note that the resulting mechanism is no longer $\varepsilon$-differentially private since* $\Pr[\mathcal{S}(0) > k/\varepsilon] = 0$ *while* $\Pr[\mathcal{S}(1) > k/\varepsilon] > 0$*, hence the ratio between these two probabilities is unbounded. However, note that the probability that $\mathcal{S}(1) > k/\varepsilon$ is exponentially small in $k$; hence, the overall probability that an adversary is able to distinguish between the two cases stays practically the same as in Example A.1. It is therefore only natural to still call this mechanism private.*

**Definition A.3 ($(\varepsilon, \delta)$-differential privacy [6])** *A mechanism $\mathcal{S}$ is said to be $(\varepsilon, \delta)$-differentially private if for all neighboring pairs of databases $\mathbf{x}, \mathbf{x}' \in \mathcal{D}^n$, and for all subsets of possible answers $\mathcal{V}$:*

$$\Pr[\mathcal{S}(\mathbf{x}) \in \mathcal{V}] \leq \Pr[\mathcal{S}(\mathbf{x}') \in \mathcal{V}] e^{\varepsilon} + \delta \ . \tag{4}$$

*The probability is taken over the coin tosses of the mechanism.*

## A.2 Privacy of Sets

While differential privacy is intended to capture the notion of individual privacy and furthermore is defined with respect to a change in a single entry, it would be somewhat disappointing to find out that it allows a change in, say, two or three entries to cause a massive change in output distribution. Fortunately, as we next show, this is not the case, but rather privacy of sets may deteriorate only linearly in the size of the set (for small sets and for small enough $\varepsilon$). Obviously, for an analysis to be meaningful, the distribution on the outputs must change with a change of many of the entries in the database the, thus privacy of sets must deteriorate, at least for large enough sets.

**Lemma A.4** *Let $\mathcal{S}$ be an $\varepsilon$-differentially private mechanism and let $\mathbf{x}, \mathbf{x}'$ be two databases such that $\mathrm{d}_H(\mathbf{x}, \mathbf{x}') = c$. Then*

$$\frac{\Pr[\mathcal{S}(\mathbf{x}) \in \mathcal{V}]}{\Pr[\mathcal{S}(\mathbf{x}') \in \mathcal{V}]} \leq e^{\varepsilon c} \ .$$

**Proof:** We prove the lemma by induction on $c$. For $c = 1$ it is simply the $\varepsilon$-differential privacy of $\mathcal{S}$. Assume correctness for $c$ and let $\mathbf{x}, \mathbf{x}'$ be two databases such that $\mathrm{d}_H(\mathbf{x}, \mathbf{x}') = c + 1$. There exists a database $\mathbf{x}''$ such that $\mathrm{d}_H(\mathbf{x}, \mathbf{x}'') = c$ and $\mathrm{d}_H(\mathbf{x}'', \mathbf{x}') = 1$. By Equation (1) and by the induction hypothesis, it follows that

$$\frac{\Pr[\mathcal{S}(\mathbf{x}) \in \mathcal{V}]}{\Pr[\mathcal{S}(\mathbf{x}') \in \mathcal{V}]} = \frac{\Pr[\mathcal{S}(\mathbf{x}) \in \mathcal{V}]}{\Pr[\mathcal{S}(\mathbf{x}'') \in \mathcal{V}]} \cdot \frac{\Pr[\mathcal{S}(\mathbf{x}'') \in \mathcal{V}]}{\Pr[\mathcal{S}(\mathbf{x}') \in \mathcal{V}]} \leq e^{\varepsilon c} e^{\varepsilon} = e^{\varepsilon(c+1)} \ .$$

□

### A.3 Composition

Another useful property of differential privacy is that even in the presence of an adaptive adversary privacy stays meaningful after $k$ rounds when $k\varepsilon$ is not too big (i.e., privacy degrades in a linear fashion, as long as $k$ and $\varepsilon$ are small enough), see [7] for the full argument. In this work we only use non-adaptive composition. Hence, the following lemma suffices for our needs (see [18] for a proof).

**Lemma A.5** *Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be $k$ mechanisms such that $\mathcal{S}_i$ is $\varepsilon_i$-differentially private for $1 \leq i \leq k$. The mechanism $\mathcal{S}$ that gives the answers of all $k$ mechanisms $\mathcal{S}_i$ (where the randomness of each mechanism is selected independently of the other mechanisms) is $\varepsilon'$-differentially private for $\varepsilon' = \sum_{i=1}^{k} \varepsilon_i$.*

## B The Attacker of [12]

There are a few attack methods described in [12], all of which follow a similar outline. Here we describe the first attack suggested by Korolova [12], which she refers to as Inference from Impressions, aimed at inferring information that a user entered on Facebook but has put into an "Only me" or "Friends Only" visibility mode. Following the notation of [12], we represent an advertising campaign as a mixture of conjunctions and disjunctions of boolean predicates. For example, campaign criteria $A = a_1 \wedge (a_2 \vee a_3)$ targets people who satisfy criteria $a_1$ (e.g. "Went to Harvard") and criteria $a_2$ (e.g. "Like skiing") or $a_3$ (e.g. "Like snowboarding").

**Input:** A user U and a feature $F$ whose value from the possible set of values $f_1, \ldots, f_k$ we would like to determine.

**Obtain initial auxiliary information:** Observe the profile information of U visible to the advertiser that can be used for targeting. Combine this information with the background knowledge on the user U, available to the attacker (implicitly, this auxiliary information includes some beliefs about all other Facebook users).

**Constructing campaigns:** Construct an ad campaign with targeting criteria $A$ combining auxiliary information about U and information visible in U's profile, so that one reasonably believes that only U matches the campaign criteria of $A$. The value of $F$ should not be specified in $A$.

1. Run $k$ advertising campaigns, $A_1, \ldots, A_k$, such that $A_i = A \wedge f_i$. Use identical and innocuous content in the title and text of all the ads. Specify a sufficiently high bids that guarantee with high probability that the ads would win an auction among other ads for which U is a match.
2. Observe the impressions received by the campaigns over a reasonable time period. If only one of the campaigns, say $A_j$, receives impressions, from a unique user, conclude that U satisfies $f_j$. Otherwise, refine campaign targeting criteria, bid, or ad content.

Figure 2: The Inference from Impressions Attack of [12].

It is instructive to note that the Inference from Impressions attack, described in Figure 2, does not require the user U to pay attention to the ad (and certainly not to click on it) in order for the attack to succeed in inferring the U's private information. It is enough that the user U connects to Facebook sufficiently often so that the ads have a chance to be displayed to U at least once over the observation time period.