Differential Privacy for Black-Box Statistical Analyses

Nitin Kohli [1] **Paul Laskowski** [1]

Abstract

We formalize a notion of a privacy wrapper, defined as an algorithm that can take an arbitrary and untrusted script and produce an output with differential privacy guarantees. Our novel privacy wrapper design leverages subsets of data and halts to overcome the issue of unknowable sensitivity.

1. Introduction

Consider a scenario involving a holder of personal data and a researcher. The researcher has written an analysis script and would like to apply it to the data in order to gain insight. The data holder is concerned with the privacy of individuals in the data, and may not necessarily trust the researcher. Furthermore, the data holder may have contractual obligations that prevent them from sharing information about individuals in the data. The researcher gives the data holder their analysis script, but the data holder does not have the resources or expertise to analyze the code for privacy threats. Instead, the data holder desires a simple way to add privacy guarantees to the researcher's script, without looking inside it, so that the output can be safely returned. This motivates the central question of this study:

> What privacy guarantees can be added to a statistical analysis script from an untrusted source, treating it in a black box fashion?

Although many organizations are interested in increasing researcher access to data, the costs given current technology can be considerable. For example, the US Census Bureau has a program to allow researchers to interact with raw data. To maintain privacy, however, the Bureau maintains a set of secure locations around the country and puts all applicants through a background check. In another example, Facebook and Social Science One teamed up to make data about shared URLs available to researchers. However, to work with the most sensitive data, researchers must be pre-approved by a university review board, pass an application process, and undergo monitoring - even at the level of individual keystrokes. The project is also exploring a solution based on differential privacy, but this would involve restricting researchers to a small set of commands, reducing the flexibility of the system.

In contrast with existing approaches, the goal of this study is to enable a researcher to work with private data in an automatic fashion, without the need for screening procedures to establish trust. We will refer to a system that achieves this as a *privacy wrapper*. Akin to a function wrapper in programming, the idea is to write an algorithm that mediates all interaction with the researcher's script, producing output that is based on the behavior of the script, while also yielding strong privacy guarantees.

1.1. Considerations for Untrusted Code

A number of stylized observations will guide our design of a privacy wrapper. First, one possible strategy involves analyzing the code submitted by the researcher to understand its privacy properties under any possible dataset, then tailoring noise to match the results of this analysis. For example, one might hope to detect if a researcher script is already differentially private, in which case the output can be returned to the researcher immediately. While this may be possible for specific scripts, it is unfortunately not possible in general. In the language of the theory of computation, privacy properties such as differential privacy are semantic and non-trivial. Rice's theorem tells us that the problem of deciding if arbitrary code meets such properties is undecidable. Instead, we must follow a different strategy for constructing a privacy wrapper, treating it as a black box.

Second, to privatize functions with real-valued outputs, researchers usually proceed by analyzing the function's *sensitivity*. In particular, the local sensitivity of a researcher script A at a dataset D is defined as the maximum amount the output may change in response to a change in one row of the data. Unfortunately, when treating a researcher script as a black box, and assuming the set of possible data entries is too large for an exhaustive search, there is no way to estimate local sensitivity. This suggests that straightforward sensitivity-based approaches are unlikely to work.

Finally, an adversary may design a script that changes its

[1]UC Berkeley School of Information. Correspondence to: Nitin Kohli <nitin.kohli@berkeley.edu>, Paul Laskowski <paul@ischool.berkeley.edu>.

behavior when a secret condition is met. For example, suppose that that an adversary is interested in a specific target, t, which may or may not be in the data. The adversary writes a script that outputs 0 unless t is in the data, in which case the script outputs some large $U >> 1$. If t is not in the data, there is no way for the wrapper to detect that the researcher script is capable of outputting any value other than 0, at least in all cases. Moreover, whatever the behavior of wrapper is when t is not in the data, it cannot change substantially when t is present in the data, or this would give away the secret information. This suggests that if the wrapper approximates the output of researcher script on the true data in some cases, it must sometimes "pretend" that certain individuals are not in the data.

These observations motivate the design of our privacy wrapper, which treats the researcher script as a black box and uses subsets of data to ensure a type of sensitivity bound. Our privacy wrapper has the following properties:

1. **Privacy:** Given any researcher script, the mechanism generated by the wrapper meets *differential privacy*.
2. **Accuracy:** The output of the wrapper is related to the researcher script in the sense that it is found by adding (predetermined) noise to the output of the script on *some* subset of data.
3. **Flexibility:** Our privacy wrapper places no limitation on the code written by a researcher. Any script that returns a real number can be used.

A disadvantage of our approach is that our wrapper requires exponential time to run. We hope to address this limitation in future research.

2. Related Works

Differential privacy was introduced by Dwork et al. as a rigorous standard for mechanisms that compute real-valued statistics from personal data (Dwork et al., 2006). The authors pioneered privacy analysis based on global sensitivity, which is defined as the maximal change in a statistic resulting from a one row change to any dataset. A number of studies have since developed differentially private mechanisms leveraging local sensitivity, which is often much smaller than global sensitivity (Nissim et al., 2007; Johnson et al., 2018; Vadhan, 2017). Such approaches cannot immediately be applied to black box researcher scripts, as neither local nor global sensitivity may be estimated.

Starting with Papernot et al., privacy researchers have followed a strategy of partitioning a dataset into pieces, then aggregating the results through noisy voting (Papernot et al., 2016; Jordon et al., 2018). Such algorithms treat a researcher script as a black box, and apply as long as the return value is categorical, rather than real-valued. Unlike studies in this lineage, we focus on real-valued statistics.

In a closely related study, Dwork and Lei develop tools for privatizing robust statistics (Dwork & Lei, 2009). The statistics they consider are white boxes, but they have unbounded global sensitivity in some cases. The authors' demonstrate that such statistics can still be privatized if an algorithm is allowed to halt when a certain sensitivity bound is breached. In our work on black boxes, we also resort to halting as a way to overcome the problem of unknowable sensitivity.

A different approach to expanding access to data involves presenting researchers with a restricted language for writing their scripts. An example is PINQ, which presents programmers with a SQL-like interface with privacy guarantees (McSherry, 2009). Kiefer et al. propose an architecture in which access to data is mediated by a privacy layer that implements differentially private mechanisms (Kifer et al., 2020).

3. Basic Definitions

Let $\mathbb{D}$ be the set of possible entries that represent one individual in a dataset. A dataset is represented as a multiset of finite size with entries in $\mathbb{D}$. Let $\mathbb{D}^\star$ be the set of all possible datasets. We say two datasets $D_1, D_2 \in \mathbb{D}^\star$ are neighbors if one can be obtained from the other by switching exactly one element.

We define an algorithm as a function $A : \mathbb{D}^\star \to \Delta(\mathbb{R} \cup \{\perp\})$. Here, $\perp$ is used to represent halt, meaning an algorithm fails to return a real value. For convenience, we define a researcher's script as a deterministic function $R : \mathbb{D}^\star \to \mathbb{R} \cup \{\perp\}$. All results in this paper can be extended to researcher scripts that are randomized by adding a sampling step to the wrapper. Let $\mathcal{R}$ be the set of all deterministic algorithms.

Given any two distributions $l_1, l_2 \in \Delta(\mathbb{R} \cup \{\perp\})$, we say these distributions are (ϵ, δ)-indistinguishable if for every measurable set $E \subseteq \mathbb{R} \cup \{\perp\}$, for any $i, j \in \{1, 2\}$, $l_i(E) \leq \exp(\epsilon) l_j(E) + \delta$.

We will denote the (ϵ, δ)-indistinguishability of l_1 and l_2 as $l_1 \sim_{\epsilon,\delta} l_2$. If l_1, l_2, and l_3 are distributions, then the following transitive property holds: If $l_1 \sim_{\epsilon_1,\delta_1} l_2$, and $l_2 \sim_{\epsilon_2,\delta_2} l_3$, then $l_1 \sim_{\epsilon_1+\epsilon_2,\delta_1+\delta_2} l_3$. (Dwork & Roth, 2014)

An algorithm, $A : \mathbb{D}^\star \to \Delta(\mathbb{R} \cup \{\perp\})$ is (ϵ, δ)-differentially private if for any neighboring datasets D_1 and D_2, $A(D_1)$ and $A(D_2)$ are (ϵ, δ)-indistinguishable.

A wrapper is a function, $W : \mathbb{D}^\star \times \mathcal{R} \to \Delta(\mathbb{R} \cup \{\perp\})$. We will say that a wrapper W *imposes* $(\epsilon, \delta)-$differential privacy if for any researcher algorithm R, the wrapper $W(\cdot, R)$ is $(\epsilon, \delta)-$differentially private. Throughout this manuscript,

we will use the more compact W_D to denote the probability distribution $W(D, R)$ when R is understood.

4. Algorithm Description

Our privacy wrapper W is based on the following definitions. Let multiset $S \in \mathbb{D}^\star$ be called $(\alpha, \lambda, T, R)-$stable if for every $X, Y \subseteq S$ of size at least T, we have $R(X), R(Y) \in \mathbb{R}$, and the two distributions $L(R(X), \lambda)$ and $L(R(Y), \lambda)$ are $(\alpha, 0)$-indistinguishable. Because α, λ, T, and R will not change in this study, we will omit them and simply refer to the set S as stable. It is worth noting that under this definition, if R yields $\perp$ for any subset of S of size at least T, then S is not stable. Further, we define $L(\mu, \lambda)$ be the Laplace distribution with mean μ and scale parameter λ.

Our wrapper may not use every element in the dataset $D \in \mathbb{D}$ when computing a private version of $A(D)$. Throughout this paper, we will use N to represent the size of a dataset, and M to represent the maximum number of entries that a privacy wrapper may exclude. To help select a size between $N - M$ and N, we define the distribution $G(\epsilon, \alpha, N, M)$ as follows: $G(\epsilon, \alpha, N, M)(n)$ equals

$$\delta \exp\left(\min\left\{(\epsilon - 4\alpha)(n - N + M) - 2\alpha, \epsilon(N - n)\right\}\right)$$

for n such that $N - M \leq n \leq N$ and 0 otherwise, where δ is a constant selected so that the total probability sums to 1. We will normally omit the parameter arguments to G for readability purposes.

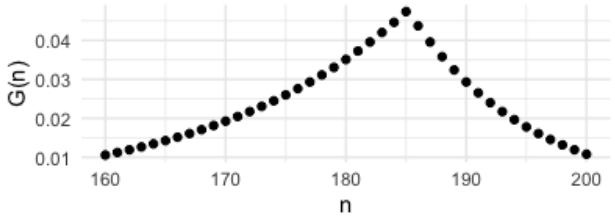

Figure 1. The distribution $G(n)$ when $N = 100, M = 40, \epsilon = 0.1$, and $\alpha = 0.01$. For these values, $\delta \approx 0.01$.

Our privacy wrapper W takes as input a researcher script $R : \mathbb{D}^\star \to \Delta(\mathbb{R} \cup \{\perp\})$ and a dataset $D \in \mathbb{D}^\star$ of size N. Additionally, Algorithm W uses the following auxiliary parameters: $\alpha, \lambda, \epsilon \in \mathbb{R}$ such that $\alpha \geq 0$, $\lambda > 0$, $\epsilon > 4\alpha$, and $M, T \in \mathbb{N}$ such that $T \leq N - 2M$.

Algorithm W:

1. Choose n from distribution G.
2. If the set of stable subsets of size n is non-empty, then select an element S at random, and output a draw from $L(R(S), \lambda)$. Otherwise, output $\perp$.

5. Analysis of Algorithm

Theorem 1. *Given $\epsilon > 4\alpha$, Algorithm W imposes (ϵ, δ)-differential privacy, where δ is the normalization factor from the distribution G.*

To prove Theorem 1, we show that for any measurable set $E \subseteq \mathbb{R} \cup \{\perp\}$, and any neighboring datasets D_1 and D_2,

$$\frac{W_{D_1}(E) - \delta}{W_{D_2}(E)} \leq \exp(\epsilon)$$

If there are no stable subsets of size $N - M$ of either D_1 or D_2, then W_{D_1} and W_{D_2} are the same distribution (giving probability 1 to $\perp$), so the bound follows immediately. If there are any stable subsets of size $N - M$ of either D_1 or D_2, choose one and call it K. Define reference distribution $H = L(R(K), \lambda)$. For any $D \in \mathbb{D}^\star$, write $W_D(\cdot|n)$ to represents the conditional probability distribution of the wrapper when n has been chosen in step 1.

Lemma 1. *H is $(2\alpha, 0)-$indistinguishable from $W_{D_1}(\cdot|n)$ and $W_{D_2}(\cdot|n)$ for any n when these algorithms do not halt.*

Proof. For $i \in \{1, 2\}$, if $W_{D_i}(\cdot|n)$ does not halt, then it is a mixture of Laplace distributions of the form $L(R(S), \lambda)$, where $S \subseteq D_i$ is a stable subset of size $n \geq N - M$. So it is sufficient to show that every distribution in the mixture is 2α-indistinguishable from H.

Since $|S| \geq N - M$, $|S \cap K| \geq N - 2M \geq T$. Also, because $S \cap K$ is a subset of a stable subset of size at least T, it is also stable. Because S is stable, and $S \cap K$ is a subset, we have $L(R(S), \lambda) \sim_{\alpha,0} L(R(S \cap K), \lambda)$. Because K is stable, and $S \cap K$ is a subset, $L(R(K), \lambda) \sim_{\alpha,0} L(R(S \cap K), \lambda)$ So then $L(R(S), \lambda) \sim_{2\alpha,0} L(R(K), \lambda)$. □

We provide two propositions that describe the behavior of the distribution G.

Proposition 1. *For any r in the support of G,*

$$G(r) \leq (\exp(\epsilon - 4\alpha) - 1) \sum_{n<r} G(n) + \exp(-2\alpha)\delta$$

Proof. Note the recurrence: for any $N - M \leq n \leq N$, $G(n + 1) \leq \exp(\epsilon - 4\alpha)G(n)$. Using this, we see that

$$\begin{aligned}\exp(\epsilon - 4\alpha) \sum_{n=N-M}^{r-1} G(n) &= \sum_{n=N-M}^{r-1} \exp(\epsilon - 4\alpha)G(n) \\ &\geq \sum_{n=N-M}^{r-1} G(n+1) = \sum_{n=N-M+1}^{r} G(n)\end{aligned}$$

Subtracting $\sum_{N-M}^{r-1} G(n)$ from both sides, we have,

$$\begin{aligned}(\exp(\epsilon - 4\alpha) - 1) \sum_{n=N-M}^{r-1} G(n) &= G(r) - G(N - M) \\ &\geq G(r) - \exp(-2\alpha)\delta\end{aligned}$$

Rearranging gives the desired result. □

Proposition 2. *For any r in the support of G,*

$$G(r) \leq (\exp(\epsilon) - 1) \sum_{n>r} G(n) + \delta$$

The proof of this proposition is similar to the previous one, so we omit it.

Given data $D \in \mathbb{D}^\star$ and researcher script R, let $m(D, R)$ be the size of the largest stable subset of D. Since any subset of a stable subset is also stable, Algorithm 1 it will output $\perp$ if it selects $n > m(D, R)$, and it will output a real number if it selects $n \leq m(D, R)$. The following lemma relates m to the idea of neighboring datasets.

Lemma 2. *When $D_1, D_2 \in \mathbb{D}^\star$ are neighbors, $m(D_1, R)$ and $m(D_2, R)$ differ by at most 1.*

Proof. Say that S_1 is a maximal stable subset of D_1. Then $S_1 \cap D_2$ is a stable subset of D_2. Further, $|S_1 \cap D_2| \geq |S_1| - 1 = m(D_1, R) - 1$. Therefore $m(D_2, R) \geq m(D_1, R) - 1$. By symmetric argument, $m(D_1, R) \geq m(D_2, R) - 1$. □

We can now complete the proof of Theorem 1. For any $D \in \mathbb{D}^\star$, write $W_D(\cdot|n)$ to represents the conditional probability distribution of the wrapper given the value of n. Let E be any measurable set in $\mathbb{R} \cup \{\perp\}$. Let $r = \max(m(D_1), m(D_2))$. For $i \in \{1, 2\}$ the law of total probability implies

$$W_{D_i}(E) = \sum_{n<r} W_{D_i}(E|n)G(n) + W_{D_i}(E|r)G(r) + \sum_{n>r} W_{D_i}(E|n)G(n)$$

We consider two cases:

Case $\perp \notin E$: We bound the privacy ratio as follows:

$$\frac{W_{D_1}(E) - \delta}{W_{D_2}(E)} \leq \frac{\sum_{n<r} \exp(2\alpha)H(E)G(n) + \exp(2\alpha)H(E)G(r) - \delta}{\sum_{n<r} \exp(-2\alpha)H(E)G(n) + 0} \leq \exp(4\alpha)\frac{\sum_{n<r} G(n) + G(r) - \exp(-2\alpha)\delta}{\sum_{n<r} G(n)}$$

Plugging in $G(r)$ from Proposition 1 and simplifying bounds the ratio by $\exp(\epsilon)$.

Case $\perp \in E$: By Lemma 1, $W_{D_i}(\cdot|n)$ is 2α-indistinguishable from $H(\cdot)$ when W_{D_i} doesn't halt. When $n < r$, W_{D_i} doesn't halt, and therefore $\exp(-2\alpha)H(E\setminus\{\perp\}) \leq W_{D_i}(E|n) \leq \exp(2\alpha)H(E \setminus \{\perp\})$. Hence,

$$\frac{W_{D_1}(E) - \delta}{W_{D_2}(E)} \leq \frac{\sum_{n<r} \exp(2\alpha)H(E \setminus \{\perp\})G(n) + G(r) + \sum_{n>r} G(n) - \delta}{\sum_{n<r} \exp(-2\alpha)H(E \setminus \{\perp\})G(n) + 0 + \sum_{n>r} G(n)}$$

Substituting for $G(r)$ from Proposition 2 and rearranging, the right-hand side is upper-bounded by

$$\exp(4\alpha)\frac{\sum_{n<r} H(E \setminus \{\perp\})G(n) + \exp(\epsilon - 2\alpha)\sum_{n>r} G(n)}{\sum_{n<r} H(E \setminus \{\perp\})G(n) + \exp(2\alpha)\sum_{n>r} G(n)}$$

Since $\epsilon > 4\alpha$, the fraction is greater than 1, so we can subtract the first terms from the numerator and denominator and maintain the inequality:

$$\frac{W_{D_1}(E) - \delta}{W_{D_2}(E)} \leq \exp(4\alpha)\exp(\epsilon - 4\alpha) = \exp(\epsilon)$$

6. Describing δ

The privacy parameter δ can be computed precisely, as δ^{-1} equals

$$\sum_{n=N-M}^{N} \exp\left(\min\Big((\epsilon{-}4\alpha)(n{-}N{+}M){-}2\alpha, \epsilon(N{-}n)\Big)\right)$$

To understand the magnitude of δ, it is helpful to have a closed form approximation. We can define one such approximation as follows:

$$\hat{\delta} = \frac{\epsilon(\epsilon - 4\alpha)\exp(2\alpha)}{4\alpha\Big[\exp\big(Q(M-1)\big) - 1\Big]}$$

where $Q = \epsilon(\epsilon - 4\alpha)(2\epsilon - 4\alpha)^{-1}$.

Proposition 3. *For a privacy wrapper with parameters ϵ and α, δ and $\hat{\delta}$ are within a multiplicative factor that approaches $\exp(2\alpha)$ as $M \to \infty$. Moreover $\hat{\delta} > \delta$, which guarantees that the privacy wrapper imposes $(\epsilon, \hat{\delta})$ differential privacy.*

We omit the proof due to space constraints.

Corollary 1. *For any fixed α and ϵ, δ is $\Theta(\exp(-M))$.*

7. Discussion

We have formalized the notion of a privacy wrapper, and presented a novel algorithm for its implementation. Our work demonstrates that differential privacy can achieved for real-valued statistics even when local sensitivity is unknowable. Future work will focus on the search for sub-exponential wrapper algorithms. We are also interested in developing metrics for comparing wrapper algorithms according to accuracy and their likelihood of halting.

References

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Johnson, N., Near, J. P., and Song, D. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539, 2018.

Jordon, J., Yoon, J., and Van Der Schaar, M. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

Kifer, D., Messing, S., Roth, A., Thakurta, A., and Zhang, D. Guidelines for implementing and auditing differentially private systems. *arXiv preprint arXiv:2002.04049*, 2020.

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, 2009.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Vadhan, S. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer, 2017.