

Differential Privacy in Metric Spaces: Numerical, Categorical and Functional Data Under the One Roof

Naoise Holohan^{*}; Douglas Leith[†]Oliver Mason[‡]

February 26, 2014

Abstract

We study Differential Privacy in the abstract setting of Probability on metric spaces. Numerical, categorical and functional data can be handled in a uniform manner in this setting. We demonstrate how mechanisms based on data sanitisation and those that rely on adding noise to query responses fit within this framework. We prove that once the sanitisation is differentially private, then so is the query response for any query. We show how to construct sanitisations for high-dimensional databases using simple 1-dimensional mechanisms. We also provide lower bounds on the expected error for differentially private sanitisations in the general metric space setting. Finally, we consider the question of sufficient sets for differential privacy and show that for relaxed differential privacy, any algebra generating the Borel σ -algebra is a sufficient set for relaxed differential privacy.

Keywords: Differential Privacy; Metric Space; Categorical Data; Functional Data; Data Sanitisation

1 Introduction

1.1 Background

The rapid expansion of the Internet and its use in everyday life, alongside the growing understanding of the potential benefits of big data [4], has pushed data privacy to the forefront of research priorities since the turn of the millennium. Whether it be online, in the supermarket or at the hospital, corporations and governments are collecting vast quantities of data about our activities, the choices we make and the people we are in order to work more efficiently, increase profits and better serve our needs as consumers and citizens [5]. The challenge of making this potentially highly sensitive data publicly available where it can be put to good use is far from trivial and it is with this problem that the field of data privacy is concerned.

^{*}Hamilton Institute, NUI Maynooth, Co. Kildare, Ireland

[†]Hamilton Institute, NUI Maynooth, Co. Kildare, Ireland

[‡]Hamilton Institute, NUI Maynooth, Co. Kildare, Ireland. Corresponding Author: email: oliver.mason@nuim.ie. Supported by the HEA, PRTL1 4 Network Mathematics Grant

Various researchers and practitioners have considered applying anonymisation techniques to data sets such as removing explicit identifiers (name, address, telephone number, social security number, etc) while leaving quasi-identifiers¹ in place. While these anonymised data sets do indeed preserve participants' privacy in isolation, auxiliary/background information make this technique extremely vulnerable to attack [22]. A study by L. Sweeney in 2000 [28] found that as much as 87% of the US population (216 out of 248 million people) could be uniquely identified using only three quasi-identifiers (5-digit ZIP code, gender and date of birth). This meant census data could be linked to "anonymised" health records to determine the health status of unsuspecting patients.

Then in 2006, American media firm AOL released 20 million Internet search queries, with user numbers in place of other quasi-identifiers to protect users' identity. This shield of anonymity was not sufficient for privacy to be protected however, and the data was quickly removed from the public domain [2]. Similarly in 2008, A. Narayanan and V. Shmatikov [26] successfully de-anonymised entries in an anonymised data set containing movie ratings of 500,000 subscribers which was released by movie streaming website Netflix. The authors used the publicly-available Internet Movie Database as background information and were able to positively identify known users despite the absence of explicit identifiers in the Netflix data set. Initial anonymisation methods such as k -anonymity [29] and l -diversity [22] have been shown to be vulnerable to attacks based on background information [22, 21]. This fact has led to the development of a wide variety of more sophisticated approaches to publishing and mining data anonymously [23, 24, 25].

The work discussed above underlines the unsatisfactory nature of ad-hoc privacy solutions and the need for a solid theoretical foundation for privacy research. With this in mind, the concept of differential privacy was proposed in [8] to provide a formal, mathematical framework for analysing privacy-preserving data publishing and mining. The premise of differential privacy is that the outputs of queries to a database are unlikely to change substantially with the addition of a new participant's information. This means that outputs will be similar whether or not an individual participates in the database.

There is now a considerable body of work on differential privacy in the theoretical Computer Science literature [9]. Many of the papers in the literature concern data or queries of some particular type or on the development of particular algorithms that satisfy differential privacy. For instance, the design of differentially private algorithms for calculating singular vectors is considered in [10], while differentially private recommender systems are developed in [11]; in both of these instances, the data are naturally modelled as real numbers. Algorithms for search problems and learning are considered in [12, 13]. A statistical perspective on differential privacy was developed in [14]; this paper considered real-valued ($[0, 1]$ in fact) queries and data. In [15], mechanisms that maximise a suitable utility function were investigated; this paper assumed discrete finite-valued data spaces, which can of course describe categorical data. The recent paper [16] addressed the problem of optimal mechanisms that add noise independent of the data; the queries considered are real-valued. To date, the only major reference on differential privacy for functional data appears to be

¹A quasi-identifier is an attribute that is not sufficient to identify an individual by itself, but can do so when combined with other quasi-identifiers (*e.g.* gender, date of birth, etc.).

[18]; in the same paper the authors emphasise the importance of being careful in selecting the measure space with respect to which probabilities are defined. In particular, if we choose our σ -algebra to be the trivial one consisting of the empty set and the entire space, then every mechanism is differentially private. The work of [20] and other similar papers on lower bounds for differentially private mechanisms considers real (or in some cases integer) valued data. Our aim is to describe a unifying framework for all major data and query types considered so far and to initiate a study of differential privacy in the formal setting of probability measures on metric spaces. This provides the machinery necessary for a rigorous discussion of random mechanisms and their accuracy. The major aims of the paper are to place the study of differentially private mechanisms within the framework of probability measures on metric spaces and to present some initial results in this direction.

1.2 Our Results

The principal contributions of this paper are the following.

- We consider differential privacy in the general framework of probability on metric spaces and highlight that it can be seamlessly applied to numerical, categorical and functional data.
- Our description shows how mechanisms based on database sanitisation and adding noise to query responses can be treated in a unified fashion.
- We describe techniques for generating families of (ϵ, δ) differentially private mechanisms from simpler mechanisms. One such example is given by sanitisation mechanisms generated from an (ϵ, δ) differentially private mechanism for the identity query. We also show how to generate differentially private mechanisms for high-dimensional databases using mechanisms for 1-dimensional databases.
- We describe lower bounds for the error in releasing a database in an (ϵ, δ) -differentially private fashion using product sanitisations. This result applies to data drawn from any metric space in contrast to previous work, which has largely focussed on real and integer valued data.
- We consider the question of testing differential privacy and describe sufficient sets for (ϵ, δ) differential privacy.

1.3 Structure of paper

We begin in Section 2 by establishing the measure-theoretic framework for differential privacy. We then consider the question of sufficient sets for differential privacy in Section 3, address sanitisation mechanisms in Section 4 and focus on product sanitisations in Section 5. Section 6 considers accuracy and we give concluding remarks in Section 7.

2 Preliminaries

We first recall some standard concepts and results from probability and measure theory [3, 27]. Given an algebra \mathcal{S} of subsets of a set Ω , we use $\sigma(\mathcal{S})$ to denote

the smallest σ -algebra containing \mathcal{S} and refer to $\sigma(\mathcal{S})$ as the σ -algebra generated by \mathcal{S} . A set Ω together with a σ -algebra of subsets of Ω is a measurable space.

A monotone class \mathcal{M} of subsets of some set Ω is defined by the following two properties: (i) if $\{A_i\}_{i=1}^\infty \subseteq \mathcal{M}$, and if $A_i \subseteq A_{i+1}$ for all i , then $\bigcup_{i=1}^\infty A_i \in \mathcal{M}$; (ii) if $\{A_i\}_{i=1}^\infty \subseteq \mathcal{M}$, and if $A_i \supseteq A_{i+1}$ for all i , then $\bigcap_{i=1}^\infty A_i \in \mathcal{M}$.

The next result, which appears as Theorem 3.4 in [3], characterises $\sigma(\mathcal{S})$ as the smallest monotone class containing \mathcal{S} .

Theorem 1 *Let \mathcal{S} be an algebra of subsets of some set Ω and let \mathcal{M} be a monotone class such that $\mathcal{S} \subseteq \mathcal{M}$. Then $\sigma(\mathcal{S}) \subseteq \mathcal{M}$.*

Given two measurable spaces (X, \mathcal{A}_X) and (Y, \mathcal{A}_Y) , subsets of $X \times Y$ of the form

$$R = \bigcup_{i=1}^p X_i \times Y_i,$$

where $X_i \in \mathcal{A}_X$, $Y_i \in \mathcal{A}_Y$ for $1 \leq i \leq p$ and $(X_i \times Y_i) \cap (X_j \times Y_j) = \emptyset$ for $i \neq j$ are known as *elementary subsets*. Let \mathcal{R} denote the collection of all elementary subsets and denote the usual product σ -algebra on $X \times Y$ by $\mathcal{A}_{X \times Y}$. The following result is Theorem 8.3 of [27].

Theorem 2 *If \mathcal{M} is a monotone class and $\mathcal{R} \subseteq \mathcal{M}$, then $\mathcal{A}_{X \times Y} \subseteq \mathcal{M}$.*

Finally, for a measure μ on a measurable space (X, \mathcal{A}_X) , we recall the following simple fact.

Proposition 1 *Suppose that $\{A_i\}_{i=1}^\infty \subseteq \mathcal{A}_X$ satisfies $A_i \subseteq A_{i+1}$ for all i , then $\lim_{i \rightarrow \infty} \mu(A_i) = \mu(\bigcup_{i=1}^\infty A_i)$.*

Similarly, if $A_i \supseteq A_{i+1}$ for all i , then $\lim_{i \rightarrow \infty} \mu(A_i) = \mu(\bigcap_{i=1}^\infty A_i)$.

2.1 Database Model

The individual entries of the databases we consider are elements of a set $D \subseteq U$ where U is a metric space with metric ρ . We equip U with the Borel σ -algebra generated by the open sets in U (in the metric topology); D then naturally inherits a σ -algebra \mathcal{A}_D . A database \mathbf{d} with n rows is given by a vector $\mathbf{d} = (d_1, \dots, d_n) \in D^n$ in which $d_i \in D$ is the i th row. Throughout, we assume that U^n (and D^n) is equipped with the usual product σ -algebra \mathcal{A}_{U^n} generated by $\{A_1 \times \dots \times A_n : A_i \in \mathcal{A}_U\}$. This ensures that projection maps $\pi_i : U^n \rightarrow U$ given by $\pi_i(x_1, \dots, x_n) = x_i$ are measurable.

It is worth highlighting the generality of this setting: the metric space D can contain numerical, categorical or functional data; moreover, it can be discrete or continuous.

Example 1 *If our data concern the hobbies or interests of people, we consider a set of all possible hobbies, denoted by \mathcal{H} . For simplicity it is not unreasonable to assume that \mathcal{H} is finite. Our data entries are then drawn from the power set $D := 2^{\mathcal{H}}$ of \mathcal{H} , which will again be a finite set. There are various natural choices of metric in this case. We could consider the discrete metric on D in which $\rho_1(A, B) = 1$ if $A \neq B$ and 0 otherwise. Alternatively, we could choose the metric given by symmetric distance: $\rho_2(A, B) = |(A \cup B) \setminus (A \cap B)|$. In*

both of these cases, the Borel σ -algebra consists of all subset of D . Note that there is no requirement that each entry in a database in D^n have the same size or cardinality, reflecting the fact that not all of us have the same number of interests or hobbies.

Example 2 In readings from field deployed sensors, each reading has a time-stamp giving rise to time-course data. Another example is in smart metering where the supplier collects data from consumers giving electricity consumption over a time-window. Data of this type is naturally represented as either a function or a sequence of real numbers. In our framework, we can take U to be a sequence space such as l_∞ or l_2 , or an appropriate function space such as $C([0, T])$ or $L_2([0, T])$, where T represents the billing period (for instance). All of these spaces have natural norms defined on them (in fact they are all Banach spaces) and can be equipped with the Borel σ -algebra generated from the open sets in the norm topology.

We say that two databases $\mathbf{d} = (d_1, \dots, d_n)$ and $\mathbf{d}' = (d'_1, \dots, d'_n)$ in D^n are neighbours, and write $\mathbf{d} \sim \mathbf{d}'$, if there is some $j \in \{1, \dots, n\}$ such that $d_j \neq d'_j$ and $d_i = d'_i$ for all $i \in \{1, \dots, n\} \setminus \{j\}$. More generally, we denote by $h(\mathbf{d}, \mathbf{d}')$ the Hamming distance between \mathbf{d} and \mathbf{d}' .

Remark: While we present our results for Hamming distance on D^n , the results of Sections 3 and 4 are also valid with respect to other metrics ρ on D^n where we define $\mathbf{d} \sim \mathbf{d}'$ if $\rho(\mathbf{d}, \mathbf{d}') = 1$. The work of Sections 5 and 6 relies on similarity being defined with respect to Hamming distance however.

For the most part, we assume that the data space D is compact. This is immediate if D is finite (as in Example 1) and is a natural assumption in most realistic situations. When D is compact, we denote by $\text{diam}(D)$ the diameter of D :

$$\text{diam}(D) = \max_{d, d' \in D} \rho(d, d'). \quad (1)$$

2.2 Query Model

We consider a very general query model. The set of all possible responses is assumed to be a metric space E_Q with metric ρ_Q and equipped with the Borel σ -algebra \mathcal{A}_Q . The query is then a measurable function, $Q : U^n \rightarrow E_Q$ and hence $Q^{-1}(A) \in \mathcal{A}_{U^n}$ for all $A \in \mathcal{A}_Q$.

Example 3 As with the data in \mathbf{d} , queries are not restricted to take numerical values in this setting. For instance, if we consider Example 1 above, then we could consider a query asking for the number of people in the database who are interested in Classical Music or Football for instance: this would clearly be a numeric query. On the other hand, we could also request the 3 most common hobbies in the database, the output of which would be a set.

We next formally introduce the concept of a *response mechanism* within this general framework. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given a collection of queries $\mathcal{Q}(n)$, a response mechanism is a set of measurable mappings

$$\{X_{Q, \mathbf{d}} : \Omega \rightarrow E_Q \mid Q \in \mathcal{Q}(n)\}. \quad (2)$$

Note that $X_{Q,\mathbf{d}}$ is an E_Q -valued random variable for each Q and \mathbf{d} .

Sanitised Response Mechanisms

If there exists some family

$$\{X_{\mathbf{d}} : \Omega \rightarrow U^n\}$$

of measurable mappings such that

$$X_{Q,\mathbf{d}} = Q \circ X_{\mathbf{d}} \tag{3}$$

for all $\mathbf{d} \in D^n$, then the mechanism given by (3) is a *sanitised response mechanism*. The motivation behind this choice of terminology is that the mechanism is generated by sanitising the database via the random variable $X_{\mathbf{d}}$ before answering the query Q , *i.e.* non-interactive data release. If the database is sanitised by adding appropriate noise, the mapping $X_{\mathbf{d}}$ takes the form $X_{\mathbf{d}} = \mathbf{d} + N$ for some U^n -valued random vector N . Clearly, in order to define a mechanism by adding noise, it is necessary for U^n to have a suitable algebraic structure: a vector space or monoid for example.

Output Perturbations

Mechanisms that provide Differential Privacy by perturbing the query response can be described as follows. Let a query $Q : U^n \rightarrow E_Q$ be given. Assume that there is a family of measurable functions $\{X_q : \Omega \rightarrow E_Q \mid q \in E_Q\}$. The output perturbation mechanism is defined as

$$X_{Q,\mathbf{d}} = X_{Q(\mathbf{d})} \tag{4}$$

In situations involving real-valued data, X_q typically takes the form $X_q = q + N$ where N represents the noise added to the query response. For set-valued queries, X_q can be defined by specifying some probability mass function on the collection of possible query responses.

2.3 Differential Privacy

In the interest of completeness and clarity, we now recall the definition of Differential Privacy and write it in the setting of this paper.

Definition 1 (Differential Privacy with Respect to a Query) *Let $\epsilon \geq 0$, $0 \leq \delta \leq 1$ be given. A response mechanism is (ϵ, δ) -differentially private with respect to a query $Q_0 \in \mathcal{Q}(n)$ if for all $\mathbf{d} \sim \mathbf{d}' \in D^n$ and all $A \in \mathcal{A}_{Q_0}$,*

$$\mathbb{P}(X_{Q_0,\mathbf{d}} \in A) \leq e^\epsilon \mathbb{P}(X_{Q_0,\mathbf{d}'} \in A) + \delta. \tag{5}$$

It is important to note that the relation $\mathbf{d} \sim \mathbf{d}'$ is symmetric, so inequality (5) is required to hold when \mathbf{d} and \mathbf{d}' are swapped.

Definition 2 (Differential Privacy) *A response mechanism is (ϵ, δ) -differentially private with respect to $\mathcal{Q}(n)$ if it is (ϵ, δ) -differentially private with respect to every query $Q_0 \in \mathcal{Q}(n)$.*

The above definitions are often referred to as *relaxed* differential privacy; the original notion of differential privacy introduced in [8] considers the case where $\delta = 0$.

The *expected error* of a mechanism (2) for a query Q on a database \mathbf{d} is given by the expectation $\mathbb{E}[\text{dist}(X_{Q,\mathbf{d}}, Q(\mathbf{d}))]$.

3 Sufficient Sets for Differential Privacy

In this brief section, we consider the following question: is there a strict subset \mathcal{S} of \mathcal{A}_Q such that if (5) is satisfied for all A in \mathcal{S} , it is guaranteed to be satisfied for all A in \mathcal{A}_Q ? We refer to such a set as a sufficient set for differential privacy.

Depending on the application, the query output space may be a subset of \mathbb{R}^n or of a sequence or function space such as $C([0, T])$. A key question for the practical deployment of differentially private mechanisms is how to determine if a mechanism is in fact (ϵ, δ) -differentially private. Testing (5) on the entire σ -algebra is clearly a prohibitively difficult task. Our next result shows that it is sufficient to test this condition on any algebra $\mathcal{S} \subset \mathcal{A}_Q$ that generates \mathcal{A}_Q .

Theorem 3 *Let a response mechanism (2) and a query (E_Q, \mathcal{A}_Q, Q) be given and let $\mathcal{S} \subset \mathcal{A}_Q$ be an algebra such that $\sigma(\mathcal{S}) = \mathcal{A}_Q$. If (5) holds for all sets $A \in \mathcal{S}$, then it holds for all sets $A \in \mathcal{A}_Q$.*

Proof: Let \mathcal{B} denote the collection of sets in E_Q for which (5) holds. By assumption, $\mathcal{S} \subseteq \mathcal{B}$. Now let A_1, A_2, \dots be any collection of sets in \mathcal{B} with $A_i \subseteq A_{i+1}$ for all i . Define $\bar{A} := \cup_i A_i$ and let $\mathbf{d}, \mathbf{d}' \in D^n$ with $\mathbf{d} \sim \mathbf{d}'$ be given. As each $A_i \in \mathcal{B}$, it follows that

$$\mathbb{P}(X_{Q, \mathbf{d}}(\omega) \in A_i) \leq e^\epsilon \mathbb{P}(X_{Q, \mathbf{d}'}(\omega) \in A_i) + \delta$$

for all i . As the sequence A_i is increasing, it now follows from Proposition 1 that

$$\begin{aligned} \mathbb{P}(X_{Q, \mathbf{d}}(\omega) \in \bar{A}) &= \lim_{i \rightarrow \infty} \mathbb{P}(X_{Q, \mathbf{d}}(\omega) \in A_i) \\ &\leq e^\epsilon \lim_{i \rightarrow \infty} \mathbb{P}(X_{Q, \mathbf{d}'}(\omega) \in A_i) + \delta \\ &= e^\epsilon \mathbb{P}(X_{Q, \mathbf{d}'}(\omega) \in \bar{A}) + \delta, \end{aligned}$$

and so $\bar{A} \in \mathcal{B}$. An identical argument shows that for any sequence $\{A_i\}$ of sets in \mathcal{B} with $A_i \supseteq A_{i+1}$ for all i , and $\bar{A} = \cap_i A_i$, $\bar{A} \in \mathcal{B}$. Taken together these two observations imply that \mathcal{B} is a *monotone class*. Moreover, $\mathcal{S} \subseteq \mathcal{B}$. The result now follows immediately from Theorem 1. \square

In the next example we show how Theorem 3 can be applied to differentially private mechanisms for functional data to obtain results such as those described in Section 3.1 of [18].

Example 4 *Suppose our query Q takes values in the space $C([0, 1])$ of continuous functions on $[0, 1]$ equipped with the norm $\|f\|_\infty = \sup\{|f(t)| : t \in [0, 1]\}$ and the σ -algebra \mathcal{A}_Q of Borel sets generated by the norm topology. Let a mechanism $X_{Q, \mathbf{d}}$ be given. Then $X_{Q, \mathbf{d}}(\omega)$ is in $C([0, 1])$ for each $\omega \in \Omega$.*

Given a positive integer k and real numbers $0 \leq t_1 < \dots < t_k \leq 1$, consider the projection $\pi_{t_1, \dots, t_k} : C([0, 1]) \rightarrow \mathbb{R}^k$ given by

$$\pi_{t_1, \dots, t_k}(f) = (f(t_1), \dots, f(t_k)).$$

These mappings are measurable with respect to the usual Borel σ -algebra on \mathbb{R}^k and hence we can define the \mathbb{R}^k -valued mechanism $X_{Q, \mathbf{d}}^{t_1, \dots, t_k} = \pi_{t_1, \dots, t_k} \circ X_{Q, \mathbf{d}}$.

We claim that if the finite-dimensional mechanisms $X_{Q, \mathbf{d}}^{t_1, \dots, t_k}$ are (ϵ, δ) -differentially private for all k and t_1, \dots, t_k , then the mechanism $X_{Q, \mathbf{d}}$ is (ϵ, δ) -differentially

private. The argument to show this is as follows. From the assumption on the finite-dimensional mechanisms, it follows immediately that (5) holds for all (so-called cylinder sets) sets A of the form

$$A = \pi_{t_1, \dots, t_k}^{-1}(B)$$

where B is a Borel set in \mathbb{R}^k . These sets form an algebra and it follows from Theorem VII.2.1 (page 212) of [17] that the σ -algebra they generate is the Borel σ -algebra of $C([0, 1])$. It follows immediately from Theorem 3 that $X_{Q, \mathbf{d}}$ defines an (ϵ, δ) -differentially private mechanism on $C([0, 1])$ as claimed.

4 Sanitisation Mechanisms and the Identity Query

A popular approach to designing differentially private response mechanisms is to add *noise* to the query response $Q(\mathbf{d})$. It is known however, that this can lead to privacy compromises by averaging a large number of responses to an identical query [9], unless the number or type of queries that can be asked is restricted. We now show that if a sanitisation mechanism (3) is (ϵ, δ) -differentially private with respect to the identity query, then it is (ϵ, δ) -differentially private with respect to *any* query. It is important to appreciate that we place no restrictions on the query Q ; it could map into a sequence space such as l_∞ and represent an infinite sequence of individual queries for example.

The identity query is defined by the identity map on the ambient space U^n . In sanitised response mechanisms, the “perturbed database” $X_{\mathbf{d}}(\omega)$ is disclosed by the response to the identity query. However, disclosing $X_{\mathbf{d}}(\omega)$ need *not* disclose the original database \mathbf{d} provided that an appropriate privacy-preserving perturbation has been applied. Importantly, if (ϵ, δ) -differential privacy is achieved with respect to query I_n , then the response to *any* query is (ϵ, δ) -differentially private.

Theorem 4 (Identity Query) *Consider a sanitised response mechanism as defined in (3). Suppose this response mechanism is (ϵ, δ) -differentially private with respect to the identity query. Then it is (ϵ, δ) -differentially private with respect to any query (E_Q, \mathcal{A}_Q, Q) .*

Proof: Let \mathbf{d}, \mathbf{d}' be two elements of D^n with $\mathbf{d} \sim \mathbf{d}'$. By assumption,

$$\mathbb{P}(X_{\mathbf{d}}(\omega) \in E) \leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'}(\omega) \in E) + \delta, \quad (6)$$

for all $E \in \mathcal{A}_{U^n}$. Let a query (E_Q, Q, \mathcal{A}_Q) and $A \in \mathcal{A}_Q$ be given. As Q is measurable, $Q^{-1}(A) \in \mathcal{A}_{U^n}$. Then, using (6),

$$\begin{aligned} \mathbb{P}(X_{Q, \mathbf{d}}(\omega) \in A) &= \mathbb{P}(Q(X_{\mathbf{d}}(\omega)) \in A) \\ &= \mathbb{P}(X_{\mathbf{d}}(\omega) \in Q^{-1}(A)) \\ &\leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'}(\omega) \in Q^{-1}(A)) + \delta \\ &= \mathbb{P}(Q(X_{\mathbf{d}'}(\omega)) \in A) + \delta. \end{aligned}$$

Thus, the mechanism satisfies (ϵ, δ) -differential privacy with respect to Q also. \square

Corollary 1 Consider a sanitised response mechanism as defined in (3) and suppose $I_n \in \mathcal{Q}(n)$. Then this response mechanism is (ϵ, δ) -differentially private with respect to I_n if and only if it is (ϵ, δ) -differentially private with respect to every query $Q \in \mathcal{Q}(n)$.

Proof: “ \Rightarrow ”: Theorem 4.

“ \Leftarrow ”: The response mechanism is (ϵ, δ) -differentially private with respect to every query $Q \in \mathcal{Q}(n)$ by assumption, therefore it must be (ϵ, δ) -differentially private with respect to the identity query I_n , since $I_n \in \mathcal{Q}(n)$. \square

Observe that the number and details of queries $Q \in \mathcal{Q}(n)$ do not need to be specified in advance for Theorem 4 to hold, and so queries may be interactive and unlimited in number. This highlights a fundamental difference between privacy mechanisms that perturb the query response (*e.g.* by adding Laplacian or Gaussian noise) vs privacy mechanisms that perturb the database itself. Namely, in the former the added noise can be averaged out by an adversary repeating a query multiple times, thereby requiring a limit to be placed on the number of queries allowed, while in the latter an averaging attack of this sort is impossible; a repeated query will simply receive the same answer each time.

5 Product Sanitisations

In this section, we derive a result that relates differentially private sanitisation-based mechanisms for n -dimensional databases in D^n to mechanisms for simple 1-dimensional databases.

Before showing how differentially private sanitisation mechanisms for databases in D^n can be constructed from simple mechanisms for databases in D , we first establish a number of technical results.

Lemma 1 Let $A_1, \dots, A_p, B_1, \dots, B_p$ be two collections of non-empty sets. Then the finite union $\bigcup_{i=1}^p (A_i \times B_i)$ can be written as

$$\bigcup_{i=1}^p (A_i \times B_i) = \bigcup_{I \subseteq \{1, \dots, p\}} (\tilde{A}_I \times \tilde{B}_I),$$

where $\tilde{A}_I = \bigcup_{i \in I} A_i$ and $\tilde{B}_I = \bigcap_{i \in I} B_i \setminus \bigcup_{i \notin I} B_i$. Moreover, $\tilde{B}_I \cap \tilde{B}_J = \emptyset$ for all $I \neq J$.

Proof: We need to prove equality and disjointness of the decomposition. Let $(a, b) \in \bigcup_{i=1}^p (A_i \times B_i)$. Then there exists at least one i^* such that $(a, b) \in A_{i^*} \times B_{i^*}$. Let $I_b := \{i : b \in B_i\} \subseteq \{1, \dots, p\}$ (note $i^* \in I_b$). Then $b \in \bigcap_{i \in I_b} B_i$, but $b \notin B_j$ for any $j \notin I_b$, otherwise j would be an element of I_b . Hence $b \in \bigcap_{i \in I_b} B_i \setminus \bigcup_{i \notin I_b} B_i$. Also $a \in \bigcup_{i \in I_b} A_i$ since $a \in A_{i^*}$. Hence $(a, b) \in \bigcup_{I \subseteq \{1, \dots, p\}} (\bigcup_{i \in I} A_i \times \bigcap_{i \in I} B_i \setminus \bigcup_{i \notin I} B_i)$.

Let $(a, b) \in \bigcup_{I \subseteq \{1, \dots, p\}} (\bigcup_{i \in I} A_i \times \bigcap_{i \in I} B_i \setminus \bigcup_{i \notin I} B_i)$. Then there exists at least one $I^* \subseteq \{1, \dots, p\}$ such that $(a, b) \in \bigcup_{i \in I^*} A_i \times \bigcap_{i \in I^*} B_i \setminus \bigcup_{i \notin I^*} B_i$. Hence $a \in A_i$ for at least one $i \in I^*$ and $b \in B_i$ for all $i \in I^*$ and so there exists at least one $i \in I^*$ such that $(a, b) \in A_i \times B_i$ and so $(a, b) \in \bigcup_{i=1}^p (A_i \times B_i)$.

Finally, we show that $\tilde{B}_I \cap \tilde{B}_J = \emptyset$ if $I \neq J$. To see this, note that if $I \neq J$, then we can without loss of generality assume that there is some index $k \in I$

that is not in J . Then any $x \in \tilde{B}_I$ must be in B_k . However, as $k \in J^C$, it follows that $x \in \cup_{i \notin J} B_k$ and hence that $x \notin B_J$. This shows that the intersection is empty as claimed. \square

For future use, we note that an analogous argument to that given above can be used to show the following.

Lemma 2 *Let $A_1, \dots, A_p, B_1, \dots, B_p$ be two collections of non-empty sets. Then the finite union $\bigcup_{i=1}^p (A_i \times B_i)$ can be written as*

$$\bigcup_{j=1}^q (\tilde{A}_j \times \tilde{B}_j),$$

where $\tilde{A}_i \cap \tilde{A}_j = \emptyset$ for $i \neq j$.

For the remainder of this section we consider a special form for the database sanitisation $X_{\mathbf{d}}(\omega)$. Suppose a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings is given. Define the mechanism $X_{\mathbf{d}}$ for $\mathbf{d} = (d_1, \dots, d_n)$ by

$$X_{\mathbf{d}}(\omega) = (X_{\mathbf{d}}^1(\omega), \dots, X_{\mathbf{d}}^n(\omega)), \quad (7)$$

where the $X_{\mathbf{d}}^i$ are independent and $X_{\mathbf{d}}^i$ has the same distribution as X_{d_i} , for all $\mathbf{d} \in D^n, i \in \{1, \dots, n\}$.

We first note the following lemma concerning such mechanisms.

Lemma 3 *Let a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings be given and let $X_{\mathbf{d}}$ be defined by (7). If $X_{\mathbf{d}}$ is (ϵ, δ) -differentially private then*

$$\mathbb{P}(X_d \in A) \leq e^\epsilon \mathbb{P}(X_{d'} \in A) + \delta,$$

for all $d, d' \in D, A \in \mathcal{A}_U$.

Proof: Let d, d' in D be given. If $d = d'$, the result is trivial. If $d \neq d'$, take $\mathbf{d} = (d, d_2, \dots, d_n)$, $\mathbf{d}' = (d', d_2, \dots, d_n)$ for any choice of d_2, \dots, d_n in D . As $X_{\mathbf{d}}$ is (ϵ, δ) -differentially private and the projection $\pi_1 : U^n \rightarrow U$ onto the first coordinate is measurable, it follows that for $A \in \mathcal{A}_U$:

$$\begin{aligned} \mathbb{P}(X_d \in A) &= \mathbb{P}(\pi_1(X_{\mathbf{d}}) \in A) \\ &= \mathbb{P}(X_{\mathbf{d}} \in \pi_1^{-1}(A)) \\ &\leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'} \in \pi_1^{-1}(A)) + \delta \\ &= e^\epsilon \mathbb{P}(X_{d'} \in A) + \delta. \end{aligned}$$

\square

We next note that the converse of this result also holds.

Theorem 5 *Consider a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings and assume that*

$$\mathbb{P}(X_d \in A) \leq e^\epsilon \mathbb{P}(X_{d'} \in A) + \delta,$$

for all $d, d' \in D, A \in \mathcal{A}_U$. Let $X_{\mathbf{d}}$ be as defined in (7). Then,

$$\mathbb{P}(X_{\mathbf{d}} \in A) \leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'} \in A) + \delta,$$

for all $\mathbf{d} \sim \mathbf{d}' \in D^n$ and all $A \in \mathcal{A}_{U^n}$.

Proof: We shall use induction on n . By assumption, the result is true for $n = 1$. Let $n > 1$ be given and assume that the result is true for all $k \leq n - 1$.

Assume that \mathbf{d} and \mathbf{d}' differ in the first element so $d_1 \neq d'_1$ but $d_j = d'_j$ for $j \neq 1$. Let

$$R = \bigcup_{i=1}^p (A_i \times B_i), \quad (8)$$

where $A_i \in \mathcal{A}_U$, $B_i \in \mathcal{A}_{U^{n-1}}$, be given. It follows from Lemma 1 that we can write

$$R = \bigcup_{i=1}^q (\tilde{A}_i \times \tilde{B}_i), \quad (9)$$

where $\tilde{A}_i \in \mathcal{A}_U$, $\tilde{B}_i \in \mathcal{A}_{U^{n-1}}$ for $1 \leq i \leq q$ and $\tilde{B}_i \cap \tilde{B}_j = \emptyset$ for $i \neq j$. Then, using the fact that the sets \tilde{B}_i are disjoint and the independence of the components of $X_{\mathbf{d}}$, $X_{\mathbf{d}'}$,

$$\begin{aligned} \mathbb{P}(X_{\mathbf{d}} \in R) &= \sum_{i=1}^q \mathbb{P}(X_{\mathbf{d}} \in \tilde{A}_i \times \tilde{B}_i) \\ &= \sum_{i=1}^q \mathbb{P}(X_{d_1} \in \tilde{A}_i) \mathbb{P}(X_{(d_2, \dots, d_n)} \in \tilde{B}_i) \\ &\leq \sum_{i=1}^q \left(e^\epsilon \mathbb{P}(X_{d'_1} \in \tilde{A}_i) + \delta \right) \mathbb{P}(X_{(d'_2, \dots, d'_n)} \in \tilde{B}_i) \\ &= e^\epsilon \sum_{i=1}^q \mathbb{P}(X_{\mathbf{d}'} \in \tilde{A}_i \times \tilde{B}_i) + \delta \mathbb{P} \left(X_{(d'_2, \dots, d'_n)} \in \bigcup_{i=1}^q \tilde{B}_i \right) \\ &\leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'} \in R) + \delta. \end{aligned}$$

If $d_1 = d'_1$, then $(d_2, \dots, d_n) \sim (d'_2, \dots, d'_n)$ and we can use Lemma 2 and the induction hypothesis to conclude that

$$\begin{aligned} \mathbb{P}(X_{\mathbf{d}} \in R) &\leq \sum_{i=1}^q \mathbb{P}(X_{d'_1} \in \tilde{A}_i) \left(e^\epsilon \mathbb{P}(X_{(d'_2, \dots, d'_n)} \in \tilde{B}_i) + \delta \right) \\ &= e^\epsilon \sum_{i=1}^q \mathbb{P}(X_{\mathbf{d}'} \in \tilde{A}_i \times \tilde{B}_i) + \delta \mathbb{P} \left(X_{d'_1} \in \bigcup_{i=1}^q \tilde{A}_i \right) \\ &\leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'} \in R) + \delta. \end{aligned}$$

Thus for any set R of the form (8), we can conclude that

$$\mathbb{P}(X_{\mathbf{d}} \in R) \leq e^\epsilon \mathbb{P}(X_{\mathbf{d}'} \in R) + \delta. \quad (10)$$

In particular, (10) holds for all elementary sets $R \in \mathcal{A}_{D^n}$. A similar argument to that used in the proof of Theorem 3 shows that the collection of all sets satisfying (10) is a monotone class; furthermore this collection of subsets contains the elementary sets. The result now follows from Theorem 2. \square

Remark: In the following subsection, we describe some simple applications of Theorem 5. It is worth noting that it applies to any database space $D \subseteq U$, and to *discrete* spaces in particular. For mechanisms of the form (7), it simplifies the task of testing the mechanism for differential privacy considerably. For instance, if D is a finite set with $|D|$ elements, then it is only necessary to check (5) for all $\binom{|D|}{2}$ pairs of elements of D and all $2^{|D|}$ subsets of D to ensure differential privacy on D^n . In general, we would have $n\binom{|D|}{2}|D|^{n-1}$ pairs of neighbouring elements and $2^{|D|^n}$ subsets to worry about!

5.1 Examples

Example 5 *Inspired by a now standard approach to designing differentially private mechanisms, we first consider a sanitisation mechanism for real-valued databases in which Laplacian noise is added to each element of the database. Recall that a Laplacian random variable $X : \Omega \rightarrow \mathbb{R}$ with mean zero and variance $2b^2$ has a probability density function (PDF) is given by*

$$f(x) = \frac{1}{2b} e^{-\frac{|x|}{b}}.$$

Let $D \subset \mathbb{R}$ be bounded; for each $d \in D$, let $X_d(\omega) = d + L(\omega)$ where $L : \Omega \rightarrow \mathbb{R}$ is a Laplacian random variable with mean zero and variance $2b^2$ such that

$$b \geq \frac{\text{diam}(D)}{\epsilon - \log(1 - \delta)}.$$

The resulting sanitised response mechanism corresponding to (7) is (ϵ, δ) -differentially private for any database in D^n .

To see this, note that by Theorem 5, it is sufficient to show that

$$\int_A \frac{e^{-\frac{|x-d|}{b}}}{2b} dx \leq e^\epsilon \int_A \frac{e^{-\frac{|x-d'|}{b}}}{2b} dx + \delta,$$

for all $d, d' \in D, A \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ -algebra on the real line \mathbb{R} . Using the triangle inequality, $|x - d'| \leq |x - d| + |\Delta|$ where $\Delta = d' - d$, and so it is sufficient to show that

$$\int_A \frac{e^{-\frac{|x-d|}{b}}}{2b} dx \leq e^{\epsilon - \frac{|\Delta|}{b}} \int_A \frac{e^{-\frac{|x-d|}{b}}}{2b} dx + \delta,$$

for all $d', d \in D, A \in \mathcal{B}(\mathbb{R})$. This last inequality will follow if $1 \leq e^{\epsilon - \frac{|\Delta|}{b}} + \delta$ or $b \geq \frac{|\Delta|}{\epsilon - \log(1 - \delta)}$ for all $\Delta \in \{d' - d : d', d \in D\}$.

Of course, keeping in mind that the l_1 sensitivity of the identity query [9] is precisely given by $\text{diam}(D)$, this example can be seen as an application of the well-known Laplacian mechanism to the identity query.

Example 6 *Consider again our earlier example where $D = 2^{\mathcal{H}}$ represents the sets of possible hobbies or interests of people. As noted earlier, it is reasonable to assume that D contains finitely many elements; we denote $|D| = m + 1$. Following Theorem 5 we will construct a mechanism for 1-dimensional databases: this can then be used to define a mechanism for databases in D^n via (7).*

For $d \in D$, consider the D -valued random variable X_d with probability mass function:

$$\mathbb{P}(X_d = d) = 1 - pm, \quad \mathbb{P}(X_d = d') = p,$$

where $d \neq d' \in D$. We make the reasonable assumption that $1 - pm > p$.

For (ϵ, δ) -differential privacy, we need the following to hold:

$$\mathbb{P}(X_d \in A) \leq e^\epsilon \mathbb{P}(X_{d'} \in A) + \delta, \quad (11)$$

where $A \subseteq D$ and $d, d' \in D$.

We claim that (11) will hold if and only if

$$1 - pm \leq e^\epsilon p + \delta. \quad (12)$$

This condition is clearly necessary as can be seen by considering the singleton set $A = \{d\}$. To see that it is also sufficient let $d, d' \in D$ and $A \subseteq D$ be given. There are 4 cases to consider.

1. $d, d' \notin A$: Then $\mathbb{P}(X_d \in A) = \mathbb{P}(X_{d'} \in A) = p|A|$ and (ϵ, δ) -differential privacy holds trivially.
2. $d, d' \in A$: Then $\mathbb{P}(X_d \in A) = \mathbb{P}(X_{d'} \in A) = p(|A| - 1) + 1 - pm = p(|A| - m - 1) + 1$ and (ϵ, δ) -differential privacy holds trivially.
3. $d' \in A, d \notin A$: Then $\mathbb{P}(X_d \in A) \leq \mathbb{P}(X_{d'} \in A)$ and (ϵ, δ) -differential privacy holds trivially.
4. $d \in A, d' \notin A$: Then

$$\begin{aligned} \mathbb{P}(X_d \in A) &= p(|A| - m - 1) + 1 \\ \mathbb{P}(X_{d'} \in A) &= p|A|. \end{aligned}$$

From (12), we have

$$\begin{aligned} 1 - pm &\leq e^\epsilon p + \delta \\ &= e^\epsilon (p|A| - p|A| + p) + \delta \\ &\leq e^\epsilon p|A| - p(|A| - 1) + \delta, \end{aligned}$$

since $|A| \geq 1$ ($d \in A$ by hypothesis). Rearranging the above inequality, we see that

$$p(|A| - m - 1) + 1 \leq e^\epsilon (p|A|) + \delta.$$

Thus we can construct an (ϵ, δ) -differentially private mechanism of the form (7) by choosing $p \geq \frac{1-\delta}{m+e^\epsilon}$.

6 Accuracy

In this section, we consider the question of accuracy for product sanitisations. The literature on the interaction between privacy and accuracy is considerable and previous work has considered examples such as count queries [19], contingency table marginals [1] and spatial data [6]. As product sanitisations are

constructed from 1-dimensional mechanisms, we focus on the error of these basic mechanisms here. These results can then be used to derive bounds for data in D^n ; the precise form these bounds will take depends on how the metric is constructed on D^n .

We wish to emphasise two points about our work: we are considering a very general class of databases that can incorporate numerical, categorical and functional data; we consider (ϵ, δ) -differential privacy and are not assuming $\delta = 0$.

As $\rho(\cdot, d)$ is a continuous function on D for any fixed d , it is measurable with respect to the Borel σ -algebra on D . It follows that $\rho(X_d, d)$ is a nonnegative-valued random variable. We define the maximal expected error \mathcal{E} as:

$$\mathcal{E} = \max_{d \in D} \mathbb{E} [\rho(X_d, d)]. \quad (13)$$

For $r > 0$ and $x \in D$, $B_r(x)$ denotes the open ball

$$B_r(x) := \{y \in D \mid \rho(y, x) < r\}.$$

We first note that for any differentially private mechanism with $\delta < 1$, $\mathcal{E} > 0$. If $\delta = 1$, then any mechanism is differentially private.

Lemma 4 *Let a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings be given, let $0 \leq \delta < 1$ and assume that*

$$\mathbb{P}(X_d \in A) \leq e^\epsilon \mathbb{P}(X_{d'} \in A) + \delta, \quad (14)$$

for all $d, d' \in D, A \in \mathcal{A}_U$. Then $\mathcal{E} > 0$.

Proof: As D is compact, we can choose u, v in D with $\rho(u, v) = \text{diam}(D)$. Let $r = \frac{\text{diam}(D)}{2}$. Then from (14), it follows that

$$\mathbb{P}(X_u \in B_r(v)) \geq e^{-\epsilon} (\mathbb{P}(X_v \in B_r(v)) - \delta).$$

As $\rho(x, u) \geq r > 0$ for all $x \in B_r(v)$, it follows that $\mathbb{E}[\rho(X_u, u)] > 0$ unless

$$\mathbb{P}(X_v \in B_r(v)) = \delta. \quad (15)$$

However, if this is the case then $\mathbb{P}(\rho(X_v, v) \geq r) = 1 - \delta > 0$ and hence $\mathbb{E}[\rho(X_v, v)] \geq r(1 - \delta) > 0$. This completes the proof. \square

We now present two simple results giving lower bounds for \mathcal{E} . The argument for the following result is inspired by that used to establish Theorem 3.3 of [20].

Theorem 6 *Let a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings satisfying (14) be given. Then*

$$\mathcal{E} \geq (1 - \delta) \left(\frac{\text{diam}(D)}{2(1 + e^\epsilon)} \right). \quad (16)$$

Proof: Without loss of generality, we may assume that \mathcal{E} is finite. Moreover, from Lemma 4 we know that $\mathcal{E} > 0$. As D is compact, there exist points u, v in D with $\rho(u, v) = \text{diam}(D)$. Set $t = \frac{\text{diam}(D)}{2\mathcal{E}}$; then $t\mathcal{E} = \frac{\text{diam}(D)}{2}$ and

the balls $B_{t\mathcal{E}}(u)$, $B_{t\mathcal{E}}(v)$ are disjoint. From Markov's inequality applied to the non-negative random variable $\rho(X_u, u)$, it follows that

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(u)) \geq 1 - \frac{1}{t} = 1 - \frac{2\mathcal{E}}{\text{diam}(D)}. \quad (17)$$

It is now immediate that

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(v)) \leq \frac{2\mathcal{E}}{\text{diam}(D)}. \quad (18)$$

From (14) we know that

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(v)) \geq e^{-\epsilon}(\mathbb{P}(X_v \in B_{t\mathcal{E}}(v)) - \delta). \quad (19)$$

Combining (18), (19) and noting that (17) also holds with u replaced by v , we see that

$$\frac{2\mathcal{E}}{\text{diam}(D)} \geq e^{-\epsilon} \left(1 - \frac{2\mathcal{E}}{\text{diam}(D)} - \delta \right)$$

and after a simple rearrangement of terms we see that

$$\mathcal{E} \geq (1 - \delta) \left(\frac{\text{diam}(D)}{2(1 + e^\epsilon)} \right),$$

as claimed. \square

The previous result applies to any compact metric space D . Now assume that D is discrete so that there exists some $\kappa > 0$ such that

$$\rho(x, y) \geq \kappa \quad \forall x, y \in D. \quad (20)$$

This combined with D being compact immediately implies that D is finite. A straightforward alteration of the argument of Theorem 6 yields the following result.

Theorem 7 *Let D be finite with $|D| = m + 1$ and $\kappa = \min_{d, d' \in D} \rho(d, d')$. Let a family $\{X_d : \Omega \rightarrow U \mid d \in D\}$ of measurable mappings be given satisfying (14). Then*

$$\mathcal{E} \geq (1 - \delta) \left(\frac{\kappa m}{(m + e^\epsilon)} \right) \quad (21)$$

Proof: It is trivial that the $m + 1$ balls $B_{t\mathcal{E}}(u)$, $u \in D$ are all disjoint where $t = \frac{\kappa}{\mathcal{E}}$. Fix some $u \in D$. By the same reasoning as in the proof of Theorem 6,

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(u)) \geq 1 - \frac{\mathcal{E}}{\kappa}. \quad (22)$$

Moreover, there must exist some $v \neq u$ such that

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(v)) \leq \frac{\mathcal{E}}{\kappa m}. \quad (23)$$

Choose one such v and apply (14) to obtain

$$\mathbb{P}(X_u \in B_{t\mathcal{E}}(v)) \geq e^{-\epsilon} (\mathbb{P}(X_v \in B_{t\mathcal{E}}(v)) - \delta). \quad (24)$$

As in the proof of Theorem 6, we can now conclude that

$$\frac{\mathcal{E}}{\kappa m} \geq e^{-\epsilon} \left(1 - \frac{\mathcal{E}}{\kappa} - \delta \right).$$

Rearranging this inequality gives us that

$$\mathcal{E} \geq (1 - \delta) \left(\frac{\kappa m}{m + e^\epsilon} \right).$$

□

Example 7 Consider again Example 6. We have shown that there exists an (ϵ, δ) -differentially private mechanism with $p = \frac{1-\delta}{m+e^\epsilon}$ where $|D| = m + 1$. If D is equipped with the discrete metric so that $\rho(d, d') = 1$ for all $d \neq d'$, then $\kappa = 1$ and for any d , the expected value of $\rho(X_d, d)$ for this mechanism is given by

$$\sum_{d \neq d'} p = mp = (1 - \delta) \left(\frac{m}{m + e^\epsilon} \right).$$

So the bound given by Theorem 7 is tight in this simple case.

7 Concluding Remarks

We have considered differential privacy in the setting of probability on metric spaces with mechanisms viewed as measurable functions taking values in output query spaces. We have demonstrated that this framework allows mechanisms based on sanitisation and output perturbation to be treated in a uniform manner; moreover we have presented examples to highlight that categorical, functional and numerical data can be treated in this setting. For sanitisation mechanisms, a formal proof that differential privacy with respect to the identity query guarantees differential privacy with respect to any measurable query has been given. We have also introduced the problem of determining sufficient sets for differential privacy, shown that a generating algebra of sets is a sufficient set and applied this fact to functional data in the space $C([0, 1])$ of continuous functions on $[0, 1]$. In the latter half of the paper, we have focussed on product sanitisations of the form (7); we have shown that these mechanisms are (ϵ, δ) -differentially private if and only if the 1-dimensional mechanism used to define them is. This result was then applied in two contexts: to provide a condition for the well-known Laplacian mechanism to be differentially private; and to give a simple example of a differentially private mechanism for discrete categorical data. Finally in Section 6, two simple results giving lower bounds for the maximal expected error for (ϵ, δ) differentially private mechanisms on metric spaces were presented.

References

- [1] 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*, pp. 273–282, 2007.

- [2] M. Barbaro and T. Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749, New York Times, August 9, 2006.
- [3] P. Billingsley. *Probability and Measure* Wiley Series in Probability and Mathematical Statistics, 1995.
- [4] D. Boyd and K. Crawford. Six Provocations for Big Data In the Oxford Internet Institutes *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
- [5] Computer Sciences Corporation (CSC). Big Data Universe Beginning to Explode 2011.
- [6] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen and T. Yu. Differentially Private Spatial Decompositions In *IEEE 28th International Conference on Data Engineering (ICDE)*, pp. 20–31, 2012.
- [7] I. Dinur, K. Nissim Revealing Information while Preserving Privacy In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp 202–210, 2003.
- [8] C. Dwork. Differential Privacy. In *Proceedings of the 33rd Annual International Colloquium on Automata, Languages and Programming*, LNCS 4051, Springer, pp. 1–12, 2006.
- [9] C. Dwork. Differential Privacy: A Survey of Results In *5th International Conference on Theory and Applications of Models of Computation*, LNCS 4978, Springer, pp. 1–19, 2008.
- [10] M. Kapralov and K. Talwar. Differentially Private Low Rank Approximation *SODA*, 2013
- [11] F. McSherry and I. Mironov. Differentially private recommender systems *Proc. KDD* 2009.
- [12] A. Beimel, P. Carmi, K. Nissim and E. Weinreb. Private Approximation of Search Problems *SIAM J. Computing*, vol. 38, no. 5, pp. 1728-1760, 2008
- [13] S. Kasiviswanathan, H. Lee, K. Nissim, S. Raskhodnikova and A. Smith. What can we Learn Privately? *SIAM J. Computing*, vol. 40, no. 3, pp. 793-826, 2011
- [14] L. Wassermann and S. Zhou. A Statistical Framework for Differential Privacy *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375-389, 2010
- [15] A. Ghosh, T. Roughgarden and M. Sundararajan Universally Utility Maximising Privacy Mechanisms *SIAM J. Computing* Vol. 41, No. 6, pp. 1673-1693, 2012.
- [16] J. Soria-Comas and J. Domingo-Ferrer Optimal Data-Independent Noise for Differential Privacy *Information Sciences*, vo. 250, pp. 200-214, 2013.
- [17] K. Parthasarathy *Probability Measures on Metric Spaces* AMS Chelsea Publishing, 2005 (Reprint).

- [18] R. Hall, L. Rinaldo and S. Wasserman Differential Privacy for Functions and Functional Data *Journal of Machine Learning Research*, Vol. 14, pp. 703-727, 2013.
- [19] C. Dwork. A Firm Foundation for Private Data Analysis *Communications of the ACM*, Vol. 54, No. 1, pp. 86-95, January 2011.
- [20] M. Hardt and K. Talwar. On the geometry of Differential Privacy In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pp. 705-714, 2010.
- [21] N. Li, T. Li, S. Venkatasubramanian t -Closeness: Privacy Beyond k -Anonymity and ℓ -Diversity In *Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE)*, 2007.
- [22] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkatasubramanian ℓ -Diversity: Privacy Beyond k -Anonymity *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007), 52 pages.
- [23] B. C. Fung, K. Wang, A. W. Fu, P. S. Yu Introduction to Privacy-Preserving Data Publishing Chapman and Hall, 2011.
- [24] D. Rebollo-Monedero, J. Forn and J. Domingo-Ferrer From t -Closeness-Like Privacy to Postrandomization via Information Theory *IEEE Transactions on Knowledge and Data Engineering*, Vol 22, No. 11, pp. 1623-1636, 2010.
- [25] F. Bonchi and E. Ferrari (Eds.) *Privacy-Aware Knowledge Discovery* Chapman and Hall, 2011.
- [26] A. Narayanan and V. Shmatikov Robust De-anonymization of Large Sparse Datasets In *2008 IEEE Symposium on Security and Privacy*.
- [27] W. Rudin. *Real and Complex Analysis* McGraw-Hill International Editions, 1987.
- [28] L. Sweeney. Forthcoming book entitled, *The Identifiability of Data* Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000
- [29] L. Sweeney. k -anonymity: A model for protecting privacy In *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Volume 10, Issue 5, 2002, pp. 557-570.