TRANSACTIONS ON DATA PRIVACY 5 (2013) 35 — 67

Practicing Differential Privacy in Health Care: A Review

Fida K. Dankar*, and Khaled El Emam*†
* CHEO Research Institute, 401 Smyth Road, Ottawa, Ontario
E-mail fdankar@ehealthinformation.ca
†Pediatrics, Faculty of Medicine, University of Ottawa, Ottawa, Ontario
E-mail kelemam@ehealthinformation.ca

Abstract. Differential privacy has gained a lot of attention in recent years as a general model for the protection of personal information when used and disclosed for secondary purposes. It has also been proposed as an appropriate model for protecting health data. In this paper we review the current literature on differential privacy and highlight important general limitations to the model and the proposed mechanisms. We then examine some practical challenges to the application of differential privacy to health data. The most severe limitation is the theoretical nature of the privacy parameter ε. It has implications on our ability to quantify the level of anonymization that would be guaranteed to patients, as well as assessing responsibilities when a privacy breach occurs. The review concludes by identifying the areas that researchers and practitioners need to address to increase the adoption of differential privacy for health data.

1 Background

We consider private data analysis in the setting of a trusted data custodian that has a database consisting of rows of data (records). Each row represents information about an individual. The custodian needs to publish an *anonymized* version of the data, a version that is *useful* to data analysts and that protects the *privacy* of the individuals in the database. The problem is known as *privacy-preserving data publishing*.

Various approaches/models have been proposed in different research communities over the past few years [1]. These models depend on the background knowledge of adversaries; as these adversaries can infer sensitive information from the dataset by exploiting background information. Because of that de-

pendence, disclosed data are not immune to attacks resulting from unforeseen auxiliary information. It may not always be possible to define reasonable bounds on the background information that an adversary may have.

Differential privacy is a fairly new privacy model that is gaining popularity. Its appeal is that it makes *almost* no assumptions about the attacker's background knowledge. Differential privacy tries to ensure that the removal or addition of any record in the database does not change the outcome of any analysis *by much*. In other words, the presence of an individual is protected regardless of the attacker's background knowledge.

A previous attempt at a context-free privacy model was made by Dalenius in 1977. He articulated a privacy goal for statistical databases: *anything that can be learned from the database about a specific individual should be learned without access to the database* [2]. Attempts to formalize Dalenius' notion centered around measuring the adversary's prior and posterior beliefs about a specific record in the database, making sure that the change is small. This however contradicts the goal of database release, which is to change or form beliefs about people/situations. An example illustrating this is the Terry Gross height example from [3]:

> *"Suppose one's exact height were considered a highly sensitive piece of information, and that revealing the exact height of an individual were a privacy breach. Assume that a database yields the average heights of women of different nationalities. An adversary who has access to the statistical database and the auxiliary information 'Terry Gross is two inches shorter than the average Lithuanian women' learns Terry Gross's height, while anyone learning only the auxiliary information learns relatively little."*

According to Dalenius' notion of privacy, learning Terry Gross's height does constitute a privacy breach even if Terry Gross did not participate in the database. That motivated Dwork to come up with a different formulation, differential privacy: *"the risk to one's privacy…should not substantially increase as a result of participating in a statistical database"*[3]. Thus, an attacker should not be able to learn any information about any participant that she cannot learn if the participant opts out of the database. Going back to the example above, since Terry Gross's height can be learned without her participation in the database, it does not constitute a breach of privacy.

Research on differential privacy is growing rapidly and gaining popularity in the computer science literature. It is even becoming challenging to justify con-

ducting research on and making improvements to other privacy models that are already in wide use. That partially led to extreme opinions around the notion of differential privacy, from being completely inadequate and can never be used in real life situation, to being the best and only viable privacy notion. It is for these reasons that we believe it is beneficial to present an overview of the current state-of-the-art on differential privacy and examine its applicability to the disclosure of health data. In doing so, we identify the areas in differential privacy that need to be explored further before it can be applicable in the health field.

In private data analysis, two settings are possible: interactive and non-interactive. In the interactive setting where the database is held by a trusted server, users pose queries about the data, and the true answer to the queries is modified to protect the privacy of the database participants. In the non-interactive setting the data custodian either computes and publishes some statistics on the data, or releases an anonymized version of the raw data.

In the first part of this paper, differential privacy is formally defined along with some of its relaxations. Section 3 describes ways to achieve differential privacy for one query, Sections 4 and 5 discuss the challenges in setting the privacy parameter and present some relaxed versions for differential privacy. Then the leading research in the two data sanitization settings, interactive and non-interactive, is presented and evaluated in Section 6, followed in Sections 7 and 8 by some conclusions on the general applicability of differential privacy on health data.

2 Differential Privacy: The Principles

Generally speaking, differential privacy requires that the answer to any query be "probabilistically indistinguishable" with or without a particular row in the database. Precisely, given two databases that differ in exactly one row, a differentially private algorithm will provide randomized outputs that follow *almost identical* probability distributions on both databases. In other words, if an adversary possesses background information about all records in the dataset except one record v, s/he won't be able to infer the presence or absence of v with high probability. The closeness of the randomized outputs on both databases is determined by a privacy parameter ε. Higher values of ε imply weaker privacy guarantees. Formally:

Given an arbitrary query f with domain $\wp$ and range P ($f:\wp \to P$), and two databases D and D', drawn from population $\wp$, that differ in exactly one record, if K_f is a randomized function used to respond to query f, then K_f gives ε-differential privacy if for any $s \subseteq Range(K_f)$

$$\Pr[K_f(D) \in s] \leq e^{\varepsilon} \Pr[K_f(D') \in s]$$

In [4], the author refers to ε as the leakage. In [5], the authors interpret the ratio $\frac{\Pr[K_f(D) \in s]}{\Pr[K_f(D') \in s]}$ as: "knowledge gain ratio from one data set over the other". Differential privacy requires that the *knowledge gain* be bounded by e^{ε}. Even if the participant removed her data from the database, the limited knowledge gain implies that *no output* would become significantly more or less likely [6]. The parameter ε is public. Values typically chosen in the literature are 0.01, or 0.1, and sometimes *ln2* or *ln3* [3], [6].

3 Achieving Differential Privacy

The notion of differential privacy has led to the introduction of several mechanisms that preserve it. The development of further mechanisms is an active area of research. In this section we explain how Laplace noise can be used to provide a differentially private output for one query. Then, in Section 6 we outline some of the leading mechanisms in the differential privacy literature.

In order to satisfy differential privacy for a query output, the authors in [4] suggests the use of Laplace distributed noise: if r is the correct answer to a query f on a database D (i.e., $f:\wp \to P$, $D \in \wp$ and $f(D)=r$), then a differentially private mechanism would output the response $r+y$, where y is the noise drawn at random from a Laplace distribution with mean 0 and scale $\Delta f / \varepsilon$. Δf represents the maximum value for $\| f(D') - f(D'') \|$ for all $D'', D' \in \wp$ differing in one row ($\Delta f = \max_{D',D''} \| f(D') - f(D'') \|$, for all $D'', D' \in \wp$ differing in one row).

Δf is the global sensitivity of the query f over domain $\wp$, it is independent of the database D, and one must consider all databases $D' \in \wp$ to determine its value.

Note that this is unlike the traditional noise addition where the amount of noise is proportional to the variance of the data. That shows a fundamental characteristic of differential privacy which is that it seeks to protect even extreme values; note that if the data is skewed, this results in high sensitivity, and

the noise would be large compared to the variance of the data, leading to potentially non-useful output [7].

The above mechanism cannot be applied to a query whose outcome is not real, as adding noise to non-real values does not make sense. In [8], the authors propose a novel mechanism that could be applied to queries with any outcome: the exponential mechanism. The authors assume the existence of a score function that evaluates the quality of the query output (the score function could be the utility of the output): Given a query f with range P (the range P could consist of nominal, categorical or integer values), the mechanism assigns probabilities to the different elements in P based on their scores, a higher score means a more desired output and hence a higher probability. If such a score function exists, then a differentially private output could be produced based on the sensitivity of the score function. The authors argue that, similar to the Laplace mechanism, if the sensitivity of the score function is low, then high quality output can be obtained. Alternative mechanisms that satisfy differential privacy have recently surfaced. These will be discussed in Section 6.

4 Setting ε

Setting a value for ε is not an easy task as it is not adequately covered in the differential privacy literature. It is difficult for ordinary data holders to measure the privacy protection of a dataset provided from a specific ε value. For example, if a user wants to know the number of smokers who experienced heart attacks in a given dataset, then what does a value of $\varepsilon = 0.01$ mean? What does it mean to have an information gain of $e^{0.01}$? Is this enough to protect the presence (and hence identifiability) of smokers with heart attacks in the dataset? Does the value of ε depend on the query, or on the universe of values?

In fact, there exists no experimental evaluation to guide the user on choosing an appropriate ε value. Dwork considers the issue as *"a social question and … beyond the scope"* of her work [6], however she provides some recommended values of: 0.01, or 0.1, and sometimes *ln2*, and *ln3* [3], [6]. Some papers used values such as : 0.5, 0.75, 1, 1.25, 1.5, 2 [9]–[11], and even 3, 4, 5, 6, 7, 8, 9,and 10 [12].

In a recent attempt at finding general guidelines for setting appropriate ε values, the authors in [13] found that ε cannot be defined in general but will always depend on the dataset in question. Similarly, the authors in [14], [15], state that, given an ε value, the probability of re-identification is not fixed, it rather

depends on data values in the data set and even on data for individuals outside the data set.

In [13], the authors concluded that one needs to define a "practical privacy standard" to be able to determine an appropriate value for ε. The privacy standard they use is "the risk of identifying an individual in the database". A value for ε is considered inappropriate if the risk of "revealing the presence of an individual" is higher than acceptable.

5 Variants of Differential Privacy

5.1 Relaxations of Differential Privacy

A relaxed version of differential privacy, described in [16], (ε,δ)*-differential privacy* requires that ε-differential privacy be satisfied with a probability of at least $1-\delta$, in other words, ε-differential privacy can be violated for some tuples, however the probability of that occurring is bounded by δ. Formally:

For any two database, D and D', drawn from a population $\wp$, that differ in exactly one record, if K_f is a randomized function used to respond to an arbitrary query f, then K_f gives (ε,δ)-differential privacy if with probability $1-\delta$, for any $s \subseteq Range(K_f)$ the following holds:

$$\Pr[K_f(D) \in s] \leq e^{\varepsilon} \Pr[K_f(D') \in s]$$

An alternative and more popular relaxation is *approximate differential privacy* or (ε,τ)*-differential privacy*, it requires that for any $s \subseteq Range(K_f)$

$$\Pr[K_f(D) \in s] \leq e^{\varepsilon} \Pr[K_f(D') \in s] + \tau$$

There are no rules provided for setting a value for τ. The decision is usually left to the data custodian, however τ is usually given a value less than 10^{-4} [17].

(ε,τ)-differential privacy is similar to ε-differential privacy but allows an additive error factor of τ [17]. They both can be satisfied by adding random noise to query answers. However they differ in the noise distribution as well as the sensitivity measurement [18].

5.2 Measuring Utility

There is no special utility definition that is adopted in the differential privacy literature. Some of the popular approaches include (α,∂)-usefulness [19], [20], relative error [10], absolute error [21], [22], error variance [10], [22], and Euclide-

an distance [23], [24]. For Euclidean distance, the utility of the mechanism is the expected Euclidean distance between the actual query answer and the outputted one. For the relative error, the utility is the ratio of the distance between the actual and outputted query answers to the actual query answer: $\frac{|K_f(D)-f(D)|}{f(D)}$. For the error variance, the utility is the variance of the Laplace mechanism: $\frac{2}{\varepsilon^2}$. More information on these and other utility measures can be found in [25], [26]. In what follows, we define (α,∂)-usefulness more precisely as it is popular in the context of differential privacy, and effective in the overall estimation of utility [19], [20], [27]:

A mechanism is (α,∂)-*useful* if every query output is within α of the correct output with a probability of at least $1-\partial$, i.e.:

$$\Pr[|K_f(D)-f(D)|\leq\alpha]\geq 1-\partial .$$

In the following, the term utility refers to (α,∂)-usefulness.

5.3 Counting and Predicate Queries

A basic class of queries is *Counting queries*. They allow the researcher to count the number of individuals in the database that satisfy a certain predicate (example: how many individuals have grey hair and are under 40 years?). Formally, given a predicate Φ, where $\Phi(v)\in\{0,1\}$ for all records $v\in\wp$, a query $f:\wp\to$ with $f(D)=\sum_{v\in D}\Phi(v)$ is a counting query. It follows that, for every counting query f, the sensitivity of f is 1. Multiple counting queries $f_1,...,f_h$ can be represented using one query F as follows: $F(D)=(f_1(D),...,f_h(D))$. In such case, the sensitivity of F would be $\Delta F\leq\sum_{i=1}^{h}\Delta f_i=h$ [26].

Another class of commonly used queries is *Predicate queries* (also referred to as statistical queries), these are normalized counting queries, in other words they count the fraction of individuals in the database that satisfy a certain predicate. Note that the sensitivity of one predicate query is $1/n$ when the database is of size n, and the sensitivity of h predicate queries is $\leq\frac{h}{n}$ [23], [26].

6 Mechanisms of Differential Privacy

6.1 Interactive Differential Privacy

6.1.1 Mechanisms

Before presenting the interactive mechanisms, we present a negative result concerning privacy from noise addition:

Result 1: Dinur and Nissim [28] showed that in a database consisting of bits, if a user wants to know the sum of some random subsets of these bits, then no private mechanism can provide accurate answers for many queries. In fact, no private mechanism can answer k queries with error $O(\sqrt{k})$, because an adversary would be able to use the queries' outputs to reconstruct $1-O(1)$ fraction of the database, a condition referred to as "blatant non-privacy".

Dwork introduced the first and most commonly used mechanism that provides ε-differential privacy in query-response situations, it is Dwork's independent Laplace mechanism [4]. In this mechanism, the data custodian adds appropriate amounts of noise to each query answer as follows:

Noise is drawn at random from a Laplace distribution with mean 0 and scale $\Delta f / \varepsilon$ and added independently to each query response (Section 3), thus, making sure that every query is perturbed appropriately. However since each query answer leaks more information and reduces privacy, what is the privacy level of several query outputs taken together? Dwork proved that the mechanism is composable, i.e., given a sequence of differentially private computations: $f_1,...,f_k$ assigned budgets privacy $\varepsilon_1,...,\varepsilon_k$ respectively (i.e., the noise added to the output of f_i is drawn at random from a Laplace distribution with mean 0 and scale $\Delta f / \varepsilon_i$), then the overall computation has a worst case parameter of $\sum_1^k \varepsilon_i$ [4].

Given its composability, two standard approaches exist for the above mechanism depending on whether the queries are known ahead of time or not:

If the queries $f_1,...,f_k$ are known ahead of time, then the standard approach is for the data custodian to assign a total privacy budget ε, the user then can divide the privacy budget among the different queries as needed (more important queries can have higher budget and hence be less noisy). However it is important to note that the type of analysis (queries) that is of interest to the analyst

is not usually known ahead of time, in fact sanitized data are usually shared with several data analysts with different goals.

On the other hand, if the queries are chosen adaptively, i.e., if each query is formulated after obtaining the result of the previous query/queries (which is a more realistic scenario), then the approach is for the data custodian to have a preset privacy budget ε. That budget will decrease every time a query is answered, and the custodian will keep on providing query answers until the budget runs out. Since different users of the database may collude, there should be one privacy budget for all users [3], [6]. Once the budget runs out, the database has to shut down.

Now, assume that each query $f_i, i \in \{1,..,k\}$ has sensitivity Δ, such that $\Delta \leq 1$, and that we want to assign a budget ε_i to each query f_i. If a fixed utility is required (supplied by fixed parameters α, ∂), then this sets a bound on the magnitude of noise allowed per query (Laplace($\frac{\Delta}{\varepsilon_i}$)), which in turn sets a bound on the value $\varepsilon = \sum_{1}^{k} \varepsilon_i$. Given the utility parameters α, ∂, this bound shows the interdependence between the two quantities: ε and k. In Dwork's independent Laplace mechanism this interdependence is as follows:

Result 2: Assume that the user is interested in predicate queries $f_1,..., f_k$ (or any set of low sensitivity queries such as sum or linear queries), and in having (α, ∂)-usefulness, if Laplace noise is added independently to every query output (Dwork's mechanism), then the privacy parameter ε (and hence the noise magnitude) grows linearly with the number of queries k answered [4].

The result implies that in order to have sublinear noise (noise to the order $O(n^c)$ with $c < 1$) the mechanism cannot answer $O(n)$ queries.

The mechanism above has bad implications on several fronts as described in [11]: When the number of users is large, the budget per user becomes limiting as lower budget requires larger noise to be added. Moreover, the issue of dividing this budget among users presents another difficulty. Even when a small number of users is expected, the budget can be limiting to creative research (for example, data mining that is not driven by testing pre-defined hypotheses). That drove research in the interactive setting to concentrate on ways to reduce the noise magnitude per query, thus allowing more queries per budget:

A line of research by Blum et al [19] tried to circumvent Result 1 by creating a mechanism that can answer arbitrary many queries while preserving differential privacy. The catch is that, the queries whose output is *useful* belong to a predefined "class of queries":

Let R be a discrete set of data items, let D be a database of size n whose rows are drawn from R ($D \in R^n$), now let C be a "concept" class, that is a set of computable functions from $R \to \{0,1\}$. The authors use the exponential mechanism [8] to create a synthetic database D' that approximates outputs corresponding to all concepts in C. In fact, for fixed usefulness parameters, the privacy parameter ε grows proportional to the VC-dimension of C, a measure of the complexity of C, which is approximately $\log_2 |C|$.

Result 3: If $|C| = k$, then for fixed usefulness parameters, the privacy parameter ε grows linearly with $\log_2 k$.

Note that the database D' can be released, or can be used interactively to answer queries.

Result 3 implies that utility can be guaranteed even for an exponential number of queries, however the mechanism works only for discrete databases (R was assumed to be discrete), generalizing it to non-discrete databases comes at the expense of its usefulness [19], moreover it is non-efficient. In fact, it is superpolynomial in the parameters of the problem (i.e., it is not bounded above by any polynomial), making it impossible for practical applications. Another drawback is that the mechanism solves the problem for the interactive database release only when the type of analysis that is of interest to the analyst is known ahead of time (set C should be known before D' can be generated).

In an attempt to overcome some of the above limitations, Dwork et al [29] presented a mechanism that is (ε, τ)-differentially private but with an improved running time (polynomial in $|R|$ and k). The mechanism's utility is better than the independent Laplace mechanism: for fixed usefulness parameters, the privacy parameter ε grows linearly with $O(\sqrt{n}2^{\sqrt{\log_2 k}})$. However, the result applies only in the non-adaptive case, i.e., the actual queries need to be known ahead of time.

Similarly, in [23] and [30] the authors addressed the issue of reducing noise and allowing more queries per a given budget, however the techniques require that all queries be known ahead of time.

In [17], the authors present a novel non-efficient mechanism, *the median mechanism*, that interactively answers k predicate queries $f_1,...,f_k$ as they arrive while guaranteeing (ε,τ)-differential privacy. Their aim was to ameliorate the limitation on the number of predicate queries that can be answered in an interactive setting. The mechanism does that by determining the correlation between different queries on the fly. In essence, the authors prove that queries can be divided into hard and easy queries, where hard queries completely determine the answers to the easy queries (up to a small error). In fact, they proved that, given a set of queries $f_1,...,f_{i-1}$ and their outputs $r_1,...,r_{i-1}$ and given an easy query f_i, the majority of databases that produce consistent results on the previous query answers $r_1,...,r_{i-1}$, would answer the current easy query f_i accurately, thus, the output of f_i would be deducible from $r_1,...,r_{i-1}$, and hence f_i would not use the privacy budget. Moreover, they proved that among any set of k queries, there are $O(\log k \log |X|)$ hard queries and the rest are easy queries. The hard queries are answered using Dwork's independent Laplace perturbations, while the easy queries are answered using the median result from databases that are consistent with previous answers. The classification of queries into hard and easy is done at a low privacy cost. The authors proved the following result:

Result 4: For fixed usefulness parameters, the privacy parameter ε, grows linearly with $O(\log_2 k \log_2 |X|)$ where X is the database domain.

That means that the median mechanism can answer exponentially more predicate queries than Dwork's independent Laplace mechanism. More importantly, the privacy and utility guarantees hold even when the queries are chosen adaptively. However the setback is that the algorithm is non-efficient, its run time is exponential in the problem's parameters, and is hence unusable in practice. For more information, the reader is referred to [17].

Recently, the authors in [31] presented a new *efficient* mechanism, *the multiplicative weights mechanism*, it achieves ε-differential privacy (rather than (ε,τ)) and for fixed usefulness parameters, the privacy parameter ε grows linearly with $O(\sqrt{k})$. In other words the algorithm achieves $O(\sqrt{k})$ noise per query, while the independent Laplace mechanism achieves $O(k)$. Moreover it is efficient and works for adaptively chosen *predicate* queries.

6.1.2 Discussion

In the interactive setting, the challenge is to find a differentially private mechanism that can (a) answer random queries (any type of queries), (b) answer a large number of queries while providing non-trivial utility for each of the queries, (c) be efficient, (d) achieve ε-differential privacy (the stronger privacy guarantee), and (e) answer queries adaptively. As discussed in the previous sub-section, many algorithms attempted to achieve some of the above requirements only to fail in others.

Dwork's independent Laplace mechanism [4] achieves requirements (a), (c), (d) and (e), Blum at al presented an elegant mechanism in [19], that achieves (b) and (d), Roth and Routhgarden presented the median mechanism [17] that achieves (b), (d) and (e), and finally, Hardt and Rothblum presented the multiplicative weights mechanism [31] that achieves (b), (c), (d) and (e).

Therefore, the result in [17], [19] and especially [31] are striking in their generality, and succeeded considerably in relaxing the limit on the number of queries per database. However their applicability is limited to counting/predicate queries and only [31] can be used in general ([19] and [17] can only be used with small size problems). On the other hand, the issue of assigning a budget for a database (i.e., choosing a value for ε) and the issue of dividing this budget among the users was not tackled in any of these papers and still presents another challenge in practical problems [11].

6.2 Non-Interactive Differential Privacy

Result 1 stated that no private mechanism can provide accurate answers for many queries. This result has bad implications on the non-interactive approach, because, contrary to expectations, any data release would provide accurate responses only to a class of *pre-defined* queries.

In differential privacy, non-Interactive noise-based mechanisms either:

- Release a new synthetic dataset that is composed of synthetic individuals whose distribution mimics that of the original participants for some pre-defined query set, or
- Release a perturbed version of the original dataset, usually in the form of a contingency table.

In this sub-section, we survey the main available mechanisms for non-interactive release. The mechanisms are divided into ones that rely *solely* on noise addition, and the more recent ones that do not.

6.2.1 Noise-Based Mechanisms

As mentioned in sub-Section 6.1, Blum, Ligett and Roth [19] studied database releases from a learning theory perspective. Given a database D and a class of concepts C (as defined in 5.1), the authors used the exponential mechanism to release a synthetic database that is useful for all queries in class C. However, the mechanism works only for discrete databases, is non-efficient, and requires that the class C of queries be known ahead of time.

The author in [6] studied the problem of histogram release while satisfying differential privacy. The histogram to be released is treated as the output of a specific query with sensitivity 1, the problem is the following:

Assume we have a database with n rows each describing one individual. Assume also that each row consists of k binary attributes: $a_1,...,a_k$. Then the database can be transformed into a contingency table, also known as a table of counts or histogram. Contingency tables describe, for each setting of the k attributes (which amounts to 2^k settings), the number of rows satisfying this setting. Table 2 shows the contingency table generated from the example shown in Table 1.

Table 1. Medical Records

	Age	**Gender**	**HIV**
1	20	M	+
2	30	F	-
3	20	M	+

Table 2. Contingency Table

20,M,-	20,M,+	20,F,-	20,F,+	30,M,-	30,M,+	30,F,-	30,F,+
0	2	0	0	1	0	0	0

Table 3. α – **Marginal table,** $\alpha = \{age, HIV\}$

20,+	20,-	30,+	30,-
2	0	0	1

The author shows how to add noise to each entry of this contingency table to make it differentially private [6].

Since the contingency table is a histogram, its sensitivity is 1 (the addition or removal of any database row affects the counts in one location by at most 1). Hence we can add independently generated noise to the 2^k cells of the contingency table with distribution $Laplace(1/\varepsilon)$.

The problem with this approach is that the amount of noise generated when computing marginals, i.e., queries that involve a large number of entries, (refer to Table 3) could be large [6],[32]. For example, the variance of the 1-way table described by any one attribute, is $2^{k-1}\varepsilon^{-2}$. This is considered unacceptable especially when $n << 2^k$ [6]. In other words, for a count query that involves the sum of a fraction of the entries in the noisy contingency table, the noise variance could be around $O(2^k)$. If the attributes $a_1,...,a_k$, are not binary and have domains of sizes $\omega_1,\omega_2,...,\omega_k$ respectively, then the contingency table size would be $\omega_1 \times \omega_2 \times ... \times \omega_k$ which could be huge as databases often have several attributes with large domains.

Alternatively, in [6], the author suggests that if low order marginals are sufficient for data analysts (these are smaller tables of count projected to a subset of the attributes, Table 3 shows an example of the marginal for the two attributes: age and HIV), then the data custodian can release a set of say C marginals. Each marginal has sensitivity 1, hence the amount of noise added to each cell of the released marginals should be proportional to $|C|/\varepsilon$. When n is large compared to $|C|/\varepsilon$, the authors suggest that this will yield good accuracy *per cell*. However, with this approach, there will be inconsistencies between the different marginals as noise is added independently.

Another (non-efficient) alternative introduced by [33] is to construct a synthetic database that is positive and integral and that preserves all low order marginals up to a small error using Fourier transforms.

In [34], the authors evaluate the mechanism of [33] using three real-life databases. They concluded that the method is not suitable for the large and sparse

tables that are usually produced by large data custodians. Moreover, the authors stated a preference for methods that use log linear models such as that of [35]–[37].

6.2.2 Alternative mechanisms

Existing techniques that satisfy differential privacy rely almost exclusively on noise addition to query outputs through the usage of Laplace noise or the exponential mechanism. However, few alternatives that rely on other means in addition to noise have recently surfaced:

In [38] the authors introduced a new efficient mechanism to release *set value data* while satisfying ε-differential privacy and (α,∂)-usefulness *for counting queries* (where the value of ∂ is a function of ε and α). Set-value data consists of a set of records; each record consists of a set of items drawn from a universe of items. An example is transaction data. The authors introduced a probabilistic top down partitioning algorithm. The algorithm starts by creating a context-free taxonomy tree and by generalizing all items under one partition. The algorithm then, recursively, generates distinct sub-partitions based on the taxonomy tree. The sub partitions have more specific representations, thus, splitting the records into more specific disjoint sets. The choice of which partition to split is based on the noisy partitions counts. Only (probabilistically) non-empty partitions are chosen, thus, limiting the overall noise added and making this approach appealing when data is sparse. The authors experimentally evaluated their approach against the histogram approach presented earlier [6] and showed that it provides better utility for counting queries. However the authors did not justify their utility parameters, and did not study the interdependence between the utility and privacy parameters. Moreover, the mechanism in [6] suffers from high noise variance for set value data (being sparse and high dimensional) and that renders its results of limited usefulness, and is therefore not a good standard for comparison.

In [32], the authors argue that contingency table release is not suitable for data with high dimensions and large domains as the noise would be relatively large compared to the cell counts, and that leads to a total destruction of utility. The authors then present a novel method that generalizes the contingency table then adds noise to the counts. The authors argue that the generalization step increases the cell counts and thus, the counts become much larger than the noise added. The algorithm initially generalizes all records into one group, then iteratively implements a sequence of specializations. At each iteration, the algorithm

probabilistically selects an attribute to specialize based on some score value (such as information gain). The algorithm terminates after a preset number of specializations. The authors performed a comparison between their algorithm and the top-down specialization approach of [39] that produces k-anonymous data. The comparison was done for a fixed value of k=5, and for ε values 0.75, 1, 2, 3, and 4. They deduced that their algorithm performs slightly better for ε values 0.75 and 1, and noticeably better for the remaining ε values. However the authors did not justify their choice for these ε values, and why these particular values should be compared to a 5-anonymity (why is any of these ε values comparable to 5-anonymity?). Additionally although the authors state that a typical ε value is less than 1 -in fact the recommended ε value is 0.01, or 0.1 [6], [12]- they did not use these recommended values in their evaluation. Moreover, the authors did not explain how to decide on an optimal generalization that would produce counts that are high enough relative to the added noise given the privacy budget ε.

Recently, the authors in [11] introduced a novel technique connecting k-anonymity to differential privacy. The technique tries to achieve (ε,δ)-differential privacy by adding a random sampling step followed by a "safe" k-anonymization. The authors show that when a random sampling step is applied to the data followed by a generalization scheme that does not depend on the database (like applying a fixed generalization scheme) followed by the suppression of all tuples that occur less than k times, then the database satisfies (ε,δ)-differential privacy.

The problem with this technique is that the usage of a data independent generalization may result in poor utility, in [40] the authors argue that adding a random sampling step impacts the utility of the data as well.

Recently, several studies on the application of differential privacy have emerged. However these studies concentrate on particular kinds of data such as search logs [41], [42], on specific applications such as recommender systems [43], private networks [44], and record linkage [45], or on certain type of low sensitivity queries (queries with sensitivity ≤ 1, such as counting and statistical queries) [10], [20], [46].

6.2.3 Discussion

In the non-interactive setting, researchers expect the data release to answer *all* their queries accurately. Moreover, they prefer non-synthetic data; in fact researchers and statisticians in several fields, such as the health sector, like to

"look at the data". Dwork mentions in [47] that "conversations with experts in this field involve pleas for a noisy table that will permit highly accurate answers to be derived for computations that are not specified at the outset". Hence, in general, the challenge is to release a non-synthetic dataset that can accurately answer any query, and the release mechanism needs to be efficient.

However, no differentially private interactive mechanism can achieve all of the above:

- In [19], [29], [31] the release can be used for a set of predicate queries that should be known in advance. The data released is synthetic, and contrary to the intuition on non-interactive data, the utility of the data suffers with the size of the query set.
- In [6], [33], the release can be used for a set of count queries, and the utility suffers when databases are large and sparse [36].
- In several other mechanisms, usability is only guaranteed for particular data types along with particular set of queries such as count queries on set value data [38].

In fact Result 1 indicates that no efficient algorithm can output a noisy table with acceptable utility to *random queries*. This result clearly states that the general applicability of differential privacy will always be limited to a class of queries. Current research presented big advances in the mechanism efficiency and in the size of the query set with utilizable outputs. However, applications remain limited to queries with low sensitivity such as count and predicate queries.

On the other hand, it is important to note that, contrary to most mechanisms that output noisy answers to submitted queries, the non-interactive mechanisms (whether releasing synthetic data or a contingency table) guarantee consistency, meaning that the answers to correlated queries will not be contradictory. For example, two complementary counting queries will add up to the number of participants in the published dataset.

6.3 Limitations of Differential Privacy

In line with the previous discussions, the literature has concentrated on the application of differential privacy on count data. In fact, the options available in differential privacy limit the user to a number of queries with low sensitivity. Low sensitivity means that the output of these queries is not "severely affected" by the removal or addition of any one record from/to the dataset (usually low sensitivity implies $\Delta f \leq 1$)

What about other queries? Recent articles [48] and [7] raise serious concerns about the utility of differentially private algorithms when applied to numeric data. They presented several examples on the application of Laplace noise for numeric data and showed that the level of noise added can be so large making the responses useless. In fact the level of noise is directly affected by the value of Δf and is independent of the actual database. Therefore, when the data is skewed, this will result in a large Δf value and subsequently a large noise variance (compared to the actual variance of the database). The authors concluded that "Like Dalenius definition of privacy before it, differential privacy is an interesting concept, but of little value in practice"[48].

The limited available experiments on differentially private mechanisms suggest that in general - for numerical data - very low privacy is required to have acceptable utility [3], [49]. In [50], the author suggests that the current formulation of differential privacy focuses on achieving privacy rather than on preserving utility, and, as mentioned before, that is due to the fact that differential privacy seeks to protect even extreme values in the database domain (through Δf).

Another limitation was mentioned by Muralidhar and Sarathy [7], [48], [51]; the authors stressed the difficulty in calculating queries' sensitivity in unbounded domains. That raises a question about not only the sensitivity calculation but about differential privacy verification as well. These two questions have been neglected in the literature, as illustrations are limited to queries with low and domain-independent sensitivities.

Another set of severe limitations is related to setting and dispensing the privacy budget, precisely, deciding on a value for the total budget ε, deciding on the number of users that will be allowed to query the dataset, and deciding on the budget portion provided to each user:

- As mentioned before, the literature on the privacy parameter ε is mostly theoretical, and there exists no experimental evaluations to guide the data holder on choosing the right ε value, or to help him/her understand what effect changing ε will have on the common notions of data utility. In [52], the author complains that their research team is running against the hurdle of choosing a good value for ε and quantifying the "*indistinguishability*" language in the differential privacy guarantee.
- Assuming that the data holder was able to overcome the hurdles of deciding on a total budget ε, then deciding on the number of users that will be allowed to query the dataset and on the budget per user is an-

other problem as interest in the dataset is not always predictable, and not all data uses require the same utility level.

- If the data holder was able to predict the number of potential users and to divide the budget among the different users (referred to as local budgets), then s/he still needs to decide on a strategy to allocate each local budget to the user's queries. A common approach is to allow each user to select their queries' budgets. In such a case, they would be able to give higher budgets to queries deemed more important. In these cases, the data custodian will keep on providing query answers until the budget runs out. However, (as iterated in Section 4), setting the right budget for every query is not an easy task, as the correlation between intuitive privacy notions and ε is not clear. Moreover, queries are not known to the user ahead of time (the formulation of new queries usually depends on the output of previous ones) hence, deciding on a budget for a current query with uncertainty about future queries is challenging. Another approach that avoids the management of local privacy budgets by users is the batch query answering mechanism presented in [18]. The set of queries are all submitted together as a batch and the mechanism adapts to the particular set of submitted queries and releases outputs that minimizes the square errors for these queries. The problem with such a strategy is that the user needs to know all the queries ahead of time.

  Another approach that deals with local budget decomposition was presented in [53]. It discusses the problem in the field of spatial data. Typical queries in spatial applications are count queries (the queries request the number of individuals that fall within a given spatial region). Given a total budget ε, the authors present a way to publish a spatial decomposition of the data using non-uniform noise parameters. To elaborate, a spatial decomposition is a hierarchical decomposition of a space into smaller areas. Building such a hierarchy privately requires knowledge of data counts at each node (i.e., the number of data points associated with the space that the node represents). Therefore, the spatial decomposition can be translated into a set of interrelated count queries, one count query for every node in the decomposition. The authors calculate the maximum number of count queries needed to perform the decomposition. Then, they describe the best way to allocate the total budget ε among these different queries in a way to achieve

the best utility possible. The problem with this budget decomposition is that it is tailored for this particular application. Moreover, it is utility driven and does not discuss the privacy implicated by each parameter (ε_i) chosen.

Another major limitation was reported by [54]. The paper pointed out that differential privacy works under the assumptions that individuals are independent from each other. If this independence assumption is violated, then differential privacy cannot adequately limit inference about individual participation in the data set. The authors in [54] illustrate with an example from social networks where the removal of one participant (record) might not hide evidence of his/her participation: "Bob's participation in a social network can cause links to form between pairs of Bob's friends" and thus, Bob's participation "affects more than just the tuple marked: Bob". In such cases, we are forced to take into consideration the adversary's background knowledge. Hence, for datasets with correlated records, a stronger privacy notion is required. This limitation comes in contrast with the highly valued property of differential privacy: its independence from the attackers background knowledge.

Finally, most existing differential privacy techniques assume the existence of one dataset with one data holder. To the best of our knowledge, [55]–[57] are the only papers to consider the problem of privately calculating the sum of numerical data held by different entities. Since aggregators wish to compute various statistics on data, and since studies can require data from physically distributed sources owned by different data holders, new techniques need to be developed to deal with this challenging aspect.

7 Differential Privacy and Health Data

In what follows, we focus on health data release, and present the challenges that differential privacy mechanisms must overcome before they can be widely adopted for health information release. Then we discuss a recent application of differential privacy on the health data.

7.1 Challenges

The disclosure of health data has a number of characteristics that need to be considered in any practical mechanism used to preserve privacy. These characteristics have to do with current practices and data sets, and the introduction of

any new mechanism for privacy protective data disclosure or analysis would have to address these issues before it can be adopted widely. The considerations below are driven by our experiences creating health data sets for secondary purposes over the last seven years, some of which have been documented, as well as empirical studies of health data sharing practices and challenges [58]–[68].

Health data contains categorical data (e.g., diagnosis codes, procedure codes, drugs dispensed, laboratory tests ordered, and geographical information about the patient and the provider), as well as numeric data (e.g., age in years, length of stay in hospital, and time since last visit). Therefore both types of variables need to be addressed. As noted earlier, the addition of Laplace noise to numeric data can distort the values significantly.

Users of health data are accustomed to data publishing – which is where the data is disclosed to the end user. There are multiple reasons. Health data is often messy, with data errors and sometimes unexpected distributions. Analysts need to look at the data to determine the appropriate transformations to apply, compute appropriate indices from the original data, and extract the appropriate cohort of patients for the analysis. This is easiest to do when one has access to the data directly. Furthermore, biostatisticians and epidemiologists will often have a suite of analysis methods and tools that are commonly used, that they have used for many years, that they understand, for which they can interpret the results correctly, that are understood and accepted within the community as valid for that type of problem and data, and for which they have code in languages such as SAS that they use. From a behavior change perspective, it would be challenging to convince data analysts to abandon their current methods and SAS code, which they may have been using for decades, in favor of an interactive system that is less understood. Therefore, at least in the short term, a non-interactive mechanism would be most suitable for this community.

The non-interactive mechanism that allows the computation of statistics without publishing the data may be suitable in some circumstances. For example, in the context of public health, on-going surveillance often relies on the computation of a well defined (and known a priori) set of statistics at regular intervals. Similarly, performance or safety reporting (e.g., the number of eligible patients that received appropriate screening and the rate of surgical site infections) involves the computation of well defined statistics. Therefore, for surveillance and performance reporting purposes differentially private statistics would be con-

gruent with the process. In such cases the primary consideration would be data utility of the differentially private statistics.

Beyond such applications, with well defined analytical needs, many other data uses would require actual data publishing to meet the needs of the analyst community.

Another important consideration is the law. The healthcare sector often has specific privacy laws in many jurisdictions. Current health privacy statutes in the US, Canada, and Europe do not specify the acceptable risk and often use the "reasonableness" standard. In practice, one relies on precedent to justify the risk thresholds that are used. For currently used privacy models, such as *k*-anonymity, there is a significant amount of precedent for different values of *k*. Data custodians have been releasing data for more than two decades, including health data. During that period guidelines, policies, court cases, and regulatory orders have come out which define what can be considered acceptable levels of risk. A data custodian, if confronted in a court case or by a regulator for example, can point to these precedents to justify their decisions. In the case of differential privacy, important parameters such as ε have no intrinsic meaning and there are few existing precedents of actual health data releases to justify the choice of any value. A data custodian needs to consider how they would justify their choice in a dispute or litigation, and it is much easier to do so under current models, such as k-anonymity, and risky (financially and reputationally) under differential privacy.

Many fields in health data sets are correlated or have natural constraints. For example, one treatment would often precede another, or certain drugs are given in combination (or never given in combination). There are correlations among drugs, and diagnoses, and between lab results and diagnoses. Independent distortions to the data that produce results that do not make sense erode the trust of the data analysts in the data and act as barriers to the acceptability of the techniques used to protect the privacy of the data. For example, if the distorted data shows two drugs that are known to interact in a way that can be damaging to a patient's health, or a drug that would never be prescribed with a particular treatment appear for the same patient, or a dose that does not make sense for a patient, then the analysts will cease to trust the data. In practice this has created challenges for introducing mechanisms that add noise to data because it is not possible to guarantee that nonsense data cannot be output.

Reference deployments of differential privacy in practice are also important. A powerful argument in convincing data analysts to use a data set that has been

transformed in some way to deal with privacy concerns is to show them actual examples where such data has produced useful and valid results. To our knowledge, thus far, there have been limited real world disclosures of differentially private health data, and consequently few examples of useful and valid analytical results.

While not often explicitly considered when designing new mechanisms to protect data, convincing the public that stewardship of their data is being conducted in a responsible way is becoming a necessary objective. For instance, patients and providers have expressed concerns about the disclosure and use of health information, and there is evidence that patients adopt privacy protective behaviors when they have concerns about how their own information is being used or disclosed, especially among vulnerable patient groups [69]. In practice this means there is an on-going need to explain in non-specialist terms the parameters of the privacy mechanisms used and how much protection they really provide. In the context of differential privacy, it is quite challenging to explain to a patient the meaning of the ε value used to disclose or provide analytical access to their data, for example. It is necessary to relate these parameters to more common notions of privacy to allow easier communication to the public.

While the above observations are limited by our experiences, we believe they represent real challenges that the differential privacy model and mechanisms need to address to ensure wider acceptability and adoption within the health domain.

7.2 Example of Health Application

In what follows, we present a recent application of differential privacy in the health sector. This is the only health care application of differential privacy that we are aware of. We will use it to illustrate the strengths and weaknesses that were discussed in the previous subsection.

Cohort identification is commonly used in clinical datasets, it involves querying the patient database to identify potential recruits for a clinical trial. Recruiting a sufficient number of subjects in clinical trials is becoming increasingly challenging, resulting in underpowered studies. This has been noted as a factor in the inability to complete trials successfully [70]. Possible causes are the marked increase in the total number of eligibility criteria and the rapid rise in the number of subjects that need to be recruited [70], [71]. Consequently, the number of trials moving to regions outside Canada and the US because of weak subject recruitment has been growing [70], [72]–[74].

One common way to identify potential recruits for a trial is to query a patient database [71]. An investigator would run a query on such a database at a recruiting site to identify the number of participants which would meet the eligibility criteria. A study would have a certain threshold of patients that must meet the eligibility criteria for the site to be considered suitable for the trial. If the number of eligible patients is larger than the threshold then the sponsor will consider the site for inclusion in the trial. Running these eligibility 'count' queries would allow a sponsor to quickly identify candidate sites for a trial that have a sufficiently large patient population. However, providing unrestricted access to count queries on a database can reveal personal information. Existing solutions to this problem (I2B2 and STRIDE [75], [76]) provide approximate counts by adding Gaussian noise to the queries outputs. However, these methods do not offer enough formal privacy guarantees (they only consider privacy loss from averaging the same query over time). In [77], the authors present a novel mechanism for the cohort identification problem with the hope of providing stronger privacy guarantees, and higher utility. They use a modified version of the exponential mechanism of [8] where they incorporate the users' preferences with respect to the magnitude and direction of the perturbation (over or under estimation). Briefly, the mechanism generates a distribution on the count values. The distribution takes the user preferences into consideration, thus, a more desired count value will have higher probability of being returned.

The authors presented a method for comparing the performance of their mechanism with I2B2. To be able to perform the comparison, the authors chose a budget for their mechanism and a parameter for the Gaussian noise in I2B2 that would guarantee a similar variance for the noisy counts in both mechanisms. They found that their method returned the true counts more often than I2B2 while guaranteeing differential privacy. However, it is not clear why equating variances is effective in comparing the two mechanisms given that their distributions are different. Furthermore, one could argue that I2B2 provided higher levels of privacy protection since it did not return the true counts as often.

In tackling the query budgeting issue (budget per query), the authors propose to have a set of predefined query budget levels, thus allowing users to choose their queries' budgets from this predefined set. When the query is expected to have a large count, higher budgets could be used. However, the authers did not specify how these values can be set. Therefore, unless the data custodian has sufficient understanding of these budget values and how much protection each

value really provides, s/he will not be able to set these values without resorting to another privacy standard (as described in Section 4).

On a similar note, the paper did not tackle the issue of choosing a total budget ε in general. However, they described a way to determine a value for ε based on the privacy standard of I2B2. In other words, an ε value that guarantees the same privacy level as I2B2 [75].

To conclude, this application provided a novel way to deal with a very important problem. Its novelty lies in the way it incorporates users preferences with respect to the magnitude and direction of the perturbation as part of the queries. However, the application did not address the basic problems with differential privacy that were mentioned earlier.

8 Conclusions

In this paper we have provided a general overview of the state-of-the-art in differential privacy and outlined some of the limitations of the model and the various mechanisms that have been proposed to implement it. We have further highlighted some practical limitations to the use of differential privacy for the disclosure of health information today. At the same time, the highlighted limitations identify elements of future research programs that could potentially lead to more adoption of differential privacy in healthcare.

The healthcare community still needs to disclose data today. While existing models have limitations, when applied properly they have met current legislation and regulations, and have provided a reasonable amount of protection of patient data. Given that ceasing the disclosure of health information is neither a realistic nor practical goal, until the theoretical and practical limitations, as well as the healthcare specific considerations have been addressed, it would be prudent for data custodians to continue using current methods for anonymizing their data.

Notation table

Notation	
ε	Total privacy budget
$\varepsilon_1, \varepsilon_2, \ldots$	Local privacy budget (per query)
$\wp$	Population
$D, D_1, D_2, \ldots$	Databases drawn from the population
ν	Record from the database
$f, f_1, f_2, \ldots$	Queries
P	Query range
$K_f, K_{f_1}, K_{f_2}, \ldots$	Randomized function for queries $f, f_1, f_2, \ldots$ respectively
s	A value from the randomized query range
$r, r_1, r_2, \ldots$	Values from the query range
$y, y_1, y_2, \ldots$	Noise drawn from a Laplace distribution
$\Delta f, \Delta f_1, \Delta f_2, \ldots$	Sensitivities of queries $f, f_1, f_2, \ldots$ respectively
(ε, τ)	Parameters for approximate differential privacy

Disclaimer and acknowledgements

Our sincere thanks to Luk Arbuckle, and Elizabeth Jonker for their helpful and insightful comments on earlier versions of this paper

References

[1] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, "Privacy-Preserving Data Publishing," *Found. Trends databases*, vol. 2, no. 1–2, pp. 1–167, 2009.

[2] T. Dalenius, "Towards a methodology for statistical disclosure control," *Statistik Tidskrift*, vol. 15, no. 429–444, pp. 2–1, 1977.

[3] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, vol. 4052, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12.

[4] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Lecture notes in computer science*, pp. 265–284, 2006.

[5] J. M. Abowd and L. Vilhuber, "How Protective Are Synthetic Data?," in *Proceedings of the UNESCO Chair in data privacy international conference on Privacy in Statistical Databases*, Berlin, Heidelberg, 2008, pp. 239–246.

[6] C. Dwork, "Differential privacy: a survey of results," in *Proceedings of the 5th international conference on Theory and applications of models of computation*, Berlin, Heidelberg, 2008, pp. 1–19.

[7] R. Sarathy and K. Muralidhar, "Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data," *Transactions on Data Privacy*, vol. 4, no. 1, pp. 1–17, 2011.

[8] F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, Washington, DC, USA, 2007, pp. 94–103.

[9] P. Kodeswaran and E. Viegas, "Applying Differential Privacy to Search Queries in a Policy Based Interactive Framework," in *Proceedings of the ACM International Workshop on Privacy and Annonymity for Very Large Datasets, November 2009*.

[10] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 8, pp. 1200–1214, 2011.

[11] N. Li, W. H. Qardaji, and D. Su, "Provably private data anonymization: Or, k-anonymity meets differential privacy," *CoRR, abs/1101.2604*, 2011.

[12] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 493–502.

[13] J. Lee and C. Clifton, "How Much Is Enough? Choosing epsilon for Differential Privacy," in *Information Security*, vol. 7001, X. Lai, J. Zhou, and H. Li, Eds. Springer Berlin / Heidelberg, 2011, pp. 325–340.

[14] S. P. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman, "The price of privately releasing contingency tables and the spectra of random matrices with correlated rows," in *Proceedings of the 42nd ACM symposium on Theory of computing*, New York, NY, USA, 2010, pp. 775–784.

[15] J. Lee and C. Clifton, "Differential identifiability," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2012, pp. 1041–1049.

[16] S. P. Kasiviswanathan and A. Smith, "A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information," *arXiv:0803.3946*, Mar. 2008.

[17] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *Proceedings of the 42nd ACM symposium on Theory of computing*, New York, NY, USA, 2010, pp. 765–774.

[18] C. Li and G. Miklau, "An adaptive mechanism for accurate query answering under differential privacy," *Proceedings of the VLDB Endowment*, vol. 5, no. 6, pp. 514–525, 2012.

[19] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to non-interactive database privacy," in *Proceedings of the 40th annual ACM symposium on Theory of computing*, New York, NY, USA, 2008, pp. 609–618.

[20] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," *Secure Data Management*, pp. 150–168, 2010.

[21] G. Cormode, M. Procopiuc, D. Srivastava, and T. T. L. Tran, "Differentially Private Publication of Sparse Data," *arXiv:1103.0825*, Mar. 2011.

[22] B. Ding, M. Winslett, J. Han, and Z. Li, "Differentially private data cubes: optimizing noise sources and consistency," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, New York, NY, USA, 2011, pp. 217–228.

[23] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the 42nd ACM symposium on Theory of computing*, 2010, pp. 705–714.

[24] Y. D. Li, Z. Zhang, M. Winslett, and Y. Yang, "Compressive mechanism: Utilizing sparse representation in differential privacy," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, 2011, pp. 177–182.

[25] D. Leoni, "Non-Interactive Differential Privacy: a Survey," in *Proc. of 1st Int. Workshop on Open Data*, Nantes, France, 2012.

[26] A. Roth, "New algorithms for preserving differential privacy," Microsoft Research, 2010.

[27] A. Blum and A. Roth, "Fast Private Data Release Algorithms for Sparse Queries," *CORD Conference Proceedings*, Nov. 2011.

[28] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-*

SIGART symposium on Principles of database systems, New York, NY, USA, 2003, pp. 202–210.

[29] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *Proceedings of the 41st annual ACM symposium on Theory of computing*, New York, NY, USA, 2009, pp. 381–390.

[30] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor, "Optimizing linear counting queries under differential privacy," in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, New York, NY, USA, 2010, pp. 123–134.

[31] M. Hardt and G. N. Rothblum, "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis," in *Foundations of Computer Science, IEEE Annual Symposium on*, Los Alamitos, CA, USA, 2010, vol. 0, pp. 61–70.

[32] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2011, pp. 493–501.

[33] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, New York, NY, USA, 2007, pp. 273–282.

[34] S. E. Fienberg, A. Rinaldo, and X. Yang, "Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables," in *Proceedings of the 2010 international conference on Privacy in statistical databases*, Berlin, Heidelberg, 2010, pp. 187–199.

[35] W. Winkler, "General Discrete-data Modeling Methods for Producing Synthetic Data with Reduced Re-identification Risk that Preserve Analytic Properties," Research Report Series Statistics 2010-02, 2008.

[36] S. E. Fienberg and A. B. Slavkovic, "A Survey of Statistical Approaches to Preserving Confidentiality of Contingency Table Entries," in *Privacy-Preserving Data Mining*, vol. 34, C. C. Aggarwal and P. S. Yu, Eds. Boston, MA: Springer US, 2008, pp. 291–312.

[37] A. Dobra, S. E. Fienberg, A. Rinaldo, A. Slavkovic, and Y. Zhou, "Albebraic Statistics and Contingency Table Problems: Log-Linear

Models, Likelihood Estimation, and Disclosure Limitation," *Emerging applications of algebraic geometry*, pp. 1–26, 2009.

[38] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing Set-Valued Data via Differential Privacy.," *PVLDB*, vol. 4, no. 11, pp. 1087–1098, 2011.

[39] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, p. 2007, 2007.

[40] H. Zang and J. Bolot, "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study," presented at the Proc. of ACM Mobicom, 2011.

[41] M. Goetz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke, "Privacy in Search Logs," *arXiv:0904.0682*, Apr. 2009.

[42] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas, "Releasing search queries and clicks privately," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 171–180.

[43] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 627–636.

[44] Y. Wang, X. Wu, J. Zhu, and Y. Xiang, "On Learning Cluster Coefficient of Private Networks," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[45] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino, "Private record matching using differential privacy," in *Proceedings of the 13th International Conference on Extending Database Technology*, New York, NY, USA, 2010, pp. 123–134.

[46] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private coresets," in *Proceedings of the 41st annual ACM symposium on Theory of computing*, 2009, pp. 361–370.

[47] C. Dwork, "An ad omnia approach to defining and achieving private data analysis," in *Proceedings of the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD*, 2007, pp. 1–13.

[48] K. Muralidhar and R. Sarathy, "Does Differential Privacy Protect Terry Gross' Privacy?," in *Privacy in Statistical Databases*, vol. 6344, J. Domingo-Ferrer and E. Magkos, Eds. Springer Berlin / Heidelberg, 2011, pp. 200–209.

[49] M. M. Baig, J. Li, J. Liu, and H. Wang, "Cloning for privacy protection in multiple independent data publications," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, New York, NY, USA, 2011, pp. 885–894.

[50] J. Domingo-Ferrer, "Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act-Discussion: A Science of Statistical Disclosure Limitation," *International Statistical Review*, vol. 79, no. 2, pp. 184–186, 2011.

[51] R. Sarathy and K. Muralidhar, "Some Additional Insights on Applying Differential Privacy for Numeric Data," in *Privacy in Statistical Databases*, vol. 6344, J. Domingo-Ferrer and E. Magkos, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 210–219.

[52] M. yin, "http://blog.myplaceinthecrowd.org/2010/05/26/recap-and-proposal-955-the-statistically-insignificant-privacy-guarantee/." 2010.

[53] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially private spatial decompositions," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, 2012, pp. 20–31.

[54] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 international conference on Management of data*, 2011, pp. 193–204.

[55] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proceedings of the 2010 international conference on Management of data*, 2010, pp. 735–746.

[56] T.-H. H. C. Elaine Shi, "Privacy-Preserving Aggregation of Time-Series Data.," *NDSS*, 2011.

[57] E. Rieffel, J. Biehl, W. van Melle, and A. J. Lee, "Secured histories: computing group statistics on encrypted data while preserving individual privacy," *arXiv preprint arXiv:1012.2152*, 2010.

[58] F. Dankar, K. E. Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 66, Jul. 2012.

[59] K. El Emam, D. Paton, F. Dankar, and G. Koru, "De-identifying a Public Use Microdata File from the Canadian National Discharge Abstract Database," *BMC Medical Informatics and Decision Making*, vol. 11, no. 53, 2011.

[60] K. El Emam, F. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J.

Bottomley, "A Globally Optimal k-Anonymity Method for the De-identification of Health Data," *Journal of the American Medical Informatics Association*, vol. 16, no. 5, pp. 670–682, 2009.

[61] K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk, "Evaluating Patient Re-identification Risk from Hospital Prescription Records," *Canadian Journal of Hospital Pharmacy*, vol. 62, no. 4, pp. 307–319, 2009.

[62] F. Dankar and K. El Emam, "A Method for Evaluating Marketer Re-identification Risk," in *Proceedings of the 3rd International Workshop on Privacy and Anonymity in the Information Society*, 2010.

[63] K. El Emam, J. Hu, S. Samet, L. Peyton, C. Earle, G. Jayaraman, T. Wong, M. Kantarcioglu, and F. Dankar, "A Protocol for the Secure Linking of Registries for HPV Surveillance," *PLoS ONE*, vol. 7, no. 7, 2012.

[64] K. El Emam and F. K. Dankar, "Protecting Privacy Using k-Anonymity," *J Am Med Inform Assoc*, vol. 15, no. 5, pp. 627–637, 2008.

[65] K. E. Emam, J. Mercer, K. Moreau, I. Grava-Gubins, D. Buckeridge, and E. Jonker, "Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak," *BMC Public Health*, vol. 11, no. 1, p. 454, Jun. 2011.

[66] K. El Emam, A. Brown, and P. AbdelMalik, "Evaluating predictors of geographic area population size cut-offs to manage re-identification risk," *Journal of the American Medical Informatics Association*, vol. 16, no. 2, pp. 256–266, 2009.

[67] K. El Emam, "Heuristics for de-identifying health data," *IEEE Security and Privacy*, pp. 72–75, 2008.

[68] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 2012, pp. 158–166.

[69] K. El Emam, *Guidelines for the de-identification of health information*. CRC Press, 2013.

[70] M. Allison, "Reinventing clinical trials," *Nature Biotechnology*, vol. 30, no. 1, pp. 41–49, Jan. 2012.

[71] OFFICE OF INSPECTOR GENERAL, "Recruiting Human Subjects: Pressures in Industry-Sponsored Clinical Research," Department of Health and Human Services, 2003.

[72] A. Silversides, "Clinical Trials: The Muddled Canadian Landscape," *CMAJ*, vol. 180, no. 1, pp. 20–22, Jan. 2009.

[73] OFFICE OF INSPECTOR GENERAL, "The Globalization of Clinical Trials: A Growing Challenge in Protecting Human Subjects," Department of Health and Human Services, 2001.

[74] F. A. Thiers, A. J. Sinskey, and E. R. Berndt, "Trends in the globalization of clinical trials," *Nature Reviews Drug Discovery*, vol. 7, no. 1, pp. 13–14, Jan. 2008.

[75] S. N. Murphy, M. E. Mendis, D. A. Berkowitz, I. Kohane, and H. C. Chueh, "Integration of Clinical and Genetic Data in the i2b2 Architecture," *AMIA Annu Symp Proc*, vol. 2006, p. 1040, 2006.

[76] H. J. Lowe, T. A. Ferris, P. M. Hernandez, and S. C. Weber, "STRIDE – An Integrated Standards-Based Translational Research Informatics Platform," *AMIA Annu Symp Proc*, vol. 2009, pp. 391–395, 2009.

[77] S. A. Vinterbo, A. D. Sarwate, and A. A. Boxwala, "Protecting Count Queries in Study Design," *J Am Med Inform Assoc*, Apr. 2012.