Optimal Distribution of Privacy Budget in Differential Privacy

Anis Bkakria, Nora Cuppens, and Frédéric Cuppens

IMT Atlantique
{anis.bkakria,nora.cuppens,frederic.cuppens}@imt-atlantique.fr

Abstract. In this paper, we first study the problem of privacy budget distribution in adaptive multi-data consumers (i.e., users) differential privacy use cases. Then, we present an extension of the classic differential privacy formal model that allows taking into consideration data consumers' information disclosure risk when distributing the privacy budget among data consumers. Finally, we define a method allowing to optimally distribute a given privacy budget among a private database's data consumers.

Keywords: Differential Privacy, Privacy Budget Distribution, Information disclosure risk

1 Introduction

In the last decade, a new paradigm called *differential privacy* has emerged as a new formal model that ensures a more robust privacy guarantees, regardless prior knowledge an adversary may possess [3]. Differential privacy model guarantee that given two databases that differ exactly in the information of a single individual *ind* (the two databases differ exactly on the record that contains the information of *ind*), a differential private data analysis mechanism will output, for the two databases, randomized results with almost identical probability distributions. Hence, regardless how much he/she knows about the other records in the database, an adversary who sees the result of the performed private analysis will not be able to guess with high confidence the database over which the private analysis was performed. Therefore, the adversary cannot guess with high confidence whether *ind* is present in the database.

Differential privacy's strong privacy guarantee comes at the price of data consumers' (e.g., individuals or entities that are going to perform data analysis over the private database) queries and analysis responses precisions. This trade-off between the level of ensured privacy and queries' responses precisions is represented in the differential privacy model through the parameter ϵ. A smaller value of ϵ means strong privacy guarantee and low queries' response precision.

The last five years have seen several papers [1, 2, 6–8, 11, 12] that study the trade-off between privacy and utility (precision) in differentially private mechanisms for different kinds of queries (e.g., counting queries [2, 7, 11], histogram

queries [12], marginal queries [1], etc). The aforementioned approaches tried to design new differential private mechanisms that allow either to enhance the precision of the responses to specific kind of queries or to reduce the quantity of privacy budget to be consumed for each query (i.e., increase the total number of queries that can be performed over the database).

Although the differential privacy model has drawn attention in quite a few areas and despite the over a hundred papers on differential privacy that are published from the security, database, machine learning, and statistics communities, some open problems remain untackled. The most obvious is how to optimally distribute the total privacy budget that can be consumed over the private database among the set of data consumers. To the best of our knowledge, all existing approaches and their developed solutions such as PINQ [9] and Airavat [10] suppose that all data consumers that can query the private database share the same privacy budget (i.e., the total budget specified by the data owner). This configuration will allow a data consumer to consume more privacy budget than other data consumers. In the case of a malicious data consumer, he/she can prevent others to query the private database by consuming the total privacy budget.

In this paper, we present an approach that extends the classic differential private model to optimally distribute the total privacy budget to be consumed over a database among the data consumers that will be allowed to query the private database. The idea of our approach is to use, for each data consumer, the risk that he/she will publish or disclose the information he/she will learn from the private database to optimally distribute the privacy budget.

2 Background on differential privacy

Informally speaking, an algorithm is differential private if a small change in its inputs does not modify considerably its outputs. Differential privacy is formalized as follows.

Definition 1 (ϵ-differential privacy [5]). *An mechanism $\mathcal{M}$ is ϵ-differentially private if for all input database d, any $d' \in \mathcal{D}^d$ and any subset of outputs $S \in Range(\mathcal{M})$, the following condition holds:*

$$Pr[\mathcal{M}(d) \in S] \leq exp(\epsilon) \times Pr[\mathcal{M}(d') \in S]$$

where $\mathcal{D}^d$ is the set of d's neighboring databases, each differing from d by at most one record and the probability is taken over the randomness of the $\mathcal{M}$.

The previous definition states that any data consumer who will observe the result of the execution of $\mathcal{M}$ over d cannot guess the presence of an individual in d with more than $100 \times (|1 - 1/exp(\epsilon)|)\%$ of confidence.

Differential privacy formal model allows computing the level of ensured privacy after performing a set of queries on a same database d.

Theorem 1 (Mechanism Composition [4]). *Given a set of k mechanisms $\mathcal{M}_1, \cdots, \mathcal{M}_k$ such that each $\mathcal{M}_i$ is ϵ_i-differential private, $i \in [1, k]$. Then, any mechanism $\mathcal{M}$ that is a composition of $\mathcal{M}_1, \cdots \mathcal{M}_k$ is $\sum_{i=1}^{k} \epsilon_i$-differential private.*

Note that in the previous definition, each mechanism can be considered as the differential private execution of a query or analysis over the d. So, if d's data curator wants to allow a data consumer to execute a set of queries $q_1, \cdots, q_k$ using the mechanisms $\mathcal{M}_1, \cdots, \mathcal{M}_k$, he/she needs to assign a privacy budget greater or equal to $\sum_{i=1}^{k} \epsilon_i$ to the data consumer.

3 Problem Statement

Let us consider that the total privacy budget for a medical database d specified by a hospital is ϵ^t. Let us also suppose that d will be used by three data consumers: A data scientist u_i working in an insurance company, a data scientists u_h working in the hospital, and a researcher u_r. The three data consumers want to perform interactively a set of queries. That is, the hospital does not know in advance the set of queries to be performed by each data consumer. So, the main question here is how to manage the usage of ϵ_t by the three data consumers?

One trivial solution is to share ϵ_t between the three data consumers. However, This solution will allow a data consumer to consume more privacy budget than others. In the worst case, he/she can prevent others to query the private database by consuming the total privacy budget (e.g., by performing sequentially a high privacy budget consuming query many time).

To avoid the previous problem, the hospital can try to distribute the total privacy budget between the data consumers. In this case, due to the fact that differential privacy's adversary model supposes that the risk/probability that (i) each data consumer will disclose the information he/she learned about d is equal to 1, ϵ_t should be distributed as follows:

$$\epsilon_{u_h} + \epsilon_{u_i} + \epsilon_{u_r} = \epsilon_t \tag{1}$$

Despite that the previous formula represents a privacy budget distribution condition, it does not specify how much privacy budget the hospital should give to each data consumer. Moreover, assumption (i) is too strong and even not valid in our case, since logically, u_i and u_h have less probability than 1 to disclose the information they learned. We strongly believe that by quantifying and taking into consideration data consumers disclosure probabilities in the differential privacy model, we could have a better distribution of the privacy budget over the data consumers that are authorized to query d.

4 Proposed Solutions

Our solution extends the differential privacy model for optimal privacy budget assignment among data consumers. That is, instead of reasoning about the in-

formation leaked to all data consumer, our approach consists of modeling the leaked information for each data consumer separately. The following definition quantifies the quantity of leaked information to each data consumer.

Definition 2. *Given a data consumer u of a private database D and its privacy budget ϵ_u. Suppose that u queries D by using a differential private mechanism $\mathcal{M}$, then the following condition holds:*

$$Pr^u[\mathcal{M}(D) \in S] \leq e^{\epsilon_u} Pr^u[\mathcal{M}(D') \in S] \tag{2}$$

where D' and D are adjacent, $S \subseteq$ Range($\mathcal{M}$), and $Pr^u[c]$ denotes the probability that c holds from u's perspective

Now, by considering the risk/probability that each data consumer will share or disclose the information he learned about the private database, we can compute, as shown in the following theorem, the probability that an adversary will learn all the information that has been released to data consumers through their performed queries.

Theorem 2 (disclosure risk-based differential privacy). *Given a set of data consumers $\mathcal{U}$ of a private database D and p_u representing the risk/probability that the data consumers $u \in \mathcal{U}$ will share or disclose the information they are going to learn about D to other parties. If we suppose that for each data consumer $u_i \in \mathcal{U}$ is attributed a privacy budget ϵ_{u_i}, then, in the worst case, the following condition holds:*

$$Pr\left[\exists \mathcal{A}, \forall \mathcal{U}' \subseteq \mathcal{U} : Pr^{\mathcal{A}}[\mathcal{M}(D) \in S] = exp(\sum_{u \in \mathcal{U}'} \epsilon_u) Pr^{\mathcal{A}}[\mathcal{M}(D') \in S]\right] \leq \prod_{u \in \mathcal{U}'} p_u \tag{3}$$

where D and D' are adjacent databases, and $\mathcal{A}$ is an adversary.

In our solution, we suppose that the data owner (or the data collector) will specify for each total privacy budget value (i.e., the value of disclosed information), the value of the maximum acceptable disclosure risk/probability level.

Definition 3 (α-risky privacy budget distribution). *Given a set of data consumers $\mathcal{U} = \{u_1, \cdots, u_n\}$ having each a disclosure probability p_{u_i}. We say that the privacy budget distribution function dist-budget: $\mathcal{U} \rightarrow \mathbb{R}$ is α-risky iff the following condition hold:*

$$\forall \mathcal{U}' \subseteq \mathcal{U} : \prod_{u \in \mathcal{U}'} p_u \leq \alpha \left(\sum_{u \in \mathcal{U}'} \textit{dist-budget}(u) \right) \tag{4}$$

where $\alpha : \mathbb{R} \rightarrow \mathbb{R}$.

The function α is going to be used by the data owner to indicate for each value of disclosed information, the value of the maximum acceptable disclosure risk/probability level. An example of the function definition could be:

$$\alpha(\epsilon^*) = \begin{cases} 1 & if \quad \epsilon^* \leq \epsilon \\ 10^{-\frac{\epsilon^* - \epsilon}{\epsilon}} & if \quad \epsilon^* > \epsilon \end{cases}$$

where ϵ is the quantity of information that an adversary can learn when the disclosure risk is 1 (i,e,. the quantity of information that an adversary can learn in the classic differential privacy model). In this example, the data owner simply requires that the acceptable disclosure probability should decrease exponentially in the amount of increased disclosed information.

Note that Theorem 2 and Definition 3 are directly related through the right side of the inequality (3) and the left side of the inequality (4). Informally speaking, given a set of data consumers $\mathcal{U}'$ and a privacy budget distribution function *dist-budget* that assigns for each data consumer a privacy budget to use for querying the private database, *dist-budget* is alpha risky if the probability that there exists an adversary $\mathcal{A}$ that knows the informations that has been released to data consumers in $\mathcal{U}'$ through their performed queries is less or equal to the maximum acceptable disclosure risk level defined by the data owner.

We now define our method for optimal privacy budget assignment among data consumers. This method will be based mainly on the data owner's trade-off between data consumers' disclosure probability/risks and the quantity of disclosed information, which we presented in Definition 3.

to meet the optimality in privacy budget sharing, we should maximize to the best the privacy budget to be attributed to each data consumer while ensuring the satisfaction of the data owner's trade-off between data consumers' disclosure probability/risks and the quantity of disclosed information. This can be formalized as follows.

Definition 4. *Given a set of data consumers $\mathcal{U} = \{u_1, \cdots, u_n\}$ having each a disclosure probability p_{u_i} and a function $\alpha : \mathbb{R} \to \mathbb{R}$ that specifies for each value of disclosed information, the value of the acceptable disclosure probability. An optimal privacy budget assignment is the solution to the following maximization problem:*

$$\begin{aligned} \underset{u \in \mathcal{U}}{Maximize} \quad & \textit{dist-budget}(u) \\ s.t. \quad & \forall u \in \mathcal{U} : \epsilon_u > 0 \\ & \forall \mathcal{U}' \subseteq \mathcal{U} : \prod_{u \in \mathcal{U}'} p_u \leq \alpha \left(\sum_{u \in \mathcal{U}} \textit{dist-budget}(u) \right) \\ & \forall u_1, u_2 \in \mathcal{U} : (p_{u_1} - p_{u_2}) \times (\epsilon_{u_1} - \epsilon_{u_2}) \leq 0 \end{aligned} \tag{5}$$

In the previous definition, the first condition states that all data consumers of the private database should have a privacy budget greater than zero. The second condition ensures that given a set of data consumers in $\mathcal{U}$, the probability that all of them will disclose the information they learn is less or equal to the disclosure

threshold specified by the function α. Finally, the last condition ensures that if a data consumer u_1 has a disclosure probability greater (respectively, lesser) than the disclosure probability of a data consumer u_2, then the privacy budget to be attributed to u_1 should be lesser or equal (respectively, greater or equal) than the privacy budget to be attributed to u_2.

5 Conclusion

This paper proposes a solution for the problem of privacy budget distribution in adaptive multi-data consumers differential privacy use cases. The solution extends the differential privacy model classic model to include data consumers information disclosure risk, and define a maximization objective function that ensures an optimal privacy budget distribution among data consumers. As a future work, we aim to define a method for computing the information disclosure risk of a data consumer and to implement and evaluate our approach on a real use case.

References

1. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K.: 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. ACM (2007)
2. Cormode, G., Procopiuc, M., Srivastava, D., Tran, T.T.: Differentially private publication of sparse data. arXiv preprint arXiv:1103.0825 (2011)
3. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: TCC. vol. 3876, pp. 265–284. Springer (2006)
4. Dwork, C., McSherry, F., Nissim, K., Smith, A.D.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings. pp. 265–284 (2006), http://dx.doi.org/10.1007/11681878_14
5. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science 9(3–4), 211–407 (2014)
6. Hardt, M., Talwar, K.: On the geometry of differential privacy. In: Proceedings of the forty-second ACM symposium on Theory of computing. pp. 705–714. ACM (2010)
7. Hay, M., Rastogi, V., Miklau, G., Suciu, D.: Boosting the accuracy of differentially private histograms through consistency. Proceedings of the VLDB Endowment 3(1-2), 1021–1032 (2010)
8. Li, C., Hay, M., Rastogi, V., Miklau, G., McGregor, A.: Optimizing linear counting queries under differential privacy. In: Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 123–134. ACM (2010)
9. McSherry, F.D.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. pp. 19–30. ACM (2009)

10. Roy, I., Setty, S.T., Kilzer, A., Shmatikov, V., Witchel, E.: Airavat: Security and privacy for mapreduce. In: NSDI. vol. 10, pp. 297–312 (2010)
11. Xiao, X., Wang, G., Gehrke, J.: Differential privacy via wavelet transforms. IEEE Transactions on Knowledge and Data Engineering 23(8), 1200–1214 (2011)
12. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G., Winslett, M.: Differentially private histogram publication. The VLDB Journal 22(6), 797–822 (2013)

Appendix A Proof of Theorem 2

Proof. Since each data consumer $u \in \mathcal{U}$ has a privacy budget ϵ_u, so in the worst case, we have:

$$\forall u \in \mathcal{U}' : Pr^u[\mathcal{M}(D) \in S] = e^{\epsilon_u} Pr^u[\mathcal{M}(D') \in S] \tag{6}$$

Let us suppose that each data consumer $u_i \in \mathcal{U}'$ performed the set of queries $\mathcal{Q}_{u_i} = \{q_1^{u_i}, \cdots, q_{n_i}^{u_i}\}$ over the database D using a differential private mechanism $\mathcal{M}$ and got the set of outputs $\mathcal{Y}_{u_i} = \{y_1^{u_i}, \cdots, y_{n_i}^{u_i}\}$. Then, in the worst case, the information learned by the data consumer u_i can be quantified as following:

$$L_{u_i}^{D \to D'}(\mathcal{Y}_{u_i}) = \ln \left(\frac{Pr[\mathcal{M}(q_1^{u_i}(D)) = y_1^{u_i} \wedge \cdots \wedge \mathcal{M}(q_{n_i}^{u_i}(D)) = y_{n_i}^{u_i}]}{Pr[\mathcal{M}(q_1^{u_i}(D')) = y_1^{u_i} \wedge \cdots \wedge \mathcal{M}(q_{n_i}^{u_i}(D')) = y_{n_i}^{u_i}]} \right) = \epsilon_{u_i} \tag{7}$$

Let us now suppose that all data consumers in $\mathcal{U}'$ disclose the set of queries they performed and the set of responses they got to an adversary $\mathcal{A}$. Then, in the worst case, the information that can be learned by the adversary $\mathcal{A}$ can be quantified as following:

$$\begin{aligned} L_{u_i}^{D \to D'} \left(\bigcup_{u \in \mathcal{U}'} \mathcal{Y}_u \right) &= \ln \left(\prod_{u_i \in \mathcal{U}} \left(\frac{\prod_{j=1}^{n_i} Pr[\mathcal{M}(q_j^{u_i}(D)) = y_j^{u_i}]}{\prod_{j=1}^{n_i} Pr[\mathcal{M}(q_j^{u_i}(D')) = y_j^{u_i}]} \right) \right) \\ &= \sum_{u_i \in \mathcal{U}'} L_{u_i}^{D \to D'}(\mathcal{Y}_{u_i}) \\ &= \sum_{u_i \in \mathcal{U}'} \epsilon_{u_i} \end{aligned} \tag{8}$$

Now, the probability that there exists an adversary who knows the set of queries performed by all data consumers in $\mathcal{U}'$ and the set of responses to those queries (an adversary how learns $\sum_{u_i \in \mathcal{U}'} \epsilon_{u_i}$ information about individual in the database D) can be computed as following:

$$\begin{aligned} Pr\left[\bigwedge_{u_i \in \mathcal{U}'} \left(\bigwedge_{j=1}^{n_i} disclose^{u_i}(q_j^{u_i}, y_j^{u_i}) \right) \right] &= \prod_{u_i \in \mathcal{U}} Pr\left[\bigwedge_{j=1}^{n_i} disclose^{u_i}(q_j^{u_i}, y_j^{u_i}) \right] \\ &= \prod_{u_i \in \mathcal{U}} p_{u_i} \end{aligned} \tag{9}$$

where $disclose^u(q, y)$ means the disclosure of the query q and its response y by the data consumer u.

Finally, based on equations 6, 8 and 9 we can deduce equation 3.