

Differentially Private k -Nearest Neighbor Missing Data Imputation

Chris Clifton

Dept. of Computer Science and CERIAS, Purdue University, West Lafayette, IN USA

Eric J. Hanson

Dept. of Mathematics, Brandeis University Waltham, MA USA

Keith Merrill

Dept. of Mathematics, Brandeis University Waltham, MA USA

Shawn Merrill

Dept. of Computer Science and CERIAS, Purdue University, West Lafayette, IN USA

Summary. Using techniques employing *smooth sensitivity*, we develop a method for k -nearest neighbor missing data imputation with differential privacy. This requires bounding the number of data incomplete tuples that can have their data complete “donor” changed by making a single addition or deletion to the dataset. The multiplicity of a single individual’s impact on an imputed dataset necessarily means our mechanisms require the addition of more noise than mechanisms that ignore missing data, but we show empirically that this is significantly outweighed by the bias reduction from imputing missing data.

1. Introduction

Missing data poses substantial challenges for data analysis and machine learning. If data were missing uniformly at random, this would not be a big issue, but in practice (as with survey data), missing data tends to be biased (Nicoletti and Peracchi, 2006; Kalton and Kasprzyk, 1982). As a result, analysis based on only collected data gives poor results.

A common solution to this problem is *missing data imputation*. The usual first step is to use known data about an individual to impute any missing values (for example, if an individual leaves their employment status unanswered but reports income from a job, their employment status can be deduced). When this fails, it is common to instead use known values from similar individuals (an approach also referred to as *allocation* (Kalton and Kasprzyk, 1982; United States Census Bureau, 2014)). This poses a privacy challenge, as a single individual’s value is reflected in multiple records, making it more difficult to keep that value private.

This privacy risk can be mitigated with *differential privacy* (Dwork et al., 2006). Differential privacy adds noise to a query result sufficient to hide the impact of any individual on the result. If an individual is a “donor” for several individuals with a missing value, then their impact on the result for many queries (how much the result would change if that individual were removed) increases with the number of individuals who take their value, requiring substantially more noise to cover their impact.

A basic approach to differential privacy requires a worst-case view: the amount of noise added is based on the worst possible scenario that can be constructed. In the case of

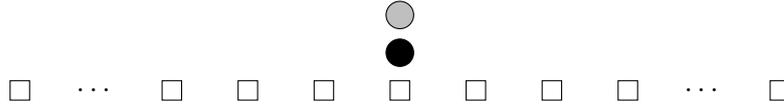


Fig. 1. Sample dataset with high global sensitivity. Each individual has attributes `SHAPE` and `COLOR`. The squares are missing `COLOR`, so the color of every square is initially imputed as black. After deleting the black circle, the color of every square is imputed as gray.

missing data imputation, we can imagine a scenario like Figure 1: If we assume the value of the attribute `color` is imputed from the nearest neighbor (in two-dimensions) with a known value, deleting the black circle will change the imputed color of every square. In particular, the count of gray individuals goes from 1 to the size of the dataset. This essentially requires sufficient noise to hide the entire range of possible answers.

Such pathological cases seem unlikely to occur in practice. *Smooth sensitivity* (Nissim et al., 2007) allows us to base our noise not on a global worst case, but on worst cases bearing resemblance to the actual dataset. As more changes to the data are needed to get to a worst case, the impact of the worst case on the noise added goes down. We will review this in Section 3.1, but for now note the key challenge: We need to account for the worst case after any number of changes to the actual data, a computationally intractable problem unless we can analytically bound the impact on a query result of a given number of changes. This is difficult, and there have been few examples of problems that can be addressed by smooth sensitivity beyond those worked out in the original paper, most quite recent: see for example synthetic graph generation (Wang and Wu, 2013), outlier detection (Okada et al., 2015), random forests (Fletcher and Islam, 2017), and PCA (Gonem and Gilad-Bachrach, 2018).

1.1. Problem Statement and Summary of Results

In this paper, we establish a *smooth upper bound* (in the sense of Nissim et al. (2007)) for k -nearest-neighbor imputation, regardless of how the data of the neighbors translates into the imputed value (e.g., averaging, majority vote, etc.). Our examples and mechanisms focus on counts, proportions, means, and variances, and the results easily extend to any queries where the impact of an individual is at most multiplicative in the number of times they are a “donor” (sums and correlations are examples).

Our use of smooth sensitivity to achieve differentially private k -nearest neighbor data imputation gives a dramatic reduction in variance compared to using the global sensitivity. While missing data imputation does give substantially higher variance than simply throwing out individuals with missing data, we do get the reduction in bias that is the primary benefit of doing missing data imputation. While this bias/variance trade-off is difficult to quantify analytically (it is heavily dependent on the specifics of the dataset), we do give a synthetic, but realistic, example in Section 6 that is reflective of an actual large-scale, high-dimensional survey.

The technical contributions of this work are: we show that any global sensitivity-based approach is untenable in Example 1; we establish computationally efficient smooth

upper bounds in Theorems 5.1-5.6 and Corollaries 5.2-5.7 that allow us to utilize smooth sensitivity; and in Section 6 we show that for numerous queries of practical interest, our mechanism gives substantially better results than simply ignoring missing data.

2. Related Work

Imputation of missing values for a *data incomplete* tuple is generally done by computing a value based on known values for that individual (edit rules), modeling the value based on known values for other individuals (e.g., using the mean), or inserting a known value from a donor *data complete* tuple (allocation) (Kalton and Kasprzyk, 1982). We first discuss allocation approaches and why they pose challenges for differential privacy. We also overview existing methods for private data imputation.

2.1. Allocation for Missing Data Imputation

Using a value from a donor is more likely to give a legal value (e.g., avoiding the family with 2.4 children). Methods such as *hot deck imputation* have a long history. In basic form, hot deck uses the last seen value for a missing data item (Bailar III and Bailar, 1978). This works if data is missing at random, but often missing data is biased towards certain classes of individuals. To address this concern, donors are chosen to be similar to the incomplete data item with regards to certain known attributes, under the assumption that individuals similar on those attributes have similar values for the missing data. Unfortunately, this can also lead to biased results, if an individual with an unusual value ends up being the donor to many missing individuals. This has given rise to complicated procedures, such as sorting hot deck data to try to get donors who are similar to the individuals with missing data (Bailar III and Bailar, 1978), or the mechanism used in the U.S. Census Bureau’s American Community Survey (United States Census Bureau, 2014). The latter identifies a similar individual as a donor, but then discards that individual for a period of time before allowing it to be reused as a donor.

Methods that use such a donor have obvious implications for differential privacy. The sensitivity of a query must account for the fact that queries covering imputed missing data may actually be multiply dependent on the value from the donor. As shown in the example in the introduction (Figure 1), this can give arbitrarily high sensitivity.

Methods such as hot deck imputation (or that use some of its concepts, such as United States Census Bureau (2014)) would intuitively seem to be well-suited for differential privacy, since an individual donor’s contributions are limited. Unfortunately, such techniques are sensitive to the order in which tuples are processed (Bailar III and Bailar, 1978). The following example shows that this makes global sensitivity arbitrarily large.

In Figure 2 we again consider individuals with attributes **shape** and **color** and impute values of **color** based on distance in two dimensions. With hot deck, a potential donor cannot impute on two consecutive individuals. Thus if we delete (from the left dataset) the leftmost black circle, the imputed value of **color** of each of the data incomplete individuals is swapped (as shown in the right dataset). The (imputed) number of gray squares is changed from 0 to the total number of squares with this single deletion. As a result, it is impractical to satisfy the differential privacy guarantee with such approaches.

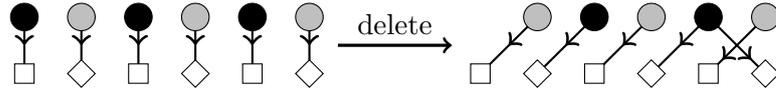


Fig. 2. High sensitivity of hot deck imputation. Each individual has attributes shape and color. The arrows represent values being imputed. Deleting the left black circle from the dataset on the left changes the imputed value of color for every data incomplete tuple (shown on the right).

2.2. Private Data Cleaning Methods

There have been some methods that propose privatized (including differentially private) mechanisms for data cleaning. However, to our knowledge these all address a different cleaning problem: correcting values that are presumed to be erroneous.

InfoClean (Chiang and Gairola, 2018) is an automated cleaning mechanism that satisfies a form of information theoretic privacy, but is not shown to satisfy differential privacy. It also addresses a somewhat different problem. The assumption is that a given tuple is known to be erroneous, and is fixed by comparing with a similar tuple retrieved from a master database; the only privacy considered is that of data in the Master database. PACAS (Huang et al., 2018) addresses a similar problem, using a k -anonymity based privacy metric.

The differentially private data cleaning methods PrivateClean (Krishnan et al., 2016) and PrivClean (Ge et al., 2018) support human-in-the-loop cleaning. Both enable an expert to specify rules for data cleaning, and ensure that the result of a query is differentially private, which may include the impact of the expert looking at data to generate the rules. As such, this is really not comparable with our approach.

2.3. Low Rank Estimation

Another common solution to missing data imputation in the differentially private setting is to view the dataset as a matrix with missing entries, and produce a differentially private matrix which is similar in to the original with respect to some norm. Examples of such approaches include McSherry and Mironov (2009); Kapralov and Talwar (2013); Jain et al. (2018). This is a powerful method, and unlike the present paper, releases a synthetic dataset on which an arbitrary number of queries can then be run. However these recent efforts have predominantly focused on *recommender systems*, which are homogeneous in their variables and do not exhibit the structure and relationships between variables which we exploit here. This stands in stark contrast with many surveys, for instance the U.S. Census Bureau’s American Community Survey (ACS) which we will focus on below. Some of these attempts have also used a weaker privacy guarantee than that of (pure) differential privacy, which we satisfy in this paper.

3. Background

Throughout this paper, we use the notation \mathcal{D} to refer to a dataset from some universe \mathcal{D} . For now, we put no assumptions on \mathcal{D} , but we will add restrictions in Section 4 that will allow us to discuss differentially private data imputation. For two arbitrary sets U and V , let $U\Delta V$ denote the *symmetric difference*; that is, $U\Delta V := (U \cup V) \setminus (U \cap V)$.

For two datasets D, D' , we denote by $d(D, D')$ the *Hamming distance* between D and D' given by $d(D, D') = |D \Delta D'|$. By a query q , we mean a map $q : \mathcal{D} \rightarrow \mathbb{R}^d$ for some positive integer d . Given $a \in \mathbb{R}^d$, we denote by $\|a\|_1$ the ℓ_1 norm of a ; that is, $\|a\|_1 := \sum_{i=1}^d |a_i|$.

A randomized algorithm $\mathcal{A} : \mathcal{D} \rightarrow \mathbb{R}^d$ is said to satisfy ε -*differential privacy* (Chawla et al., 2005) if for all $D, D' \in \mathcal{D}$ such that $d(D, D') = 1$ and for all measurable $S \subset \mathbb{R}^d$,

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

Note that we are working in what is referred to as the *unbounded* differential privacy setting, where a difference between D and D' is from adding or removing an individual.

3.1. Smooth Sensitivity

The goal of this section is to provide an overview of Nissim, Raskhodnikova, and Smith's results on *smooth sensitivity* (Nissim et al., 2007). For convenience, we have included proofs of the statements we will use. All of the proofs in this section have been adapted from those in the original paper. Some of the constants in the statements below are tighter bounds than the original, and are better suited to our needs. We begin with an overview of the problem they solved.

Definition 3.1. Let q be an arbitrary query. Then the *global sensitivity* of q is:

$$GS_q := \max_{D, D': d(D, D')=1} \|q(D) - q(D')\|_1,$$

where the maximum is over all datasets D, D' which are neighbors in the Hamming distance.

The global sensitivity of a query is often used to achieve differential privacy via the *Laplace mechanism*.

Theorem 3.2. (Chawla et al., 2005) Let $\varepsilon > 0$ and let q be a query taking values in \mathbb{R}^d . Then the mechanism $\mathcal{A}_q(D) = q(D) + (X_1, \dots, X_d)$, where the X_i are i.i.d. from

$$f(x) = \frac{\varepsilon}{2GS_q} \exp(-|x|\varepsilon/GS_q),$$

satisfies ε -*differential privacy*.

For queries with low global sensitivity (e.g., counts, proportions), the Laplace mechanism provides a strong privacy guarantee without losing data utility. Unfortunately, many important classes of queries (e.g., medians) have high or unbounded global sensitivity. This is often due to the existence of pathological datasets that are likely very different from real-world data. To avoid the necessity of adding noise to cover these unrealistic scenarios, it is natural to consider instead using *local sensitivity*.

Definition 3.3. Let q be an arbitrary query and let D be a dataset. The *local sensitivity* of q at D is:

$$LS_q(D) := \max_{D': d(D, D')=1} \|q(D) - q(D')\|_1.$$

The local sensitivity of a query is obviously bounded above by the global sensitivity, and can be considerably lower. However, one should generally be wary of mechanisms based solely on local sensitivity, as there are queries for which adding noise proportional to the local sensitivity cannot satisfy differential privacy (such as median, see Nissim et al. (2007)). As we shall see, the mechanisms proposed in this paper *are* based solely on (an upper bound of) the local sensitivity, a computationally convenient fact.

We prove our proposed mechanisms are differentially private using Nisse, Rashodnikova, and Smith's compromise between local and global sensitivity, formed by computing a *smooth upper bound*.

Definition 3.4. (Nissim et al., 2007, Definition 2.1) Let $\beta > 0$ and let q be an arbitrary query. A β -smooth upper bound is a function $S : \mathcal{D} \rightarrow \mathbb{R}^+$ such that

$$\begin{aligned} \forall D \in \mathcal{D} : \quad & S(D) \geq LS_q(D) \\ \forall D, D' \in \mathcal{D}, d(D, D') = 1 : \quad & S(D) \leq e^\beta \cdot S(D'). \end{aligned}$$

The following gives a general construction for turning a bound on local sensitivity into a β -smooth upper bound. The values $U_k(D)$ below can be thought of as an upper bound on the maximum local sensitivity of any dataset (Hamming) distance k from D .

Theorem 3.5. (Nissim et al., 2007, Claim 3.2) Let q be a query. For $k \in \mathbb{N}$, let $U_k : \mathcal{D} \rightarrow \mathbb{R}$ so that

$$\begin{aligned} \forall D \in \mathcal{D} : \quad & LS_q(D) \leq U_0(D), \\ \forall k \in \mathbb{N}, \forall D, D' \in \mathcal{D}, d(D, D') = 1 : \quad & U_k(D) \leq U_{k+1}(D'). \end{aligned}$$

Then there is a β -smooth upper bound on the local sensitivity of q given by:

$$SS_{\beta,q}(D) = \max_k e^{-\beta k} U_k(D).$$

This proof is taken from the preliminary version of Nissim et al. (2007).

PROOF. By definition, we have

$$LS_q(D) \leq U_0(D) \leq SS_{\beta,q}(D),$$

so let $d(D, D') = 1$. Then

$$\begin{aligned} SS_{\beta,q}(D) &= \max_k e^{-\beta k} U_k(D) \leq e^\beta \max_k e^{-\beta(k+1)} U_{k+1}(D') \\ &\leq e^\beta \max_k e^{-\beta k} U_k(D') = SS_{\beta,q}(D'). \end{aligned}$$

In order to construct differentially private mechanisms based on this definition and the theorem, we need the following result.

Theorem 3.6. (Nissim et al., 2007, Lemma 2.6) Let $f(x)$ be a pdf on \mathbb{R}^d and let $\varepsilon > 0$. Let q be a query taking values in \mathbb{R}^d . Suppose there exist parameters $\alpha, \beta > 0$ such that for all measurable subsets $B \subset \mathbb{R}^d$,

$$\begin{aligned} \forall \Delta \in \mathbb{R}^d, \|\Delta\|_1 \leq \alpha : \quad & \Pr_{X \sim f} [x \in B] \leq e^{\varepsilon/2} \cdot \Pr_{X \sim f} [x \in B + \Delta] \\ \forall \lambda \in \mathbb{R}, |\lambda| \leq \beta : \quad & \Pr_{X \sim f} [x \in B] \leq e^{\varepsilon/2} \cdot \Pr_{X \sim f} [x \in e^\lambda \cdot B]. \end{aligned}$$

Then for any β -smooth upper bound, $S(D)$, of LS_q , the algorithm

$$\mathcal{A}(D) = q(D) + \frac{S(D)}{\alpha} \cdot X$$

is ε -differentially private, where X is a random variable with pdf $f(x)$. In this case, the distribution corresponding to f is called an (α, β) -admissible noise down distribution.

PROOF. Let $B \subset \mathbb{R}^d$ be measurable and let $D, D' \in \mathcal{D}$ be neighboring datasets. Then

$$\begin{aligned} \Pr[\mathcal{A}(D) \in B] &= \Pr_{X \sim f} \left[x \in \frac{\alpha(B - q(D))}{S(D)} \right] \leq \Pr_{X \sim f} \left[x \in \frac{\alpha(B - q(D'))}{S(D)} \right] \cdot e^{\varepsilon/2} \\ &\leq \Pr_{X \sim f} \left[x \in \frac{\alpha(B - q(D'))}{S(D')} \right] \cdot e^\varepsilon = \Pr[\mathcal{A}(D') \in B] \cdot e^\varepsilon. \end{aligned}$$

The first inequality holds because

$$\frac{\alpha \|q(D') - q(D)\|_1}{S(D)} \leq \frac{\alpha \|q(D') - q(D)\|_1}{LS_q(D)} \leq \alpha$$

by the definition of local sensitivity. The second inequality holds because

$$\left| \frac{S(D)}{S(D')} \right| \leq e^\beta$$

by the definition of a β -smooth upper bound.

We now give our primary example of an admissible noise down distribution, which we refer to as the *generalized Cauchy distribution*.

Definition 3.7. Let $\gamma > 1$. The *generalized Cauchy distribution* with parameter γ has a density given by

$$f(x) \propto \frac{1}{1 + |x|^\gamma}.$$

Theorem 3.8. (Nissim et al., 2007, Lemma 2.7) Let $\varepsilon > 0$ and $\gamma > 1$. Then the *generalized Cauchy distribution* with parameter γ is $\left(\frac{\varepsilon}{2(\gamma-1)}, \frac{\varepsilon}{2(\gamma-1)}\right)$ -admissible (note that in this case, $\alpha = \beta$). In particular, if q is an real-valued query and S is a $\left(\frac{\varepsilon}{2(\gamma-1)}\right)$ -smooth upper bound of the local sensitivity of q , then the mechanism

$$\mathcal{A}(D) = q(D) + \frac{2(\gamma-1) \cdot S(D)}{\varepsilon} \cdot X,$$

where X is sampled from a *generalized Cauchy distribution* with parameter γ , satisfies ε -differential privacy.

PROOF. Let $\mu \in \mathbb{R}$ with $|\mu| \leq \frac{\varepsilon}{2(\gamma-1)}$. Let $f_0(x) = \frac{1}{1+|x|^\gamma}$. We must show that $\frac{f(x)}{f(x+\mu)} = \frac{f_0(x)}{f_0(x+\mu)} \leq e^{\varepsilon/2}$, or equivalently $\ln(f_0(x)) - \ln(f_0(x+\mu)) \leq \varepsilon/2$. First observe

$$f_0(x) = \begin{cases} \frac{1}{1+x^\gamma} & x \geq 0 \\ \frac{1}{1+(-x)^\gamma} & x < 0 \end{cases} \quad f_0'(x) = \begin{cases} \frac{-\gamma x^{\gamma-1}}{(1+x^\gamma)^2} & x \geq 0 \\ \frac{\gamma(-x)^{\gamma-1}}{(1+(-x)^\gamma)^2} & x < 0 \end{cases}$$

Now, by the mean value theorem, there exists $z \in (x, x + \mu)$ such that

$$|\ln(f_0(x)) - \ln(f_0(x + \mu))| = |\mu \cdot (\ln \circ f_0)'(z)|.$$

If $z > 0$, we have

$$|(\ln \circ f_0)'(z)| = \left| \frac{f_0'(z)}{f_0(z)} \right| = \frac{\gamma z^{\gamma-1}}{1 + z^\gamma} = \frac{\gamma}{z^{1-\gamma} + z}.$$

Now consider that the function $g(z) = z^{1-\gamma} + z$ is minimized at $z_0 = (\gamma-1)^{1/\gamma}$. Moreover,

$$g(z_0) = (\gamma-1)^{1/\gamma} \cdot \frac{\gamma}{\gamma-1} = \frac{\gamma}{(\gamma-1)^{1-1/\gamma}} \geq \frac{\gamma}{\gamma-1}.$$

Therefore,

$$|(\ln \circ f_0)'(z)| = \frac{\gamma}{g(z)} \leq \frac{\gamma}{g(z_0)} \leq \gamma - 1.$$

Likewise, if $z \leq 0$, we have

$$|(\ln \circ f_0)'(z)| = \left| \frac{f_0'(z)}{f_0(z)} \right| = \frac{\gamma}{(-z)^{1-\gamma} + (-z)} \leq \gamma - 1.$$

Thus in any case, we have

$$\ln(f_0(x)) - \ln(f_0(x + \mu)) \leq (\gamma - 1)\mu \leq \varepsilon/2.$$

Now let $\lambda \in \mathbb{R}$ with $|\lambda| \leq \frac{\varepsilon}{2(\gamma-1)}$. We must show that $\frac{f(x)}{e^\lambda f(e^\lambda x)} \leq e^{\varepsilon/2}$. If $\lambda \geq 0$, then

$$\frac{f(x)}{e^\lambda f(e^\lambda x)} = \frac{1 + |e^\lambda x|^\gamma}{e^\lambda(1 + |x|^\gamma)} \leq (e^\lambda)^{\gamma-1} \leq e^{\varepsilon(\gamma-1)/2(\gamma-1)} < e^{\varepsilon/2}.$$

Likewise, if $\lambda < 0$, then

$$\frac{f(x)}{e^\lambda f(e^\lambda x)} = \frac{1 + |e^\lambda x|^\gamma}{e^\lambda(1 + |x|^\gamma)} \leq e^{-\lambda} \leq e^{\varepsilon/2(\gamma-1)} < e^{\varepsilon/2}.$$

The following is a reformulation of this theorem, which will feature in Section 5.

Corollary 3.9. *Let $\varepsilon, \beta > 0$, let q be a real-valued query, and choose γ so that $\gamma \leq 1 + \frac{\varepsilon}{2\beta}$. Then for $S(D)$ a β -smooth upper bound of the local sensitivity of q , the mechanism*

$$\mathcal{A}(D) = q(D) + \frac{S(D)}{\beta} \cdot X,$$

where X is sampled from the generalized Cauchy distribution with parameter γ , satisfies ε -differential privacy.

Remark 3.10. We make two remarks about interpreting the theorem in this way.

- (a) While this mechanism works for any values of ε and β , it is inadvisable to use small values of γ . In particular, the generalized Cauchy distribution only has well-defined variance when $\gamma > 3$. Above this value, the variance is a decreasing function in γ .

- (b) Technically, the corollary holds for any γ such that $1 < \gamma \leq 1 + \frac{\epsilon}{2\beta}$. However, γ should be taken to have the maximal value possible since the variance of the generalized Cauchy mechanism (when it is defined) is decreasing in γ .

While beautiful theoretical results, it is often computationally intractable to obtain differential privacy using Theorems 3.6 and 3.8. Without a bound on the change in local sensitivity between two datasets arbitrarily far apart, an exhaustive search of all nearby neighbors needs to be done until the smooth upper bound is found.

In Section 5, we compute a smooth upper bound for the local sensitivity of several queries under k -nearest neighbor allocation. We then examine more closely the effect of the parameters γ and β on the variance of the noise added.

4. Deterministic Data Imputation

The goal of this section is to establish the necessary theoretical framework to apply differential privacy to a dataset with (k -nearest neighbor) imputation. We do so by examining the relationship between the local sensitivity of a query and the ability to change the donor(s) of a data incomplete tuple.

4.1. Theoretical Framework

We start with a few (heavily theoretical) definitions, but the well-known motivating examples are listed in Example 4.2. Let T denote the set of all possible tuples, in which missing values are allowed. Throughout this section, we fix a set of attributes A . We let T_c denote the subset of T consisting of tuples with no missing values and T_i the subset of T consisting of tuples missing responses to A . We refer to T_c as the set of *complete tuples* and T_i as the set of *incomplete tuples*. We assume that these are the only two possibilities (although the results still hold under the existence of other types of tuples).

From now on, we assume that a dataset cannot contain two identical tuples[†]. This allows us to define $D_c := D \cap T_c$ and $D_i := D \cap T_i$ so that $D = D_c \cup D_i$ and $D_c \cap D_i = \emptyset$.

For any positive integer k , we define $\binom{T_c}{k} := \{S \subset T_c : |S| = k\}$ and denote by $\text{ord}(T_c)$ the set of (total) orders on T_c . For A a set of attributes, we refer to the set of possible responses to A as $\text{resp}(A)$ and the responses given by $y \in T_c$ as $\text{resp}_y(A)$.

Definition 4.1. (a) A *deterministic k -nearest neighbor imputation scheme* for the set of attributes A is a pair (f, g) where $f : T_i \rightarrow \text{ord}(T_c)$ and $g : \binom{T_c}{k} \rightarrow \text{resp}(A)$.

- (b) Fix a pair (f, g) as above and let $D = D_c \cup D_i$ be a dataset. Then for $x \in D_i$, we denote by $\text{don}_D(x)$ the set of k elements of D_c which are smallest with respect to the ordering $f(x)$. We call $\text{don}_D(x)$ the *donor set* (or just *donor* if $k = 1$) of x . The value imputed to x for the attributes in A is then $g(\text{don}_D(x))$.

The definition of a deterministic k -nearest neighbor imputation scheme is intentionally very abstract. This allows us to state the results of this section with some generality. The following are some common examples that fit into this framework.

[†]This assumption, which is necessary for the proofs that follow, is not difficult to achieve in reality. For example, multiple individuals can be identical up to some unique record label.

Example 4.2.

- (a) *Nearest Neighbor*: Let $k = 1$. For $x \in T_i$, define $f(x)$ to be the order (with tie-breakers if necessary) on T_c given by some distance metric (e.g., Hamming distance on a certain set of attributes). For $y \in T_c$, define $g(y) = \text{resp}_y(A)$. Then (f, g) is the imputation scheme which copies the responses to A from the nearest data complete tuple in the dataset.
- (b) *Mean*: For $x \in T_i$, define $f(x)$ to be the order (with tie-breakers if necessary) on T_c given by some distance metric (e.g., Hamming distance on a certain set of attributes). For $Y \in \binom{T_c}{k}$, define

$$g(Y) = \frac{1}{k} \sum_{y \in Y} \text{resp}_y(A).$$

Then (f, g) is the imputation scheme which imputes the average of the responses to A from the k nearest data complete tuples in the dataset.

- (c) *Majority*: For $x \in T_i$, define $f(x)$ to be the order (with tie-breakers if necessary) on T_c given by some distance metric (e.g., Hamming distance on a certain set of attributes). For $Y \in \binom{T_c}{k}$, define $i(Y)$ to be the most common value of $\text{resp}_y(A)$ over $y \in Y$. Then (f, g) is the imputation scheme which imputes the most common response to A from the k nearest data complete tuples in the dataset.

Remark 4.3.

- (a) Given a dataset $D = D_c \cup D_i$ and a data incomplete tuple $x \in D_i$, it is necessary for the order $f(x)$ to be on T_c (the set of *all possible* data complete tuples) rather than D_c (the set of data complete tuples actually in D). Otherwise, if we make a change to the dataset by adding a new data complete tuple $y \in T_c \setminus D_c$, we would have no way to determine whether y should replace (one of) the donor(s) of x .
- (b) As in the above examples, it is often not necessary to specify or compute the entire function $f : T_i \rightarrow \text{ord}(T_c)$. Indeed, specifying a metric to determine the nearest neighbor (for example Hamming distance on a certain set of attributes with the difference in record labels as a tiebreaker) is enough that one could construct the function f if desired. This will be the approach used in our empirical analysis.
- (c) Definition 4.1 is made so that a single change to the dataset can only change at most one donor of a data incomplete tuple. That is, if $d(D, D') = 1$ and $x \in D_i \cap D'_i$, then $|\text{don}_D(x) \cap \text{don}_{D'}(x)| \geq k - 1$.

From now on, we fix an imputation scheme (f, g) .

Definition 4.4.

- (a) Let D be a dataset. Given $y \in D_c$, its *set of donees* is

$$\text{don}_D^{-1}(y) := \{x \in D_i \mid y \in \text{don}_D(x)\}.$$

That is, $\text{don}_D^{-1}(y)$ consists of those incomplete tuples in D whose values are imputed based on y . For convenience, we will define $\text{don}_D^{-1}(y) = \emptyset$ for $y \in T_c \setminus D_c$.

(b) Let D, D' be datasets. The *donee change* between D and D' is

$$c(D, D') := [D_i \Delta (D')_i] \cup [\text{don}_D^{-1}(D \cup D') \Delta \text{don}_{D'}^{-1}(D \cup D')].$$

The first term represents the data incomplete tuples added or deleted as we transform D into D' . The second term represents those incomplete tuples that have their donor set change as we transform D into D' . We note that $c(D, D') = c(D', D)$.

4.2. Bounding Donee Changes

We now turn our attention to studying how the function $c(D, D')$ behaves as $d(D, D')$ increases. More precisely, we study the function

$$L_\ell(D) := \max_{\{D': d(D, D')=\ell\}} |c(D, D')|.$$

Remark 4.5.

- (a) Our motivation for studying this function is to construct a smooth upper bound for the local sensitivity of several queries based upon these values. Our definition of $L_\ell(D)$ enables us to use Theorem 4.6 later in Section 5 to create mechanisms requiring only the computation of $L_1(D)$ for a given dataset.
- (b) The quantity $L_1(D)$ will play a particularly important role in our analysis. Recall that from the unbounded definition of differential privacy, the only changes we allow to the dataset are the addition or deletion of a tuple. $L_1(D)$ can therefore be seen as a maximum over the impacts of such a change.

The following result describes how the function $L_\ell(D)$ changes as the distance ℓ and dataset D change, and forms the basis of our mechanisms.

Theorem 4.6. *Let D be any dataset. Then:*

- (a) For all $\ell \in \mathbb{N}$, $L_\ell(D) < L_{\ell+1}(D)$.
- (b) For all $\ell \in \mathbb{N}$, $L_\ell(D) \leq \ell L_1(D)$.
- (c) For any dataset D' , $L_1(D') \leq L_{d(D, D')+1}(D)$.

PROOF. (a) Let D' be a dataset realizing $|c(D, D')| = L_\ell(D)$, and let $x \in T_i$ such that $x \notin D \cup D'$. Then $D' \cup \{x\}$ is a dataset satisfying $d(D, D' \cup \{x\}) = \ell + 1$ and

$$L_{\ell+1}(D) \geq c(D, D' \cup \{x\}) = c(D, D') + 1 > c(D, D') = L_\ell(D).$$

(b) Let D' be a dataset realizing $L_\ell(D) = |c(D, D')|$, and let $D \Delta D' = \{t_1, \dots, t_\ell\}$. We will prove this statement by induction on ℓ . For $\ell = 1$, there is nothing to show, so assume the statement holds for $\ell - 1$. We then have

$$c(D, D') = c(D, D_{\ell-1}) \cup (c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})),$$

where $D_{\ell-1}$ is the dataset obtained from D by adding/removing the tuples $t_1, \dots, t_{\ell-1}$. Hence by the induction hypothesis

$$\begin{aligned} L_\ell(D) &= |c(D, D')| \\ &= |c(D, D_{\ell-1})| + |c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})| \\ &\leq (\ell - 1)L_1(D) + |c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})|. \end{aligned}$$

Therefore it suffices to show that

$$|c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})| \leq L_1(D).$$

First observe that if $t_\ell \in T_i$, then $|c(D_{\ell-1}, D')| = 1$, and we are done. Thus we can assume that $t_\ell \in T_c$.

As a set, $c(D', D_{\ell-1})$ represents those incomplete tuples whose donor set changes in the move from $D_{\ell-1}$ to D' and had not changed previously. To simplify notation, denote

$$D \pm \{t_\ell\} = \begin{cases} D \cup \{t_\ell\}, & t_\ell \notin D \\ D \setminus \{t_\ell\}, & t_\ell \in D \end{cases}.$$

We claim

$$c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1}) \subseteq \{x \in D_i : t_\ell \in (\text{don}_D(x)) \Delta (\text{don}_{D \pm \{t_\ell\}}(x))\}.$$

To see this, let $x \in c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})$. We first note that since t_ℓ is data complete, we have $(D_{\ell-1})_i = (D')_i$. Next, observe that $x \in D_i$. Indeed, if $x \notin D_i$, then x must be equal to some t_j and therefore $x \in c(D, D_{\ell-1})$, a contradiction. Since $x \in D_i$ and it does not change donors in the first $\ell - 1$ steps, we have $\text{don}_D(x) = \text{don}_{D_{\ell-1}}(x)$.

Now if $t_\ell \in D$, then $t_\ell \in \text{don}_{D_{\ell-1}}(x) = \text{don}_D(x)$ and we are done. Thus assume $t_\ell \in \text{don}_{D'}(x)$. If $\text{don}_{D'}(x) \setminus \{t_\ell\} \not\subseteq D$ then x changes donors between D and $D_{\ell-1}$, a contradiction. Therefore $t_\ell \in \text{don}_{D \pm \{t_\ell\}}(x)$. This proves the claim.

We conclude that $|c(D', D_{\ell-1}) \setminus c(D, D_{\ell-1})| \leq |c(D, D \pm \{t_\ell\})| \leq L_1(D)$, as needed.

(c) This proof is similar to that of part (2), with one change to be explained later. Let D' be another dataset, and set $\ell := d(D, D')$. Let D'' be the neighboring dataset of D' which realizes $|c(D', D'')| = L_1(D')$. As before, set

$$D \Delta D' = \{t_1, \dots, t_\ell\}, \quad D' \Delta D'' = \{t\}.$$

We now break the proof into cases, some of which are trivial. If $t \in T_i$, then $L_1(D') = |c(D', D'')| = 1$, and there is nothing to show. Thus assume that $t \in T_c$. We now have two cases to consider.

(c1) If $t \neq t_j$ for all j , then as in case (2), we have

$$c(D', D'') \subseteq c(D, D''),$$

and hence

$$L_1(D') = |c(D', D'')| \leq |c(D, D'')| \leq L_{d(D, D'')}(D) \stackrel{(a)}{\leq} L_{\ell+1}(D),$$

where the last inequality holds because $d(D, D'') \leq d(D, D') + 1$ by the triangle inequality and the function $s \mapsto L_s(D)$ is strictly increasing.

(c2) If $t = t_j$, since the order of those moves does not matter, we can assume without loss of generality that $t = t_\ell$. In other words, the biggest change one can make to the dataset D' is to add (resp. remove) the data complete item we just removed (resp. added) in the previous step. Therefore, $c(D', D'') = c(D_{\ell-1}, D')$, and so

$$L_1(D') = |c(D', D'')| = |c(D_{\ell-1}, D')| \leq |c(D, D')| \leq L_\ell(D) \stackrel{(a)}{\leq} L_{\ell+1}(D).$$

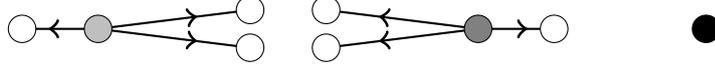


Fig. 3. A greedy algorithm cannot be used to compute $L_\ell(D)$. See Example 4.7 below.

We now provide an example that shows we can not use a “greedy algorithm” to compute $L_\ell(D)$. More precisely, consider a pair of datasets D, D' with $d(D, D') = 1$ and $L_1(D) = |c(D, D')|$. We would like to be able to say that

$$L_\ell(D) \leq L_1(D) + L_{\ell-1}(D' \setminus c(D, D')).$$

That is, we would like to be able to delete all individuals whose donor has changed in transitioning from D to D' , compute $L_{\ell-1}$ on this new dataset, and recover $L_\ell(D)$ from this value. The following example shows that this is not the case.

Example 4.7. We consider the dataset shown in Figure 3. As in previous examples, individuals have attributes **shape** and **color**. Values of **color** are imputed based on distance in two dimensions and arrows represent donor/donee relationships.

We observe that $L_1(D) = 4$, obtained by adding a black tuple between the four white tuples. Likewise, $L_1(D' \setminus c(D, D')) = 1$, obtained by deleting either gray circle.

On the other hand, we have that $L_2(D) = 6$. This can be realized by deleting both the light gray and dark gray circles, resulting in all imputing from the right black circle.

In light of this example, it is not computationally feasible to compute $L_\ell(D)$ for all $\ell \in \mathbb{N}$. We instead use Theorem 4.6 to design differentially private mechanisms based only on the value of $L_1(D)$. We discuss computing this value in Section 6.2.

5. Differentially Private Imputation-based Mechanisms

In this section we describe mechanisms satisfying differential privacy release query results on imputed datasets. Explicitly, we will use Theorems 3.5 and 4.6 to bound the smooth sensitivity of several queries in terms of the quantity $L_1(D)$. We then use the mechanism in Corollary 3.9 to achieve differential privacy.

We will assume from now on that all queries are answered on the imputed dataset. We also remark that these mechanisms only provide benefit when used for queries which are impacted by imputation. In other words, as we have fixed a set of attributes A in our definition of data-complete and data-incomplete tuples, any query which does not involve the attributes in A can and should be answered using standard techniques.

5.1. Counts and Proportions

We first consider counting queries and proportion queries.

Theorem 5.1. *Let q be a counting query. Then for $\beta \geq \ln 2$,*

$$SS_{\beta,q}(D) = 1 + L_1(D)$$

is a β -smooth upper bound on $LS_q(D)$.

PROOF. Let D be a dataset. Let $D \pm \{t\}$ be a neighboring dataset where we have either added or deleted the tuple t . Then t is a donor for at most $L_1(D)$ tuples (in whichever of D and $D \pm \{t\}$ contains t). This means the largest change between $q(D)$ and $q(D \pm \{t\})$ is $1 + L_1(D)$. That is, $LS_q(D) \leq 1 + L_1(D)$.

Now let

$$U_k(D) = 1 + 2^k L_1(D).$$

Then for any dataset D' with $d(D, D') = 1$ and any $k \in \mathbb{N}$, we have

$$U_k(D) = 1 + 2^k L_1(D) \leq 1 + 2^{k+1} L_1(D') = U_{k+1}(D')$$

by Theorem 4.6. Thus by Theorem 3.5,

$$\max_k e^{-\beta k} \left[1 + 2^k L_1(D) \right]$$

is a β -smooth upper bound on $LS_q(D)$. This converges if and only if $\beta \geq \ln 2$, in which case it converges to $1 + L_1(D)$.

We emphasize that as a consequence of this theorem, $1 + L_1(D)$ is a $(\ln 2)$ -smooth upper bound on $LS_q(D)$, not just an upper bound. Corollary 3.9 then implies the following (taking $\beta = \ln 2$ and $\gamma = 1 + \frac{\epsilon}{2\beta}$).

Corollary 5.2. *Let q be a counting query and $\epsilon > 0$. Then the mechanism which returns*

$$q(D) + \frac{1 + L_1(D)}{\ln 2} \cdot X,$$

where X is sampled from the generalized Cauchy distribution with parameter $\gamma = 1 + \frac{\epsilon}{2\ln 2}$, satisfies ϵ -differential privacy.

Remark 5.3.

- (a) The noise added in Corollary 5.2 above depends only on ϵ and $L_1(D)$. Moreover, it is quite possible that $GS_q \neq LS_q(D) = L_1(D) + 1$, in which case the local sensitivity and the smooth sensitivity coincide. In such a case, the noise added is proportional to the *local* sensitivity of q . Such a phenomenon is also possible for the median query discussed in Nissim et al. (2007) for certain values of β and certain non-pathological datasets, even if no differentially private mechanism can be based *solely* on the local sensitivity. We also emphasize that the reason β does not appear in our formula is that increasing β beyond $\ln 2$ has no impact on the computation of our smooth sensitivity.
- (b) There is a well-known relaxation of differential privacy called (ϵ, δ) -differential privacy. It is shown in Nissim et al. (2007) that if the Gaussian is used above instead of the generalized Cauchy, the resulting mechanism satisfies this weaker guarantee. Thus our methods can be straightforwardly adapted to this setting.

We can extend this result to proportion queries since they are quotients of counting queries. Thus a data-user interested in a proportion could query the numerator and denominator separately and compute the answer as post-processing. We emphasize that in the case that inclusion in the subpopulation of interest is separate from the attributes in A (for example, if the proportion is computed over the whole dataset), then the global sensitivity of the denominator is 1 and the standard Laplace mechanism can be used.

5.2. Means

We now address the computation of the mean of an attribute computed over a subpopulation of the dataset. As we are using unbounded differential privacy, releasing a privatized version of the mean is non-trivial, even if computed over the whole dataset. Our solution is to compute the size of the subpopulation (a counting query) first and to use this as a parameter in the second mechanism (leveraging sequential composition).

We consider an attribute Y taking values in some range $[a, b]$. We assume $a \geq 0$, but all arguments generalize to allow $a < 0$ in a straightforward way. The values a and b can be thought of as bottom- and top-coded values; without these, the local sensitivity of the mean is unbounded (regardless of whether there is data imputation). For any tuple x , we denote by $Y_D(x)$ the (possibly imputed) value of the attribute Y for the tuple x .

For a dataset D , we denote $S \cap D$ the subpopulation over which we wish to compute the mean. We remark that if $d(D, D') = 1$, it is possible that $|(S \cap D) \Delta (S \cap D')| > 1$ if inclusion in the subpopulation is partially determined by attributes in A . We will give two different versions of each result depending on whether or not this is the case.

By abuse of notation, given two datasets D, D' with $D \Delta D' = \{t\}$, we denote by

$$Y(t) = \begin{cases} Y_D(t) & t \in D \\ Y_{D'}(t) & t \notin D \end{cases} \quad \text{don}^{-1}(t) = \begin{cases} \text{don}_D^{-1}(t) & t \in D \\ \text{don}_{D'}^{-1}(t) & t \notin D \end{cases}.$$

If $t \in T_i$ (i.e., it is data-incomplete), then $\text{don}^{-1}(t) = \emptyset$. Also observe $\text{don}^{-1}(t) \leq L_1(D)$.

Theorem 5.4. *Let s be an estimate of $|S \cap D|$ and let $q(D) = \frac{1}{s} \sum_{x \in S \cap D} Y_D(x)$.*

- (a) *If the subpopulation $S \cap D$ does not depend on the (possibly imputed) values of the attributes in A , then for $\beta \geq \ln 2$, a β -smooth upper bound on $LS_q(D)$ is given by*

$$SS_{\beta,q}(D) = \frac{b + L_1(D)(b - a)}{s}.$$

- (b) *If the subpopulation $S \cap D$ is determined by the attributes in A , then for $\beta \geq \ln 2$, there is a β -smooth upper bound on $LS_q(D)$ given by*

$$SS_{\beta,q}(D) = \frac{b[1 + L_1(D)]}{s}.$$

PROOF. (a) Let D, D' be neighboring datasets with $D \Delta D' = \{t\}$. If t is not in the subpopulation of interest, then there is nothing to show. If t is in the subpopulation of interest, then $S \cap D$ and $S \cap D'$ differ by precisely 1 element. We then have

$$\begin{aligned}
|q(D) - q(D')| &= \frac{1}{s} \cdot \left| \sum_{x \in S \cap D} Y_D(x) - \sum_{x \in D'} Y_{D'}(x) \right| \\
&= \frac{1}{s} \cdot \left| \pm Y(t) + \sum_{x \in S \cap \text{don}^{-1}(t)} (Y_D(x) - Y_{D'}(x)) \right| \\
&\leq \frac{1}{s} \cdot \left[Y(t) + \sum_{x \in S \cap \text{don}^{-1}(t)} |Y_D(x) - Y_{D'}(x)| \right] \\
&\leq \frac{b + L_1(D) \cdot (b - a)}{s}.
\end{aligned}$$

Now let

$$U_k(D) = \frac{b + 2^k L_1(D) \cdot (b - a)}{s}.$$

Then for any $k \in \mathbb{N}$, we have

$$U_k(D) = \frac{b + 2^k L_1(D) \cdot (b - a)}{s} \leq \frac{b + 2^{k+1} L_1(D') \cdot (b - a)}{s} = U_{k+1}$$

by Theorem 4.6. Thus by Theorem 3.5,

$$\max_k e^{-\beta k} \left[\frac{b + 2^k L_1(D) \cdot (b - a)}{s} \right]$$

is a β -smooth upper bound on $LS_q(D)$. This converges if and only if $\beta \geq \ln 2$, in which case it converges to $\frac{b + L_1(D) \cdot (b - a)}{s}$.

The proof of (b) is similar. The key difference is that now any tuple in $\text{don}^{-1}(t)$ can move into or out of the subpopulation of interest with the addition or deletion of t . Thus their contributions to the sum can each change by b , rather than $b - a$ as before.

We observe that if Y is a binary attribute, then $b = 1$ and $a = 0$, so we recover the counting query result from the previous section.

By Corollary 3.8, we then have the following (taking $\beta = \ln 2$ and $\gamma = 1 + \frac{\varepsilon}{2\beta}$)

Corollary 5.5. *Let s be an estimate of $|S \cap D|$, $q(D) = \frac{1}{s} \sum_{x \in S \cap D} Y_D(x)$ and $\varepsilon > 0$.*

(a) *If the subpopulation $S \cap D$ does not depend on the attributes of A , then the mechanism that returns*

$$q(D) + \frac{b + L_1(D) \cdot (b - a)}{s \ln 2} \cdot X,$$

where X is sampled from the generalized Cauchy distribution with parameter $\gamma = 1 + \frac{\varepsilon}{2\ln 2}$, satisfies ε -differential privacy.

(b) *If the subpopulation $S \cap D$ depends on the attributes of A , then the mechanism that returns*

$$q(D) + \frac{[1 + L_1(D)] \cdot b}{s \ln 2} \cdot X,$$

where X is sampled from the generalized Cauchy distribution with parameter $\gamma = 1 + \frac{\epsilon}{2 \ln 2}$, satisfies ϵ -differential privacy.

As in the previous section, we observe that for some datasets, the noise may be proportional to the local sensitivity.

5.3. Variances

Recall the sample variance of the attribute Y computed over the subpopulation $S \cap D$ is

$$S_Y(D) = \frac{1}{|S \cap D| - 1} \sum_{x \in S \cap D} (Y_D(x) - \bar{Y}(S \cap D))^2.$$

We suppose that the mean $\bar{Y}(S \cap D)$ has already been computed using the mechanism in Section 5.2. We denote the returned value by \bar{Y} , which we will take as a parameter in our next mechanism. Moreover, this means the quantity $|S \cap D|$ has already been computed as well. As before, we denote the returned value s . As in the previous section, we assume the attribute Y takes values in $[a, b]$. Moreover, we assume that our estimate of \bar{Y} is also in the interval $[a, b]$.

Theorem 5.6. *Let s be an estimate of $|S \cap D|$ and \bar{Y} an estimate of $\bar{Y}(S \cap D)$ with $a \leq \bar{Y} \leq b$. Let*

$$q(D) = \frac{1}{s-1} \sum_{x \in S \cap D} (Y_D(x) - \bar{Y})^2 \quad m = \max \{(a - \bar{Y})^2, (b - \bar{Y})^2\}.$$

Then for any $\beta \geq \ln 2$, there is a β -smooth upper bound on $LS_q(D)$ given by

$$SS_{\beta,q}(D) = \frac{m}{s-1} [1 + L_1(D)].$$

We note that the upper bound is the same regardless of whether inclusion in $S \cap D$ is independent of the attributes in A .

PROOF. Let D, D' be neighboring datasets with $D \Delta D' = \{t\}$. As before, there is nothing to show if t is not in the subpopulation of interest. Thus assume it is and choose $x \in \text{don}^{-1}(t)$. Suppose that with the addition or deletion of t , the tuple x moves into (resp. out of) the population of interest. Then x contributes $(Y_D(x) - \bar{Y})^2$ to the sum in $q(D)$ (resp. $q(D')$) and 0 to the sum in $q(D')$ (resp. $q(D)$). Thus the change in its contribution is bounded above by m . If, on the other hand, t does not move into (resp. out of) the population of interest (as is the case when inclusion in $S \cap D$ is independent of the attributes in A), then its contribution to the sum changes from $(Y_D(x) - \bar{Y})^2$ to $(Y_{D'}(x) - \bar{Y})^2$. This change is readily bounded above by m .

Thus by analogous reasoning to in the previous section, we have that

$$|q(D) - q(D')| \leq \frac{m}{s-1} [1 + L_1(D)].$$

Now let

$$U_k(D) = \frac{m}{s-1} \cdot [1 + 2^k L_1(D)].$$

Then for any $k \in \mathbb{N}$, we have

$$U_k(D) = \frac{m}{s-1} \cdot [1 + 2^k L_1(D)] \leq \frac{m}{s-1} \cdot [1 + 2^{k+1} L_1(D')] = U_{k+1}$$

by Theorem 4.6. Thus by Theorem 3.5,

$$\max_k e^{-\beta k} \left[\frac{m}{s-1} [1 + 2^k L_1(D)] \right]$$

is a β -smooth upper bound on $LS_q(D)$. If and only if $\beta \geq \ln 2$ this converges to

$$\frac{m}{s-1} [1 + L_1(D)].$$

By Corollary 3.8, we then have the following (taking $\beta = \ln 2$ and $\gamma = 1 + \frac{\epsilon}{2\beta}$).

Corollary 5.7. *Let s be an estimate of $|S \cap D|$ and \bar{Y} an estimate of $\bar{Y}(S \cap D)$ with $a \leq \bar{Y} \leq b$ and let $\epsilon > 0$. The the mechanism which returns*

$$q(D) + \frac{m \cdot [1 + L_1(D)]}{(s-1) \ln 2} \cdot X,$$

where X is sampled from the generalized Cauchy distribution with parameter $\gamma = 1 + \frac{\epsilon}{2\ln 2}$, satisfies ϵ -differential privacy.

As in the previous section, we observe that for some datasets, the noise may be proportional to the local sensitivity. This type of construction can be extended to correlation coefficients, taking means and variance as inputs to the query.

6. Empirical Demonstration

We show concrete examples of these results using the 1940 U.S. Census dataset released for testing disclosure avoidance methodologies (Ruggles et al., 2018). We test the impact of imputation on proportion and mean queries, showing the bias/variance trade-off that imputation is designed to improve. We used the 1940 Census data as a ground truth. For simplicity and efficiency we show results for the state of Minnesota; as U.S. Census Bureau imputation is done at a state level or finer, this reflects real-world use.

Our experiments impute wage income of individuals, as this and the variables used for imputation in modern counterparts were present in 1940 Census data. This data is of practical importance and therefore it is valuable to elucidate the impact that imputation and this form of differential privacy would have on the resulting data.

6.1. Data Creation

While the 1940 Census data does contain missing data, we want to compare against a known ground truth. We instead ignore actual missing data and instead model missing values from data complete tuples to give a ground truth. We leverage the similarity of the variables between the 1940 Census and current American Community Survey (ACS) Public-Use Microdata Samples to simulate missing values.

We first mapped schemas of the 2016 and 1940 data due to differences in specificity between the two surveys. For example, the relationship to householder differs significantly between the 2016 data and the 1940 data, with the ACS providing 18 different responses for “RELP” while the 1940s data “RESPOND” provides only 8 options. This mapping was a joint refinement of the possible responses to various attributes, and was done strictly on the domain of the variables without reference to the data. For example, in homogenizing the relationship to householder attribute, we mapped the responses “boarder”, “roommate”, “other non-related adult” to the single value of “non-related adult” in the 1940 data. In some cases the 1940 data provided more granularity. E.g., the “EMPSTATD” attribute for employment status responses for “not in labor force” in 1940 provide reasons - “housework”, “unable to work”, “schooling”, “other”. Homogenization on both sides merged similar groups to the greatest extent possible.

We trained a model to learn the probability that the income value was missing in the 2016 data based on the homogenized attributes, validating the model on the 2017 1-year PUMS data. We applied the trained model to each data complete record in the 1940 dataset, predicting the likelihood that the data would be missing. Each run of the experiment generated both a new random sample of the data (simulating a sample-based survey), and flagged a new set of tuples as missing income based on the modeled likelihoods. Thus our experiments capture variance based on sampling, randomness in missing data, and the noise required to satisfy differential privacy.

6.2. Imputation

The imputation uses the same variables that ACS currently uses to impute wage per Minnesota Population Center (2018). We use (1-)Nearest Neighbor, which requires a definition of distance in order to determine the “closest” neighbor. As our data has a mix of categorical and ordinal attributes, we had to modify the attributes again to allow for a more meaningful idea of distance. We categorized all of the attributes as either ordinal or categorical, and the ordinal attributes were untouched. The categorical variables were split into a number of binary variables equal to the number of options, e.g., relationship to householder was split into 7 binary values. We then used Euclidean distance for calculating distances between records. This has the effect of allowing the ordinal variables to preserve their distance while treating categorical attributes as edit distance, albeit with a penalty that the resulting distance is 2 instead of 1. This means that the imputation has a small preference for small changes to AGE or SCHOOL over changes to the categorical values. The only modification to an ordinal variable was group ages by decade. This limits the scope of the attribute and facilitates efficient calculation of $L_1(D)$, the crucial value required to determine the impact of the imputation on privacy.

Our imputation first creates groups of donors and non-donors and then uses a tie-breaker to create an ordering for all records in the group. In our case the tie-breaker used was the row number in the original database and records were assigned to the lowest donor with a value higher than the record. To facilitate discussion of the possible options, we will first discuss our use of equivalence classes in the experiments. Given the known attributes that are used in the imputation, it is possible to create equivalence classes of all donors and non-donors that match on those attributes, and our discussion of “groups” in this section will rely on that definition. Once we have divided the non-donors

into groups, we can identify whether any group has at least one donor or not.

We identify 4 cases of which $L_1(D)$ is the maximum. Note that in Remark 4.5 we interpreted $L_1(D)$ as a maximum over two possibilities; it will be convenient here to further expand on the impact of an addition. The 4 possible options are outlined below:

- (a) Remove an existing donor
- (b) Add a new donor to a group with an existing donor
- (c) Add a new donor to a group without an existing donor (but containing non-donors)
- (d) Add a new donor in a unique location

The impact of case (a) is maximized by removing the maximal donor of the existing dataset. Adding a record to a group with an existing donor cannot change more records than deleting the maximal donor in that group. The tie-breaker described above is easily calculated and ensures that the impact of case (b) never exceeds the impact of case (a).

Option (c) requires a more nuanced calculation. By placing a new donor in an existing group that lacks a donor, that donor now donates to the entire group. The rest of the calculation comes from the impact this new donor can have on other groups that don't have an existing donor. For each other group without an existing donor, we calculate the distance to this new donor and then determine what, if any, records would have been imputed on by a maximal donor in this new world.

The calculation for a move of type (d) requires exploring every possible equivalence class not currently represented in the data and determining the impact of a new donor placed there. Even for the relatively limited scope of our attributes, the brute-force calculation required for determining this value exactly was too expensive. As such, we bound this number from above in a way that *for our experiments* was smaller than the contributions from moves (a)-(c). We first find pairs of groups without an existing donor that are pairwise close to each other, The sizes of the unions of such groups provides an upper-bound for the maximum impact of option (d). As discussed previously, a large part of the motivation in grouping age by decade was to guarantee that this upper bound for case (d) was small. This ensures the value of $L_1(D)$ we computed is exact.

6.3. Experiment

Given the lack of viable differentially private approaches to dealing with missing data, we compare against ignoring missing data and a Nearest Neighbor imputation using global sensitivity with the Laplace mechanism. We provide two example queries: what proportion of the records had an income that was below the “poverty line,” and the mean income. As there is no “official” poverty line for the 1940 data we used \$658 as determined by Barrington (1997) as the necessary income to support a family of four above the poverty level in 1940.

It is known that ignoring missing data for income leads to biased results (Nicoletti and Peracchi, 2006); this example is useful since it shows that the increased variance of our mechanism is more than made up for by the reduced bias. Our comparison to a naive global sensitivity differential privacy scheme requires some explanation about the optimistic assumptions made. It is easy to see that global sensitivity for an imputation scheme will provide unusable amount of noise (see Figure 1). Our version of global

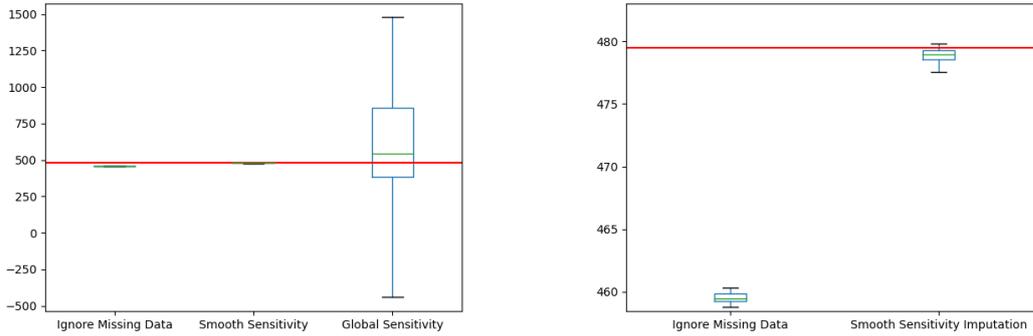


Fig. 4. Mean individual income, ages 20-59, with $(6 \ln 2)$ -differential privacy. Box reflects inter quartile range, whisker is 5-th and 95-th percentiles. Full width line is the true value. “Ignore Missing Data” means missing data was discarded and the Laplace mechanism was used, “Smooth Sensitivity” (resp. “Global Sensitivity”) means missing data was imputed and our smooth sensitivity-based mechanism (resp. the Laplace mechanism) was used.

sensitivity assumes that the maximum number of records a donor can donate to is the number of missing records; i.e., we limit the impact from a single person donating to the entire state to a single individual donating to every value that was missing from the state. This is an optimistic form of global sensitivity for the problem, but we included it to highlight that imputation creates significant challenges for differential privacy.

As we see in Figure 4, the global sensitivity of imputation requires unreasonable noise; better results are obtained by ignoring missing values. Our global sensitivity scheme has a Mean Squared Error of 389899.14 compared to simply ignoring the missing data having a MSE with 397.25 and our nearest neighbor scheme has an MSE of 1.3. Ignoring missing data results in a value substantially lower than the true result. Our smooth sensitivity imputation method gives results close to the true mean, with only slightly higher variance than that induced by the random variation in which individuals have missing values. Global sensitivity results were similar on other queries, and are omitted to highlight comparison of our method with ignoring missing data.

Figure 5 shows the same query, but just for individuals in their 20s or 40s. For the 20 year old query, ignoring the data provides a MSE of 116.9 compared to our imputation strategy providing a MSE of 7.3. For the 40-year olds ignoring provides a MSE of 1180.3 and our mechanism provides a MSE of 11.1. We see that missing data has a larger impact on queries of those in their 40s, but imputation still largely removes this impact. Figure 6 shows a different query on the same data; the proportion of individuals who make enough to support a family of four above the poverty line. We again see significantly better results for smooth sensitivity imputation. Ignoring the missing data for 20-year olds provides a MSE of 33×10^{-6} while our imputation strategy has a MSE of 1.4×10^{-6} . For the 40 year olds, ignoring the missing data provides a MSE of 4.6×10^{-4} and our imputation strategy provides a MSE of 7.5×10^{-7} .

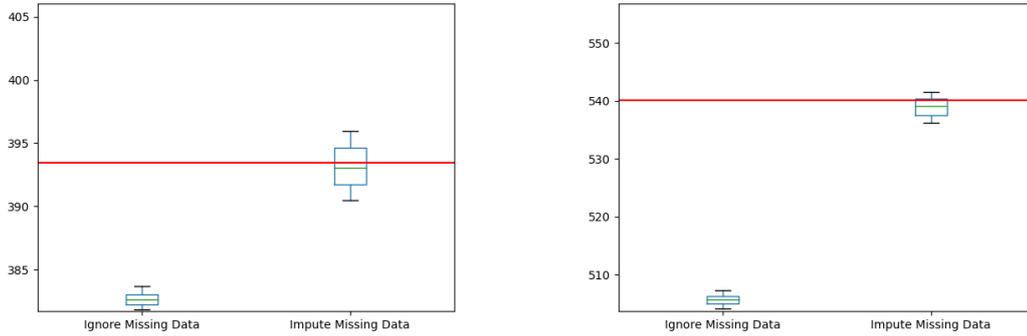


Fig. 5. Mean individual incomes for persons 20-29 years old (left) and 40-49 years old (right), with $(6 \ln 2)$ -differential privacy.

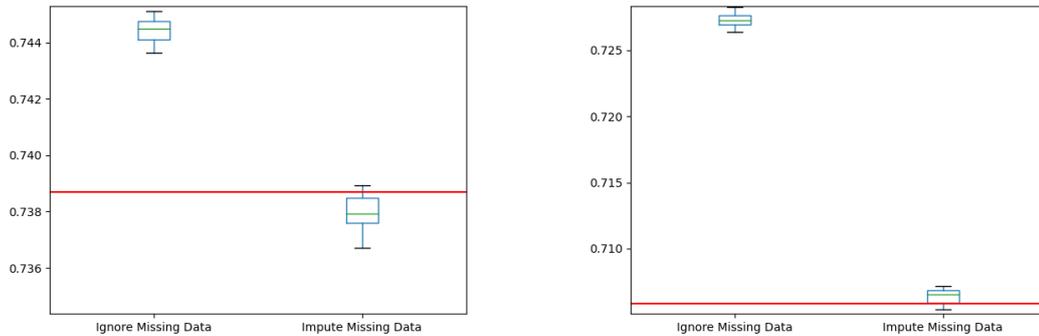


Fig. 6. Proportions of adults age 20-29 (left) and 40-49 (right) who make enough to support a family of 4 above the poverty level, with $(6 \ln 2)$ -differential privacy.

7. Conclusions and Future Work

The problem of missing survey data presents an interesting privacy challenge for data curators: ignoring the missing values tends to yield biased results, but imputation methods can dramatically increase an individual’s impact on the dataset, thus increasing the likelihood of reidentification. Global sensitivity-based mechanisms require an untenable amount of noise, even under unreasonable assumptions. We have advocated for an approach based on smooth sensitivity to mitigate these issues. To do this, we developed a smooth upper bound which is far more computationally tractable in many cases than computing the local sensitivity an arbitrary number of steps away; a technical contribution to the differential privacy literature in its own right.

The extent to which this approach can be pushed requires analysis of other queries where changes give a bounded $L_1(D)$. One could also look for similar quantities dependent only on the present dataset D from which similar bounds could be derived.

Each of our mechanisms answers a single query privately on a dataset with missing data. For circumstances in which a data curator seeks to answer a large quantity of queries on such data, it would be interesting to investigate the compatibility of our mechanism with approaches like the high dimensional matrix mechanism, PriView, or differentially private synthetic data releases.

Acknowledgements

We wish to thank Mark E. Asiala and Edward C. Castro, Jr. for helpful discussions, particularly in identifying appropriate uses cases for Section 6; and Rolando Rodriguez and Mark Fleischer for extensive suggestions on the manuscript. This work supported by the U.S. Census Bureau under CRADA CB16ADR0160002. The views and opinions expressed in this writing are those of the authors and not the U.S. Census Bureau.

References

- Bailar III, J. C. and Bailar, B. A. (1978) Comparison of two procedures for imputing missing survey values. In *Proceedings of the Survey Research Methods Section*, 462–467. American Statistical Association. URL: <http://www.asasrms.org/Proceedings>.
- Barrington, L. (1997) Estimating earnings poverty in 1939: A comparison of orshansky-method and price-indexed definitions of poverty. *The Review of Economics and Statistics*, **79**, 406–414. URL: <http://www.jstor.org/stable/2951387>.
- Chawla, S., Dwork, C., McSherry, F., Smith, A. and Wee, H. (2005) Toward privacy in public databases. In *Theory of Cryptography Conference*. Cambridge, MA. URL: <http://research.microsoft.com/research/sv/DatabasePrivacy/public.ps>.
- Chiang, F. and Gairola, D. (2018) Infoclean: Protecting sensitive information in data cleaning. *Journal of Data and Information Quality (JDIQ)*, **9**, 22. URL: <https://doi.org/10.1145/3190577>.
- Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In *Proc. of the 3rd Theory of Cryptography Conf.*, 265–284.
- Fletcher, S. and Islam, M. Z. (2017) Differentially private random decision forests using smooth sensitivity. *Expert Systems with Applications*, **78**, 16–31. URL: <https://doi.org/10.1016/j.eswa.2017.01.034>.
- Ge, C., Ilyas, I. F., He, X. and Machanavajjhala, A. (2018) Private exploration primitives for data cleaning. *Tech. Rep. 1712.10266*, arXiv. URL: arxiv.org/abs/1712.10266.
- Gonem, A. and Gilad-Bachrach, R. (2018) Smooth sensitivity based approach for differentially private pca. In *Proceedings of Algorithmic Learning Theory*, vol. 83, 438–450. URL: <http://proceedings.mlr.press/v83/gonem18a.html>.
- Huang, Y., Milani, M. and Chiang, F. (2018) PACAS: Privacy-aware, data cleaning-as-a-service. In *2018 IEEE International Conference on Big Data (Big Data)*. URL: <https://doi.org/10.1109/BigData.2018.8622249>.

- Jain, P., Thakkar, O. D. and Thakurta, A. (2018) Differentially private matrix completion revisited. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* (eds. J. G. Dy and A. Krause), vol. 80 of *Proceedings of Machine Learning Research*, 2220–2229. PMLR. URL: <http://proceedings.mlr.press/v80/jain18b.html>.
- Kalton, G. and Kasprzyk, D. (1982) Imputing for missing survey responses. In *Proceedings of the Survey Research Methods Section*, 22–33. American Statistical Association. URL: <http://www.asasrms.org/Proceedings>.
- Kapralov, M. and Talwar, K. (2013) On differentially private low rank approximation. In *Proc. 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1395–1414. New Orleans. URL: <https://doi.org/10.1137/1.9781611973105.101>.
- Krishnan, S., Wang, J., Franklin, M. J., Goldberg, K. and Kraska, T. (2016) Private-clean: Data cleaning and differential privacy. In *Proceedings of the 2016 International Conference on Management of Data*, 937–951. San Francisco, California: ACM. URL: <https://doi.org/10.1145/2882903.2915248>.
- McSherry, F. and Mironov, I. (2009) Differentially-private recommender systems: Building privacy into the netflix prize contenders. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France.
- Minnesota Population Center (2018) Introduction to data editing and allocation. <https://usa.ipums.org/usa/flags.shtml>.
- Nicoletti, C. and Peracchi, F. (2006) The effects of income imputation on microanalyses: evidence from the european community household panel. *Journal of the Royal Statistical Society Series A*, **169**, 625–646. URL: <https://doi.org/10.1111/j.1467-985X.2006.00421.x>.
- Nissim, K., Raskhodnikova, S. and Smith, A. (2007) Smooth sensitivity and sampling in private data analysis. In *STOC*, 75–84.
- Okada, R., Fukuchi, K., Kakizaki, K. and Sakuma, J. (2015) Differentially private analysis of outliers. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD 2015)*, 458–473. Porto, Portugal. URL: https://doi.org/10.1007/978-3-319-23525-7_28.
- Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J. and Sobek, M. (2018) IPUMS USA: Version 8.0 extract of 1940 Census for U.S. Census Bureau disclosure avoidance research [dataset]. <https://doi.org/10.18128/D010.V8.0.EXT1940USCB>.
- United States Census Bureau (2014) American Community Survey design and methodology (January 2014). *Tech. Rep. Version 2.0*. URL: <https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html>.
- Wang, Y. and Wu, X. (2013) Preserving differential privacy in degree-correlation based graph generation. *Transactions on Data Privacy*, **6**, 127–145. URL: <http://www.tdp.cat/issues11/abs.a113a12.php>.