Information Sciences 305 (2015) 256–268

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof

Naoise Holohan [a], Douglas J. Leith [a], Oliver Mason [b,*]

[a] *School of Computer Science and Statistics, Trinity College Dublin, Ireland*
[b] *Hamilton Institute/Department of Mathematics & Statistics, Maynooth University-National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland*

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 19 June 2014
Accepted 27 January 2015
Available online 3 February 2015

Keywords:
Differential privacy
Metric space
Categorical data
Functional data
Data sanitisation

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.

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 [5], 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 [7]. 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.

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-identifiers[1] in place. While these anonymised data sets do indeed preserve participants' privacy in isolation, auxiliary/background information can make this technique extremely vulnerable to attack [19]. A study by Sweeney in 2000 [25] 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.

* Corresponding author. Tel.: +353 (0)1 7086274; fax: +353 5(0)1 7086269.
E-mail address: oliver.mason@nuim.ie (O. Mason).

[1] A 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.).

http://dx.doi.org/10.1016/j.ins.2015.01.021
0020-0255/© 2015 Elsevier Inc. All rights reserved.

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' identities. 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, Narayanan and Shmatikov [20] successfully de-anonymised entries in an anonymised data set containing movie ratings of 500,000 subscribers which was released by the 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. Anonymisation methods such as k-anonymity [26] and l-diversity [19] have been shown to be vulnerable to attacks based on background information [19,18].

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 [9] 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 [10]. 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 [16], while differentially private recommender systems are developed in [23]; in both of these instances, the data are naturally modelled as real numbers. Algorithms for search problems and learning are considered in [3,17]. A statistical perspective on differential privacy was developed in [27]; this paper considered real-valued ($[0,1]$ in fact) queries and data. In [13], mechanisms that maximise a suitable utility function were investigated; this paper assumed discrete finite-valued data spaces, which can describe categorical data. The recent paper [24] addressed the design of optimal mechanisms that add noise independently of the data; the queries considered are real-valued. To date, the only major reference on differential privacy for functional data appears to be [14]; 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 [15] and other similar papers on lower bounds for differentially private mechanisms considers real (or in some cases integer) valued data.

The principal aim in this paper is to develop a unifying, abstract framework for all major data and query types considered so far. It is not our intention to add to the considerable body of work done on differential privacy for specific data types or on output perturbations for specific query types (counting queries, linear queries, etc.). Rather, we identify the minimal technical requirements necessary for a discussion of differential privacy and initiate a programme of developing a theory based on these. Such an approach has a number of advantages. Isolating the core assumptions behind results provides insight into precisely why they hold and in some instances, proofs are clarified. Moreover results developed at an abstract level can be widely applied as their derivation is not tied to any one particular application domain. Throughout the paper, we emphasise this point with examples of real-valued, functional and categorical data.

Metrics have previously been used in the context of differential privacy in the papers [12,6], albeit in a different manner to that employed here. The standard definition of differential privacy (see Section 2.5 for details) measures similarity between databases using *Hamming distance*: namely the number of places or entries in which they differ. The work of these earlier papers considers generalisations of this in which arbitrary metrics may be used to quantify similarity of databases. It is worth noting that the results of Sections 3 and 4 may be readily extended to more general measures of similarity in the same spirit. However, our interest lies in analysing the standard definition of relaxed differential privacy and metrics are used here to quantify the error of a mechanism.

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 with specific examples of practical relevance that it can be seamlessly applied to numerical, categorical and functional data. A major advantage of such an abstract framework is that results derived in generality can be applied uniformly to a wide variety of different applications.
- Our description shows how mechanisms based on database sanitisations and output perturbations to query responses can be treated in a unified fashion. Moreover, the fundamental differences between these two classes of mechanisms can be seen clearly within the framework developed here (see the Remark after Theorem 4).
- We describe techniques for generating families of (ϵ, δ) differentially private mechanisms from simpler mechanisms. One such example is given by sanitised response mechanisms generated from an (ϵ, δ) differentially private database sanitisation. We also show how to generate differentially private sanitisations for high-dimensional databases using sanitisations for 1-dimensional databases. This approach provides a simple paradigm that can be applied to any type of data.