Information Processing Letters 129 (2018) 1–4

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

A lower bound on the release of differentially private integer partitions

Francesco Aldà *, Hans Ulrich Simon

Horst Görtz Institute for IT Security and Faculty of Mathematics, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany

ARTICLE INFO

Article history:
Received 10 August 2016
Received in revised form 14 June 2017
Accepted 1 September 2017
Available online 5 September 2017
Communicated by B. Doerr

Keywords:
Randomized algorithms
Differential privacy
Integer partitions

ABSTRACT

We consider the problem of privately releasing integer partitions. This problem is of high practical interest, being related to the publication of password frequency lists or the degree distribution of social networks. In this work, we show that any ε-differentially private mechanism releasing a partition of a sufficiently large non-negative integer N must incur a minimax risk of order $\Omega(\sqrt{N/\varepsilon})$. Moreover, for small values of N, we provide an optimal lower bound of order $\Omega(N)$.

1. Introduction

The trade-off between the protection of data privacy and the release of accurate data analysis has attracted a great deal of attention in recent years and has become one of the most challenging lines of research in statistics. Introduced in the seminal work of Dwork et al. [1], *differential privacy* has emerged as the de facto standard for privacy-preserving statistical analysis. It formally guarantees that the presence or absence of an input record does not significantly influence the aggregate statistics output, regardless of the adversary's side knowledge or computational power.

Since its introduction, differential privacy has seen a number of different applications. A very recent one deals with the private release of integer partitions [2]. Throughout this paper, whenever we write that a mechanism releases a partition of a non-negative integer N we always assume that the input is a partition of N but the output is allowed to be a partition of any non-negative integer. In their work, Blocki et al. [2] show that the exponential mechanism of McSherry and Talwar [3] can be used to release the partition of a non-negative integer N with L_1-error $O(\sqrt{N}/\varepsilon)$. Blocki et al. [2] also propose an approximate instantiation of the exponential mechanism which attains the same error bound but, in contrast to the latter, achieves computational efficiency by relaxing the privacy guarantees provided to (ε, δ)-differential privacy. The release of integer partitions is of high practical relevance. For instance, Blocki et al. [2] use their algorithm to publish a password frequency list from a password dataset of 70 million Yahoo! users [4]. In fact, a password frequency list is the partition of the number of passwords in a dataset. The security community is highly interested in such lists since they enable a better understanding of how passwords are chosen by users. Moreover, they can be used to accurately estimate security risks and design new password defenses. As observed by Blocki [5], integer partitions have other applications, besides password frequency lists. The degree distribution of a graph G with V vertices and E edges is a partition of the integer $2E$, and has been studied under various privacy models [6–8].

Besides their upper bound on utility, Blocki et al. [2] give empirical evidence that the error of their mechanism seems to scale with $1/\sqrt{\varepsilon}$ instead of $1/\varepsilon$ for large values of N. In this work, we demonstrate that this is ac-

* Corresponding author.
E-mail addresses: francesco.alda@rub.de (F. Aldà), hans.simon@rub.de (H.U. Simon).

http://dx.doi.org/10.1016/j.ipl.2017.09.001
0020-0190/

tually the best accuracy we can hope for, proving that any ε-differentially private mechanism which releases a partition of the integer N must incur a minimax risk of order $\Omega(\sqrt{N/\varepsilon})$ if $N \geq 1/(2\varepsilon)$. Moreover, we show that the bound becomes $\Omega(N)$ if $N < 1/(2\varepsilon)$. Since an ε-differentially private mechanism which always returns the partition of 0 incurs an L_1-error of at most N, our lower bound for small values of N is optimal.[1]

Standard techniques for proving lower bounds in differential privacy, e.g. packing [9] or information-theoretic [10] arguments, cannot be fully leveraged when dealing with integer partitions. The volume argument introduced by Hardt and Talwar [9], despite being applicable to related problems like releasing a noisy sorted sequence, does not seem to be appropriate for the special sequences that represent integer partitions. On the other hand, exploiting bounds on the mutual information [10] leads to results weaker than the ones which are obtained by more direct arguments. Our proof is based on a relatively simple application of Assouad's Lemma [11], a well-known tool for establishing lower bounds in statistics which has proved very useful in other differential privacy applications [12,13]. Despite its simplicity, it allowed us to significantly improve the result of Blocki [5], who recently showed a weaker lower bound of order $\Omega\left(\frac{\sqrt{N}}{\log N}\right)$. Moreover, Blocki [5] conjectured that for *sufficiently large* values of N the upper bound of Blocki et al. [2] could actually be improved to $O(\sqrt{N/\varepsilon})$. If this conjecture holds, then our lower bound would actually be tight in this regime, too.

2. Databases, integer partitions, differential privacy

Databases as histograms. A database is a collection of records from a (possibly infinite) universe $\mathcal{X} = \{x_1, x_2, \ldots\}$. In general $\mathcal{X}$ could be uncountable, but for the purpose of this paper it suffices to restrict our attention to countable universes only. Moreover, it is convenient to represent databases by their histograms, i.e., a database D over $\mathcal{X}$ is viewed as a sequence $(n_1, n_2, \ldots) \in \mathbb{N}^{\mathcal{X}}$ where n_i represents the number of elements in D of type x_i. Note that a database of size N is a sequence D of non-negative integers which sum up to N. Note that at most N members of the sequence D are non-zero. Throughout the paper, we identify a finite sequence $n_1, \ldots, n_k$ of non-negative integers with the infinite sequence $n_1, \ldots, n_k, 0, 0, \ldots$. Thus, databases are always given by infinite sequences of non-negative integers (even when the underlying universe $\mathcal{X}$ is finite). As usual, the L_1-distance between two databases D, D' is defined as $\|D - D'\|_1 = \sum_{i\geq 1} |n_i - n'_i|$. Two databases are said to be *adjacent* if their L_1-distance equals 1.

Integer partitions as sorted histograms. We define a *partition of an integer* N as a non-increasing sequence $f = (f_1 \geq f_2 \geq \ldots \geq f_N)$ such that $f_1, f_2, \ldots, f_N$ are non-negative integers that sum up to N. We will often identify a partition f of the integer N by an infinite sequence by setting $f_k = 0$ for all $k > N$. If $D = (n_i)_{i\geq 1}$ is a database of size N and, for $k = 1, \ldots, N$, f_k is the k-largest member of the sequence D, then we say that $f(D) = (f_1, \ldots, f_N)$ is the *partition of the integer* N that is *induced by* D. Note that the integer partition induced by D provides less information than D itself because it hides which member of the universe actually occurs most frequently (or second-most frequently and so on). In the sequel, $\mathcal{P}_N$ denotes the set of all partitions of the integer N and $\mathcal{P} = \bigcup_{N\geq 0} \mathcal{P}_N$.

Let D, D' be two databases and let $f(D), f(D')$ denote the corresponding integer partitions. The L_1-distance between D and D' is in general not the same as the L_1-distance between $f(D)$ and $f(D')$. For instance, the L_1-distance between $D = (1, 2)$ and $D' = (2, 1)$ is 2 but the corresponding integer partitions coincide, i.e., $f(D) = f(D') = (2, 1)$. The following result is quite obvious:

Proposition 1.

1. *Let $f(D)$ and $f(D')$ be the integer partitions that are induced by the databases D and D', respectively. Then $\|f(D) - f(D')\|_1 \leq \|D - D'\|_1$.*
2. *For every pair of integer partitions f, f', there exist databases D and D' such that D induces f, D' induces f' and $\|f - f'\|_1 = \|D - D'\|_1$.*

Definition 1 *([1]).* Let $\mathcal{X}$ be a universe and $\mathcal{R}$ be a (possibly infinite) set of responses. A mechanism $\mathcal{M} : \mathbb{N}^{\mathcal{X}} \to \mathcal{R}$ (meaning that, for every database $D \in \mathbb{N}^{\mathcal{X}}$, the response $\mathcal{M}(D)$ returned by $\mathcal{M}$ is an $\mathcal{R}$-valued random variable) is said to provide *ε-differential privacy* for $\varepsilon > 0$ if, for every pair (D, D') of adjacent databases and for every measurable $S \subseteq \mathcal{R}$, we have

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] \ .$$

It is well known that this implies that $\Pr[\mathcal{M}(D) \in S] \leq e^{d\varepsilon} \Pr[\mathcal{M}(D') \in S]$ for databases D, D' with L_1-distance d [1].

In this paper, we consider ε-differentially private mechanisms $\mathcal{M}_{N,\varepsilon} : \mathbb{N}^{\mathcal{X}} \to \mathcal{P}$ that take as input a database of size N and output an integer partition in $\mathcal{P}$. As observed before, each database D of size N induces an integer partition $f(D) \in \mathcal{P}_N$. We can then define the following risk function:

$$R(D, \mathcal{M}_{N,\varepsilon}) = \sum_{\hat{f} \in \mathcal{P}} \|\hat{f} - f(D)\|_1 \Pr[\mathcal{M}_{N,\varepsilon}(D) = \hat{f}] \ .$$

The minimax risk is then given by

$$R^{\star} = \inf_{\mathcal{M}_{N,\varepsilon}} \sup_{D \in \mathbb{N}^{\mathcal{X}} :\ \|D\|_1 = N} R(D, \mathcal{M}_{N,\varepsilon}) \ .$$

3. Some prerequisites from statistics

In this section, we remind the reader of some very general notions from measure theory, but for sake of simplicity we restrict our presentation to the discrete case only. This will hold throughout the rest of the paper.

Let P, Q be two discrete probability distributions over a space S. We say that P is absolutely continuous with

[1] In this paper, when we write optimal we always mean optimal modulo a constant factor.