Gradually Releasing Private Data under Differential Privacy

Fragkiskos Koufogiannis, Shuo Han, George J. Pappas
{fkouf, hanshuo, pappasg}@seas.upenn.edu*

January 22, 2015

1 Background and Problem Motivation

Aggregating individuals' data and computing statistics over a population are key ingredients to enable the Internet of Things [1]. Constructing traffic maps from individuals' GPS traces [2] and performing demand response in smart grids [3], [4] are two examples that involve such data aggregation. Using these statistics, individuals can perform their activities more efficiently; they may choose to avoid heavily congested routes or charge their electric vehicle during non-peak hours. However, accessing private data for the purpose of performing data aggregation has raised serious privacy concerns. An adversary can potentially extract information about individuals' data from aggregate statistics, especially when side information is available [5]. The framework of differential privacy was developed in order to mitigate these concerns and provide strong privacy guarantees [6], [7]. Given a desired privacy level, a noisy version of the aggregated value is publicly released to prevent an adversary from confidently extracting information about the private data.

For a fixed privacy level, [8] provides tools to build a private mechanism that approximates the desired aggregate quantity. Using these tools as primitives, applications of privacy-aware data aggregation have emerged [9], [10]. In these applications, the privacy level, parametrized by the constant $\epsilon \in [0, \infty)$, is assumed to be constant; parameter ϵ is a designer's choice and is set to a fixed value throughout the life of the aggregation system. Lower values of the parameter ϵ correspond to stronger privacy guarantees. Therefore, the value $\epsilon = 0$ translates to total privacy and the value $\epsilon = \infty$ means no privacy. Forever fixing the privacy level ϵ is a severe limitation. In practice, a varying privacy level can be useful as motivated by the following examples.

For instance, limited techniques exist for choosing a reasonable privacy budget ϵ. For small values of ϵ, substantial amounts of noise are injected and the performance of the resulting privacy-aware mechanism

*This work was supported in part by the TerraSwarm Research Center, one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP) a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

can be dramatically degraded. Thus, one strategy for choosing a privacy level is through a privacy-performance trade-off curve; the minimum privacy budget ϵ that achieves an acceptable performance is chosen. However, the characterization of this trade-off is impossible for real-life systems. One approach past this hurdle is booting the system with maximum privacy ($\epsilon = 0$) and relaxing the privacy budget until a desired performance is achieved. Another example of a varying privacy budget is a potential market of private data. Aggregating agencies initially access private data under ϵ_1 privacy guarantees, but they later decide to *"buy some more private data"*, relax privacy to a budget ϵ_2, and enjoy better accuracy. To the best of our knowledge, there is no previous approach on gradually releasing private data under differential privacy.

For completeness, we mention the setting of differential privacy under continuous observations which was first studied in [11]. In that setting, the privacy level remains fixed while more data are being released. This scenario is radically different than the one currently presented. In this work, both the private data and the quantity of interest are held fixed and the privacy level is varying.

Compositional theorems [8] provide a trivial, yet highly unsatisfying, approach to the aforementioned problems. Given an initial privacy budget ϵ_1, a noisy but privacy-preserving response y_1 is generated. Later, the privacy budget is increased to a new value ϵ_2 and a response y_2 is published. An adversary has possibly observed both values y_1 and y_2. Thus, compositional theorem suggests that only $\epsilon_1 + \epsilon_2$ privacy guarantees hold. Conversely, if an effective privacy level of value ϵ_2 were desired, the second response y_2 needs to be $\epsilon_2 - \epsilon_1$ private; thus, potentially noisier than the first response. In fact, [12] exploits compositional theorems and obtains tight bounds on the effective privacy level.

In this work, we prove that gradually releasing private data can be efficiently performed. Gradually relaxing the privacy level enables fine-tuning of the parameters of a privacy-aware system after it is bootstrapped or buying private data in multiple chunks. Moreover, we prove that releasing private data in multiple steps, instead of in a single step, does not incur any performance loss. This result is proven for multi-dimensional, isotropic identity queries and can be applied to many existing privacy-aware systems without additional modifications. Finally, we conjecture that gradually releasing private data is an intrinsic property of differential privacy.

2 Main Results

The problem explored in this work is that of gradually releasing data under differential privacy. Specifically, we are interested in designing composite mechanisms $(Q_{\epsilon_1}, \ldots, Q_{\epsilon_n})$, where $0 \leq \epsilon_1 \leq \ldots \leq \epsilon_n$, with the following properties:

- Each mechanism Q_{ϵ_i} should be an ϵ_i-private mechanism that *efficiently* approximates a query q.
- Any prefix $(Q_{\epsilon_1}, \ldots, Q_{\epsilon_i})$ of the composite mechanism should satisfy ϵ_i-privacy guarantees.

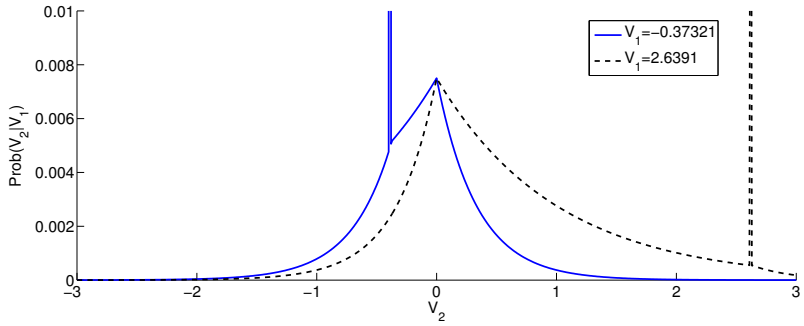

Figure 1: The distribution of each coordinate of V_2 conditioned on the value of the corresponding coordinate V_1 contains an atom and possesses a piece-wise exponential density. The privacy levels are $\epsilon_1 = 1$ and $\epsilon_2 = 2$.

The first result addresses identity queries in high-dimensional Euclidean spaces equipped with an l_1-norm adjacency relation. The version of differential privacy adapted to metric spaces [13] is used. This version is only slightly tighter than the original differential framework and, thus, results remain valid under the original framework. Specifically, Theorem 1 provides the probability distribution of additive noise samples that allow gradual release of private data.

Theorem 1. *Consider privacy levels ϵ_1, ϵ_2 with $\epsilon_2 \geq \epsilon_1 > 0$. Let Q_{ϵ_1} and Q_{ϵ_2} respectively be ϵ_1-private and ϵ_2-private mechanisms with respect to the metric space $\mathbb{R}^n_{l_1}$ that use oblivious additive noise:*

$$Q_{\epsilon_1} u = u + V_1 \text{ and } Q_{\epsilon_2} u = u + V_2, \text{ with } (V_1, V_2) \sim g \in \Delta\left(\mathbb{R}^{2n}\right), \tag{1}$$

where Δ denotes the set of probability measures. Then, the probability distribution $g = l^n_{\epsilon_1,\epsilon_2}$ is such that both Q_{ϵ_1} and Q_{ϵ_2} have optimal mean-squared-error:

$$l^n_{\epsilon_1,\epsilon_2}(v_1, v_2) = \prod_{i=1}^{n} \left\{ \frac{\epsilon_1^2}{2\epsilon_2} e^{-\epsilon_2|y_i|} \delta(x_i - y_i) + \frac{\epsilon_1(\epsilon_2^2 - \epsilon_1^2)}{4\epsilon_2} e^{-\epsilon_1|x_i - y_i| - \epsilon_2|y_i|} \right\}, \tag{2}$$

where $v_1 = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$ and $v_2 = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}$.

The probability density (2) allows sampling each coordinate independently. The first noise sample x_i is drawn from the Laplace distribution, whereas the second noise sample y_i is drawn according to the conditional distribution depicted in Figure 1. Coordinates V_1 and V_2 are marginally Laplace-distributed with parameters $\frac{1}{\epsilon_1}$ and $\frac{1}{\epsilon_2}$, respectively. Moreover, the mechanism that releases $(u + V_1, u + V_2)$ is ϵ_2-private, which is crucial for performing gradual release of private data.

The second result extends Theorem 1 to the case of relaxing privacy in multiple levels. In particular, Theorem 2 establishes that it is indeed possible to gradually release sensitive data in arbitrarily many steps:

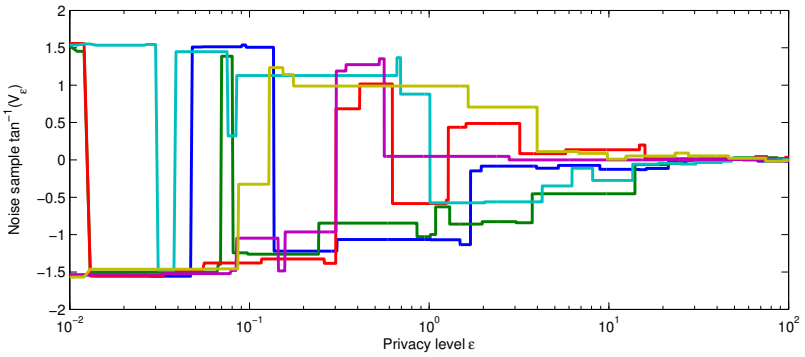

Figure 2: Gradual release of identity queries is achieved with the use of the stochastic process V_ϵ for $\epsilon \geq 0$. The component-wise noise samples are samples of this process. For tight values of privacy, high values of noise are returned, whereas, almost zero samples are returned for large values of ϵ. The process V_ϵ is Markov; future samples depend only on the current value of the process which eases implementation. Furthermore, the process is lazy; the value of the process is changed only a few times. Thus, only a small fraction of the dimensions are expected to have updated noise samples. Therefore, the communication burden of updating the privacy budget requires limited re-communications in a massive aggregation scheme. For a compressed illustration, the $\tan^{-1}$ of the noise value is drawn.

Theorem 2. *Consider m privacy levels $\{\epsilon_i\}_{i=1}^m$ with $0 < \epsilon_1 \leq \ldots \leq \epsilon_m$ and mechanisms Q_{ϵ_i} of the form:*

$$Q_i u = u + V_i, \text{ with } (V_1, \ldots, V_m) \sim g \in \Delta(\mathbb{R}^m) \tag{3}$$

where $(Q_1, \ldots, Q_i)$ is ϵ_i-private, for any $i \in \{1, \ldots, m\}$. Then, the probability distribution $g = l_{\epsilon_1,\ldots,\epsilon_m}$ has the property that each mechanism Q_i achieves optimal mean-squared-error:

$$l_{\epsilon_1,\ldots,\epsilon_m}(v_1, \ldots, v_m) = l_{\epsilon_1}(v_1) \prod_{i=1}^{m-1} \frac{l_{\epsilon_i,\epsilon_{i+1}}(v_i, v_{i+1})}{l_{\epsilon_i}(v_i)} \tag{4}$$

The probability distribution $l_{\epsilon_1,\ldots,\epsilon_m}$ is highly structured. In fact, the distribution establishes that gradually releasing private data can be performed in a Markov fashion. Initially, the first noise sample V_1 is drawn from the Laplace distribution. Subsequently, the distribution of each noise sample V_i is fully specified by the targeted privacy level ϵ_i and the last noise sample V_{i-1} and privacy level ϵ_{i-1}. The typical form of the conditional distribution is shown in Figure 1. Therefore, there is no computational complexity incurred by the number of steps. The owner of the sensitive data needs to store only the most recently released noise sample and the corresponding privacy level. In fact, neither the exact number of steps nor the future values of privacy levels are required a priori. Form another point of view, these properties are exactly the Markov and the non-anticipating property of the stochastic process V_ϵ, which is plotted in Figure 2.

References

[1] Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer networks*, 54(15):2787–2805, 2010.

[2] Daniel B Work, Sébastien Blandin, Olli-Pekka Tossavainen, Benedetto Piccoli, and Alexandre M Bayen. A traffic model for velocity data assimilation. *Applied Mathematics Research eXpress*, 2010(1):1–35, 2010.

[3] Pedram Samadi, A Mohsenian-Rad, Robert Schober, Vincent WS Wong, and Juri Jatskevich. Optimal real-time pricing algorithm based on utility maximization for smart grid. In *IEEE International Conference on Smart Grid Communications*, 2010.

[4] Na Li, Lijun Chen, and Steven H Low. Optimal demand response based on utility maximization in power networks. In *IEEE Power and Energy Society General Meeting*, pages 1–8, 2011.

[5] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

[6] Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, 2006.

[7] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, 2008.

[8] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science*, 2007.

[9] Michael Backes and Sebastian Meiser. Differentially private smart metering with battery recharging. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 194–212. Springer, 2014.

[10] Fragkiskos Koufogiannis, Shuo Han, and George J Pappas. Computation of privacy-preserving prices in smart grids. In *IEEE Conference on Decision and Control*, 2014.

[11] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 715–724. ACM, 2010.

[12] Jason Reed and Benjamin C Pierce. Distance makes the types grow stronger: a calculus for differential privacy. In *ACM Sigplan Notices*, 2010.

[13] Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In *Privacy Enhancing Technologies*, pages 82–102. Springer, 2013.