# Differential Identifiability*

Jaewoo Lee
Dept. of Computer Science
Purdue University, West Lafayette, IN, USA
jaewoo@cs.purdue.edu

Chris Clifton
Dept. of Computer Science / CERIAS
Purdue University, West Lafayette, IN, USA
clifton@cs.purdue.edu

## ABSTRACT

A key challenge in privacy-preserving data mining is ensuring that a data mining result does not inherently violate privacy. $\epsilon$-Differential Privacy appears to provide a solution to this problem. However, there are no clear guidelines on how to set $\epsilon$ to satisfy a privacy policy. We give an alternate formulation, *Differential Identifiability*, parameterized by the probability of individual identification. This provides the strong privacy guarantees of differential privacy, while letting policy makers set parameters based on the established privacy concept of individual identifiability.

## Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: Privacy; H.2.8 [**Database Applications**]: Data mining

## General Terms

Security

## Keywords

Differential Privacy; Identifiability

## 1. INTRODUCTION

Privacy-preserving data mining has seen many advances, and today we can construct many data mining models without disclosing the input data [1, 21]. One key challenge remains: does the produced model inherently violate privacy? *Differential Privacy* [7] provides a means to address this challenge. The basic idea is to add enough noise to the outcome (e.g., the model resulting from training) to hide the contribution of any single individual to that outcome.

*Support for this work was provided by MURI award FA9550-08-1-0265 from the Air Force Office of Scientific Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*KDD* '12, August 12–16, 2012, Beijing, China
Copyright 2012 ACM 978-1-4503-1462-6 /12/08 ...$15.00.

This appears to address the challenge perfectly. From a data mining perspective, the first privacy issue is revealing information about an individual in the training data. Differential privacy essentially hides an individual by ensuring that the resulting model is nearly indistinguishable from the one without that individual – *for any individual*. Formally, given the query function $f$, $\forall S \subseteq Range(\mathcal{M})$ an $\epsilon$-differentially private mechanism $\mathcal{M}$ satisfies:

$$\frac{Pr[\mathcal{M}_f(D_1) \in S]}{Pr[\mathcal{M}_f(D_2) \in S]} \leq e^{\epsilon} \quad (1)$$

where $D_1$ and $D_2$ are two databases differing by at most one element.[1] In data mining terms, $\mathcal{M}_f$ is the learning algorithm, $f$ is the resulting model, and $D_1$ and $D_2$ are nearly identical training databases. Since the (randomized) learning algorithm guarantees that the results are indistinguishable for **any** two training sets that differ by one individual, we can claim that the resulting model does not disclose information about any single individual in the training set.

Unfortunately, indistinguishability only holds for $\epsilon = 0$, which makes the resulting model useless. The larger the value of $\epsilon$, the less noise needed in the model, and the more a single individual can impact that model (thus risking disclosure of information about that individual). The problem is that $\epsilon$ limits how much one individual can affect the resulting model, not how much information is revealed about an individual. This does not match legal definitions of privacy, which require protection of *individually identifiable data* [8, 11].

We instead propose a parameterization based on the probability that an individual contributes to the model. This corresponds to legal definitions, and we can use regulations such as the U.S. HIPAA safe harbor rule [11] to determine a probabilistic intent of the regulation. The HIPAA safe harbor rule requires removal of names, addresses, and identifying numbers. However, it does allow geographic units as small as 20,000 people, age in years (if less than 90), gender, and ethnicity. Given that less than 1.7% of the U.S. population is male and 85 or older[2], knowing the age, gender, and address of someone over 85 would allow us to limit them to one of 68 people (on average) in safe-harbor de-identified data. In practice, the privacy provided could be much worse; e.g., in a college town there may be few older people. How-

[1]There are several formulations of the definition of differential privacy, such as a difference bound rather than the ratio given in Equation 1; we present the one from [7]. The differences between definitions is not critical to this paper.
[2]U.S. Census, 2010 data

ever, from this we can deduce that the goal of the privacy policy is met if we limit the estimate of the probability that an individual is in the data to approximately 1.5%.

The $\epsilon$ in $\epsilon$-differential privacy does not correspond to such a probability; it has been shown that for a given value of $\epsilon$, the probability of identification can vary depending on data values, or even on values of individuals not in the data set [13]. Nor does $\epsilon$ correspond to ability to infer private data values for an individual [4]. We give a definition *$\rho$-differential identifiability* that provides the same guarantees as differential privacy (proof against an arbitrarily strong adversary), but the parameter $\rho$ bounds the probability estimate that an individual contributed to the resulting model.

Differential Identifiability is a subtle but important variation of differential privacy, and this paper shows that the general Laplacian noise addition mechanism for differential privacy can be adapted to provide differential identifiability. However, the mathematical formulation shows significant differences; there is no direct translation from $\rho$ to $\epsilon$; the relationship depends on additional information outside the scope of setting policy. The result is a method corresponding to real-world privacy policy (e.g., $\rho = .015$ provides privacy equivalent to what appears intended by the HIPAA safe harbor rules), with the strong formal guarantees of the adversary model used in differential privacy.

## 2. RELATED WORK

While many privacy definitions have been proposed, the common goal is to allow learning information about a group of individuals while protecting information of each individual in the group. Samarati and Sweeney proposed $k$-anonymity [19, 20]. $k$-anonymity ensures that the identifying information for at least $k$ tuples is identical, ensuring that individuals cannot be uniquely re-identified.

While $k$-anonymity prevents linking a record to an individual, it may still disclose sensitive information. $\ell$-diversity [14] shows that $k$-anonymous tables are vulnerable to homogeneity. Their definitions require that each equivalence class with identical identifying information have at least $\ell$ distinct values for each sensitive attribute.

While these and related approaches prevent the adversary from uniquely identifying an individual's record, they assume little prior knowledge available to the adversary. $\delta$-presence [18] protects against determining if an individual is in a dataset even if the adversary has full knowledge of values of individuals. Generalization/suppression bounds the adversary's probability of inferring that an individual is in the database to the range $\delta = (\delta_{min}, \delta_{max})$.

A big advantage of the above methods is that the outcomes are guaranteed true, even though specificity may be lost through generalization/suppression. However, they assume the adversary's knowledge is limited. For example, in $k$-anonymity and $\ell$-diversity, protection targets are not met if the adversary knows sensitive values of other individuals.

Differential privacy [7] avoids modeling prior knowledge of the adversary. To do this, it gives up the correctness guarantees of the above methods, instead providing a noisy result that hides the impact of any single individual on the result. The idea is that what is learned from a dataset with a particular individual can also be learned from a dataset without that individual. Therefore, it hides presence or absence of an individual in the database by making the response generated by two datasets (one with the individual and the other without the individual) indistinguishable. As stated previously, we follow in this model, but provide a parameterization based on the risk of identifying an individual.

**Table 1: Notations**

| | |
|---|---|
| $U$ | Universe |
| $D \subset U$ | Database to be queried |
| $D' = D - i^*$ | Subset of $D$ missing one individual |
| $i \in U$ | Data associated with an individual |
| $I(i)$ | Identity of an individual corresponding to $i$ |

## 3. PRELIMINARIES

We now give background on differential privacy, the problem of re-identification, and introduce a possible worlds adversary and sensitivity model. We first introduce notation.

### 3.1 Notation

A database $D$ can be modeled as a (multi)set. Each element $x_i$ takes a fixed value from the universe $U$. Each entry in $U$ corresponds to an individual whose privacy must be protected; $I(i)$ denotes the identity of the individual corresponding to database entry $i$. The set of individuals who contributed their data to $D$ is denoted by $\mathcal{I}_D = \{I(i) | i \in D\}$. Let $D' \subset D$ denote a database having one less element than $D$ (i.e., $|D'| = |D| - 1$). As with differential privacy, a query function $f$ can be any function that extracts information from a database. Examples from differential privacy include aggregate queries (count, mean, sum, ...) used by a data mining algorithm [5, 15, 22], a data anonymizer [17], or a learning algorithm such as ID3 [10] or logistic regression model [3]. These notations are summarized in Table 1.

Another concept borrowed from differential privacy is sensitivity, which measures the maximum impact a single individual can have on the query result:

DEFINITION 1 (SENSITIVITY). *The sensitivity of a query function $f$ for bounded differential privacy [6] is defined as*

$$\Delta f = \max_{x, x' \subset U} \left| f(x) - f(x') \right|$$

*where $x'$ can be obtained by replacing one element in $x$ with another.*

To attempt to dispel confusion, we note that there are different ways of defining sensitivity, based on the definition of neighboring databases $D$ and $D'$ differing by removal of an individual (*unbounded*) or replacement of an individual (*bounded*). Unless we explicitly state otherwise, we use bounded differential privacy in this paper. Further discussion of these differences can be found in [12].

### 3.2 Problem Statement

The goal of privacy-preserving data mining is to release the mining outcome without revealing identities of individuals in the database. Precisely, given $\rho \in [0, 1]$, the adversary's expectation that any individual corresponding to $i \in U$ *not previously known to be in* $D$ is in $D$ is at most $\rho$. This can be thought of as a game between a privacy mechanism and an adversary. The privacy mechanism $\mathcal{M}$ builds the model $f(D)$ and adds noise to produce the perturbed result (*response*) $R = \mathcal{M}_f(D)$. The adversary tries to identify individuals whose data is in $D$ from the given response $R$. If

$R$ enables the adversary to state any new individual belongs to $D$ with confidence exceeding $\rho$, privacy is breached.

## 3.3 Adversary Model

This paper assumes the same strong adversary as differential privacy. The adversary's prior knowledge is represented as a triplet $\mathcal{L} = \langle U, D', \mathcal{I}_{D'} \rangle$. The adversary has complete knowledge on the universe; every value in $U$ is known. The adversary knows every tuple in $D$ except one. In other words, the adversary has $D'$. The adversary also has $\mathcal{I}_{D'}$, the identities of the individuals corresponding to those tuples. The only piece of information the adversary does not have is who the $n^{th}$ individual of $\mathcal{I}_D$ is. The adversary also knows the privacy mechanism $\mathcal{M}_{\{}$, i.e., how the mechanism works and the noise distribution. This type of adversary is called an *informed adversary* [6].[3]

In our model, the goal of adversary is to determine the membership of the unknown individual in $D$ with high confidence. To find out the identity of the missing individual, the adversary interacts with a randomized mechanism $\mathcal{M}$; issues a query and receives a noisy response $R = \mathcal{M}_f(D)$. The adversary uses the response and prior knowledge to reduce uncertainty about the missing individual.

To measure the adversary's confidence in making an inference, we use a possible worlds model. The adversary considers the set of all possible databases. Given the prior knowledge $\mathcal{L}$ and $R$, the adversary creates a set of all possible databases, called *possible worlds*, that may be the $D$ that generated the perturbed answer $R$. Since the adversary only needs to determine the membership of one missing individual, each possible world is the union of $D'$ and one data entry from $U$. Given the adversary's prior knowledge $\mathcal{L}$, the set of all possible worlds, denoted by $\Psi$, is

$$\Psi = \{D' \cup \{i\} | i \in U \wedge i \notin D'\}$$

Notice that exactly one of the worlds in $\Psi$ is the true database that produced $R$. When the adversary knows $k$ rows of $D$, the size of $\Psi$ is $|\Psi| = \binom{|U|-k}{n-k}$. Let $\Psi_{\{i\}}$ and $\Psi^C_{\{i\}}$ denote the set of possible worlds that contains and doesn't contain a data entry $i$, respectively.

Before seeing the response $R$, we assume every possible world is equally likely to be $D$ (we discuss relaxing this assumption at the end of Section 4.1.) Once the adversary receives $R$, for each possible world $\omega \in \Psi$, the adversary computes the probability that $\omega$ is the original database $D$ that generated the perturbed response $R$:

$$Pr\left[\omega = D | \mathcal{M}_f(D) = R\right]$$

These probabilities give the adversary an updated belief on each possible world. Among all possible worlds, the one with the highest probability will be the adversary's "best guess". If the mechanism allows the adversary to make a correct guess on the missing data entry $i$ with high confidence, the privacy of the individual corresponding to $i$ is at risk. Therefore, the goal of our privacy mechanism is to bound the probability that the adversary identifies an individual's presence in the database.

[3]As with differential privacy, we assume the adversary does *not* know $U - D$, the individuals not in the dataset, as knowing $U - D$ and $D'$ reveals $D$ and there is no privacy left to protect. Assuming a large universe $U$, knowledge of a few individuals not in $D$ has little impact, so the assumption that the adversary knows $D'$ is a stronger practical assumption.

## 3.4 Sensitive Range

To determine the noise needed to hide the impact of any individual, differential identifiability (and differential privacy) use the *sensitivity* of the result to the contribution of any single individual. The greater the variation from the contribution of a single individual, the more noise must be added. Given the query function $f$, the contribution of an individual can be stated as the change in the range of $f$ due to having that individual in its domain. The formal definition of individual contribution is:

DEFINITION 2 (CONTRIBUTION OF AN INDIVIDUAL). *The contribution of an individual corresponding to $i$ to a query function $f$, $\mathcal{C}_f(i)$, is defined as:*

$$\mathcal{C}_f(i) = \max_{\omega_1, \omega_2} \|f(\omega_1) - f(\omega_2)\|$$

*where $\omega_1 \in \Psi_{\{i\}}$ and $\omega_2 \in \Psi^C_{\{i\}}$.*

To capture the largest contribution that can be made by any single individual in the universe, the *sensitive range* is defined as the range of a query function over the domain $\Psi$. This is the maximum distance over the range of $f$ between two possible worlds, *for any set of possible worlds* (since we don't know which individual the adversary doesn't know.) The maximum contribution is the largest contribution that can be made by a single individual:

DEFINITION 3 (SENSITIVE RANGE $\mathcal{S}(f)$). *The sensitive range of a query function $f$ is the range of $f$.*

$$\mathcal{S}(f) = \max_{\omega_1, \omega_2 \in \Psi} \|f(\omega_1) - f(\omega_2)\|$$

*where $\Psi$ is the set of possible worlds under the prior knowledge $\mathcal{L}$.*

Note that $\mathcal{S}(f) = \max_i \mathcal{C}_f(i)$.

## 3.5 Disclosure Risk

Our goal is to hide identities of data contributors; we measure the adversary's computed probability that any given individual is in the database. This is called *identifiability risk*. In an interactive privacy mechanism, the information to be disclosed is the answer to an aggregate query that does not relate to a specific individual. However, even though published statistics are presented in aggregated form, it is still possible they leak some information about an individual. This is especially true when the database contains an individual whose contribution to the query answer is significantly larger (or smaller) than that of others.

To illustrate, we give an example where the mean $\mathcal{M}$ released by a differentially private mechanism enables the adversary to guess the missing element with high probability. Given $D = \{1, 2, 3, 10\}$ drawn from $U = \{1, 2, 3, 4, 5, 10\}$, the sensitivity of the query function $f$ = mean is $\frac{9}{4}$. Assume the adversary already knows $\{1, 2, 3\} \subset D$. The possible worlds are $\omega_1 = \{1, 2, 3, 4\}$, $\omega_2 = \{1, 2, 3, 5\}$, and $\omega_3 = \{1, 2, 3, 10\}$. Assume that $\epsilon = 2$ and the response $R = 5.041$. The adversary computes the probability $Pr[\mathcal{M}_f(\omega_i){=}R], 1{\leq}i{\leq}3$ that $R$ came from each distribution, and compares their ratio. $Pr[\mathcal{M}_f(\omega_3){=}5.041] = 0.1762$ is much larger than the other two, $Pr[\mathcal{M}_f(\omega_1){=}5.041] = 0.0464$ and $Pr[\mathcal{M}_f(\omega_2){=}5.041] = 0.0580$. The adversary concludes that the missing element is 10 with confidence $\frac{Pr[\mathcal{M}_f(\omega_3)=R]}{\sum_i Pr[\mathcal{M}_f(\omega_i)=R]} = \frac{0.1762}{0.1762+0.0464+0.0580} = 0.6278$. (For the same $R$ and a smaller value of $\epsilon$ the confidence would be lower; see [13].)

# 4. DIFFERENTIAL IDENTIFIABILITY

Given the above, we can now provide a formal definition that satisfies the problem statement given in Section 3.2.

DEFINITION 4 ($\rho$-DIFFERENTIAL IDENTIFIABILITY). *Given a query function $f$, a randomized mechanism $\mathcal{M}$ is said to satisfy $\rho$-differential identifiability if for all databases $D$, $\forall D' = D - i^*$, and $\forall i \in U - D'$:*

$$Pr[I(i) \in \mathcal{I}_D | \mathcal{M}_f(D) = R, D'] \leq \rho$$

A randomized mechanism $\mathcal{M}$ satisfying the above definition ensures that the identifiability risk of any individual in the universe is less than or equal to $\rho$. Basically, every possible world becomes indistinguishable within a factor of $\rho$. The parameter $\rho$ in our work can be interpreted as the degree of indistinguishability between possible worlds, where the possible worlds differ by (any) one individual – providing an upper bound on the confidence that the individual is the difference between the worlds (and thus identifiable.) This differs from the $\epsilon$ in differential privacy, which measures the difference in the query result given different possible worlds, not the difference in the likelihood of those worlds.

## 4.1 Achieving Differential Identifiability

We now show how to calibrate noise to achieve $\rho$-differential identifiability, given the sensitive range of a query function. As with the mechanism for differential privacy introduced in [6], noise $Y$ is added to every query response, $R = f(D) + Y$, where $Y$ is an i.i.d. random variable drawn from a Laplace distribution.

Let $\Gamma(i)$ be the identifiability risk for the individual $I(i)$. $\Gamma(i)$ represents the degree to which an adversary believes $I(i)$ is in $D$ given $\mathcal{M}_f(D) = R$. An upper bound on $\Gamma(i)$ can be obtained as follows:

$$\Gamma(i) = Pr[I(i) \in \mathcal{I}_D | \mathcal{M}_f(D){=}R, D'] \tag{2}$$

$$= Pr[D{=}D'{\cup}\{i\} | \mathcal{M}_f(D){=}R, D'] \tag{3}$$

$$= \frac{Pr[D{=}D'{\cup}\{i\}]}{Pr[\mathcal{M}_f(D){=}R]} \cdot Pr[\mathcal{M}_f(D){=}R | D{=}D'{\cup}\{i\}] \tag{4}$$

$$= \frac{Pr[D{=}D'{\cup}\{i\}] \cdot Pr[\mathcal{M}_f(D' \cup \{i\}){=}R]}{\sum_{\omega \in \Psi} Pr[\omega] \cdot Pr[\mathcal{M}_f(\omega){=}R]} \tag{5}$$

$$= \frac{e^{-\frac{|R-f(D)|}{\lambda}}}{e^{-\frac{|R-f(D)|}{\lambda}} + \sum_{\omega \in \Psi, \omega \neq D} e^{-\frac{|R-f(\omega)|}{\lambda}}} \tag{6}$$

$$= \frac{e^{-\frac{|R-f(D)|}{\lambda}}}{e^{-\frac{|R-f(D)|}{\lambda}} + e^{-\frac{|R-f(\omega_{j_1})|}{\lambda}} + \ldots + e^{-\frac{|R-f(\omega_{j_{m-1}})|}{\lambda}}} \tag{7}$$

Dividing numerator and denominator by $e^{-\frac{|R-f(D)|}{\lambda}}$ gives

$$\leq \frac{1}{1 + e^{-\frac{|f(D)-f(\omega_{j_1})|}{\lambda}} + \ldots + e^{-\frac{|f(D)-f(\omega_{j_{m-1}})|}{\lambda}}} \tag{8}$$

Since $\forall k, 1 \leq k \leq m, |f(D) - f(\omega_{i_k})| \leq \mathcal{S}(f)$, simple application of triangle inequality yields

$$\leq \frac{1}{1{+}(m{-}1) \cdot e^{-\frac{\mathcal{S}(f)}{\lambda}}} \tag{9}$$

where $m = |\Psi| = |U| - |D'|$.

Since $(m-1) \cdot \exp(-\frac{\mathcal{S}(f)}{\lambda}) \leq (m-1)$, it is trivial to see that the lower bound of Equation (9) is

$$\frac{1}{1 + (m-1)} = \frac{1}{m} = \frac{1}{|\Psi|} = \frac{1}{|U| - |D'|}$$

This implies that it is impossible to protect the privacy of individuals in the database with the probability less than an adversary's probability of a correct random guess.

To find $\lambda$ that ensures $\forall i, \Gamma(i) \leq \rho$, it is sufficient to satisfy the following inequality

$$\frac{1}{1 + (m-1) \cdot \exp\left(-\frac{\mathcal{S}(f)}{\lambda}\right)} \leq \rho \tag{10}$$

$$1 + (m-1) \cdot \exp\left(-\frac{\mathcal{S}(f)}{\lambda}\right) \geq \frac{1}{\rho} \tag{11}$$

$$\exp\left(-\frac{\mathcal{S}(f)}{\lambda}\right) \geq \frac{1-\rho}{(m-1)\rho} \tag{12}$$

Since $\rho \leq 1$, taking the natural log of both sides yields

$$-\frac{\mathcal{S}(f)}{\lambda} \geq \ln \frac{1-\rho}{(m-1)\rho} \tag{13}$$

$$\frac{\mathcal{S}(f)}{\lambda} \leq \ln \frac{(m-1)\rho}{1-\rho} \tag{14}$$

When $\frac{(m-1)\rho}{1-\rho} \leq 1$, i.e.,$\rho \leq \frac{1}{m}$, (14) can never be satisfied. For $\rho \geq \frac{1}{m}$,

$$\lambda \geq \frac{\mathcal{S}(f)}{\ln \frac{(m-1)\rho}{1-\rho}} \tag{15}$$

The above leads to (and serves as a proof of):

THEOREM 1. *For an arbitrary adversary $\mathcal{A}$, if $\lambda = \frac{\mathcal{S}(f)}{\ln \frac{(m-1)\rho}{1-\rho}}$, $\mathcal{M}$ satisfies $\rho$-differential identifiability where $m = |\Psi|$.*

The construction given assumes that the prior probability of an individual being in $D$ is the same for all individuals. In practice, some individuals may have a higher prior. This may make providing $\rho$-differential identifiability impossible: If the prior probability $Pr[D = D' \cup \{i\}] > \rho$, then the privacy goal is inherently violated and no privacy mechanism can restore it. For less severe cases, the value for $m$ in Equation (15) can be replaced with $1/\max_\Psi(Pr[D = D' \cup \{i\}])$. This essentially says that the adversary's best guess is no better than the average calculated in Equation (5).

While this seems to require additional knowledge of the capabilities of an adversary, it actually shows that differential privacy does not provide guarantees on individual identifiability. Even if an adversary already has sufficient prior knowledge to identify an individual, differential privacy blindly produces the same (noisy) result as if the adversary had no such prior knowledge. The protection of differential privacy measures only the impact of an individual on the output, not the ability to identify an individual [13], or even the ability to infer data values for an individual [4].

Under the assumption that every possible world is equally likely and the number of possible worlds is known, there is a relationship between our definition and differential privacy. Any $\epsilon$-differentially private mechanism satisfies $\frac{1}{1+(m-1)e^{-\epsilon}}$-differential identifiability. This is easy to see once we apply

**Table 2: Possible worlds**

| Possible world | mean | median |
|---|---|---|
| $\omega_1 = \{1, 2, 3\}$ | 2 | 2 |
| $\omega_2 = \{1, 3, 4\}$ | $\frac{8}{3}$ | 3 |
| $\omega_3 = \{1, 3, 5\}$ | 3 | 3 |
| $\omega_4 = \{1, 3, 6\}$ | $\frac{10}{3}$ | 3 |
| $\omega_5 = \{1, 3, 7\}$ | $\frac{11}{3}$ | 3 |
| $\omega_6 = \{1, 3, 8\}$ | 4 | 3 |
| $\omega_7 = \{1, 3, 9\}$ | $\frac{13}{3}$ | 3 |
| $\omega_8 = \{1, 3, 10\}$ | $\frac{14}{3}$ | 3 |

the differential privacy guarantee ,$\forall j, Pr[\mathcal{M}_f(\omega_j) = R] \geq e^{-\epsilon} Pr[\mathcal{M}_f(D) = R]$, to Equation (6). Alternatively, $\epsilon \leq \ln \frac{(m-1)\rho}{1-\rho}$. This provides an upper bound on $\epsilon$ for the given $\rho$, not an optimal value. Note that even setting this upper bound for $\epsilon$ requires knowing the number of possible worlds $m$, making it difficult to develop a policy for a value of $\epsilon$ to prevent re-identification.

Since the upper bound on the identifiability risk is computed by approximating the distances between the original database and other possible worlds, $|f(D) - f(\omega_k)|$, with $\mathcal{S}(f)$, the calibrated noise could be greater than what is actually required. For example, assume $U = \{1, 2, \cdots, 10\}$, $D = \{1, 2, 3\}$, and the subset of $D$ known to the adversary is $D' = \{1, 3\}$. Since the adversary already knows 1 and 3 belong to $D$ and only needs to determine the membership of one missing element, the possible worlds generated by the adversary would be Table 2. Given the query function $f_{mean}(X)$ returning the mean of elements in $X$, the sensitive range is:

$$\mathcal{S}(f_{mean}) = \left| \frac{14}{3} - 2 \right| = \frac{8}{3}$$

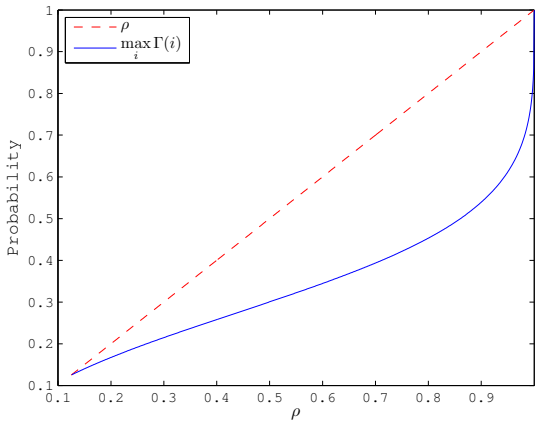
Assuming that the privacy goal is to ensure that the probability of identifying any individual in the database is no greater than $\frac{1}{3}$ (i.e., $\rho = \frac{1}{3}$),

$$\lambda = \frac{\frac{8}{3}}{\ln \frac{7 \cdot \frac{1}{3}}{1 - \frac{1}{3}}} = \frac{\frac{8}{3}}{\ln \frac{7}{2}} = \frac{8}{3 \ln \frac{7}{2}}$$

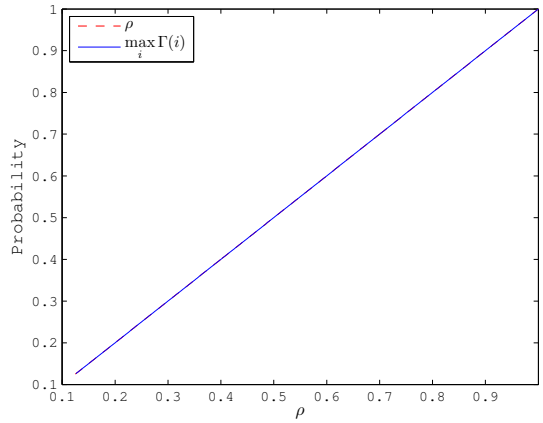
Assume the response to the adversary $R = 2$. This is the worst case, as it maximizes the distance between the expectation that $\omega_1 = D$ and other possible worlds. Given $\lambda = \frac{8}{3 \ln \frac{7}{2}}$ and $R = 2$, the identifiability risk for $I(2)$ is

$$\Gamma(2) = \frac{\exp\left(-\frac{|R - f(\omega_1)|}{\lambda}\right)}{\sum_{k=1}^{8} \exp\left(-\frac{|R - f(\omega_k)|}{\lambda}\right)} = 0.2294 < \frac{1}{3}.$$

The actual identifiability risk is less than the threshold $\rho$, thus it is possible to choose noise from a tighter distribution while still satisfying the privacy constraint. Figure 1(a) shows the change of identifiability risk by varying $\rho$. However, the bound is tight for some function $f$. Consider the case where the query function is median. Notice that the median of every possible world is 3 except $\omega_1 = D$. In this case, the sensitive range is $\mathcal{S}(f_{median}) = |3 - 2| = 1$ and $|f(D) - f(\omega_k)| = \mathcal{S}(f)$ for $2 \leq k \leq 8$. Therefore, Equation (15) holds with equality. In other words, the amount of noise computed by the upper bound is the amount actually required. Figure 1(b) shows the tightness of the bound.




(a) $f = mean$




(b) $f = median$

**Figure 1: Change of $\Gamma(i)$ by varying $\rho$**

THEOREM 2. *The upper bound on the identifiability risk, $\Gamma(i)$, is tight, that is there exists a case where the bound holds with equality.*

PROOF. The conditions under which Equation (8) and (9) hold with equality are

1. $R \geq f(D) \geq f(\omega_i) \vee R \leq f(D) \leq f(\omega_i)$ and
2. $\forall \omega, \omega \neq D, |f(D) - f(\omega)| = \mathcal{S}(f)$

respectively. Consider the case where the size of database to publish is one less than that of universe (i.e., $|D| = |U| - 1$). In this case, there exists only two possible worlds, namely $\omega_1$ and $\omega_2$. Without loss of generality, assume $f(\omega_1) < f(\omega_2)$. If $\omega_1 = D$, $Pr[R \leq f(D)] = \frac{1}{2}$. If $\omega_2 = D$, $Pr[R \geq f(D)] = \frac{1}{2}$. Clearly there always exists an $R$ that satisfies the first condition. According to the definition of sensitive range, $\mathcal{S}(f) = |f(\omega_1) - f(\omega_2)|$. Hence, the second condition is always satisfied. Therefore, for any function $f$, there exists a case that satisfies both conditions at the same time. □

## 5. EXAMPLE

We now look at practical applicability of differential identifiability. Due to space constraints, we only show the case

**Table 3: Description of Adult database**

| Attribute | Max. | Min. |
|---|---|---|
| age(AG) | 90 | 17 |
| education-num(EN) | 16 | 1 |
| capital-gain(CG) | 99999 | 0 |
| capital-loss(CL) | 4356 | 0 |
| hours-per-week(HW) | 99 | 1 |

where $f$ is a simple aggregate query. This mechanism can be applied to more complex queries (such as a data mining model) in the same manner as the Laplace noise mechanism for differential privacy.

We use the Adult Database from the UCI Machine Learning Repository [9], comprised of 48,842 individuals from the 1994 U.S. Census, as our example database. This database contains 9 categorical and 5 numerical attributes. Only the numerical attributes (shown in Table 3) are used in this example. Since the example is census data, we assume the universe is all US residents. Assume we wish to release the $\rho$-differentially identifiable mean of hours-per-week. Let $DB$ denote the example database, $D = \Pi_{hours-per-week}(DB)$ ($\Pi$ is relational projection), and $D'$ be the database obtained by removing an element from $D$. We assume that the maximum and minimum hours-per-week in the universe are 99 and 1, respectively. This is a reasonable assumption, since Census data typically uses top and bottom-coding to prevent rare values from being identifiable (clearly, [0, 168] could also be used as bounds.)

To determine the noise distribution, we must calculate $\mathcal{S}(f)$. Assume the adversary knows hours-per-week of every individual except one (i.e., the adversary has $D'$). The possible worlds the adversary would generate are $\omega_1 = D' \cup \{1\}, \omega_2 = D' \cup \{2\}, \cdots, \omega_{99} = D' \cup \{99\}$. Thus, the sensitive range $\mathcal{S}(f) = |f(\omega_{99}) - f(\omega_1)| = \frac{98}{48842} \approx 0.0020$. For $\rho = \frac{1}{10}$, this gives

$$\lambda = \frac{\mathcal{S}(f)}{\ln \frac{(m-1)\rho}{1-\rho}} = \frac{\frac{98}{48842}}{\ln \frac{98 \times 0.1}{1-0.1}} = \frac{49}{24421 \cdot \ln \frac{98}{9}} \approx 8.4032 \times 10^{-4}$$

Table 4 shows the parameters and three responses for both mean and median of each of the three attributes for $\rho = 0.1$ and 0.001. (This is assuming the adversary sees only a single one of these responses, returning all would require sharing the "privacy budget" as with differential privacy.) Notice that when $\rho = 0.001$, the amount of noise required to enforce $\Gamma(i) \leq \rho$ for the attributes age, education-num and hours-per-week becomes infinite; the released statistics are essentially random noise and do not give any information about individuals in the database. This is because a random guess that $D$ contains an individual with (say) 12 years of education is almost certainly correct.

To demonstrate the reliability of results from differential identifiability, Figure 2 shows the results of 1000 queries for mean as $\rho$ is varied. This illustrates how often a querier would be seriously misled by the differentially identifiable result. The vertical axis gives the noise ratio:

$$\text{Noise ratio (NR)} = \frac{R - f(D)}{U_{range}}$$

where $R$ is a response and $U_{range} (= U_{max} - U_{min})$ is the range of the domain. Note that the scale is $\times 10^{-3}$, so except for the rightmost trial on each plot, all the responses are close to the true answer. The band and cross near the middle represent the median and mean of responses, respectively. The boxplot is omitted when the magnitude of noise added to the response is typically greater than any possible value in the domain (i.e., $NR > 1$). From the figure, it is clear that the smaller the value of $\rho$ (i.e., the higher the desired privacy), the more noise is required; responses become less useful. When the value of $\rho$ is close to the probability of a random guess being correct, the mean and median of responses are not in range even with 1000 trials. Note that for everything except years of education, answers for $\rho = 1.5\%$ are highly likely to be close to the true answer.

Figure 3 shows the same information for differential privacy as the value of $\epsilon$ varies. (Only Capital-gain and Capital-loss are shown due to space constraints; the other figures are similar.) Note the similarity of the figures; it is clear that $\epsilon$-differential privacy and $\rho$-differential identifiability are comparable in terms of privacy. However, setting $\epsilon$ to achieve this privacy is not an easy problem. While the plots appear similar, the scale of $\epsilon$ is very different across the two plots. Interpretation of the semantics of epsilon with respect to re-identification is very difficult.

The effect of database size on the noise needed is shown in Figure 4. For this experiment, two databases of different sizes (1000 and 10,000) are constructed by randomly sampling from the original dataset. The red (left) column of each pair shows the noise ratio for the database of size 1000; the right (blue) is for the database of size 10,000. (The $\rho$ values for each pair are the same.) We see that, for $f = mean$, the noise needed shrinks with the size of the dataset. However, for $f = sum$, the noise is independent of dataset size. This is because the change in a sum based on one individual is the same regardless of the number of individuals, whereas for mean the impact shrinks with the number of individuals. Thus for mean it becomes harder to identify an individual from the released statistics as the database grows.

# 6. CONCLUSION

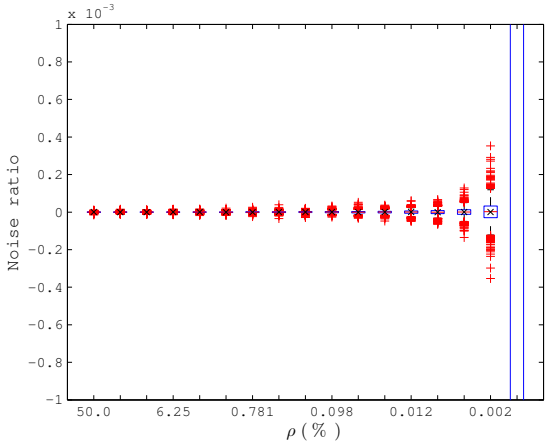
Differential identifiability (as with differential privacy) can often be satisfied with little impact on the resulting model. Data mining models should not be too dependent on a single individual; this would suggest that the model would not generalize well to unseen data. In addition to the general Laplace noise addition method of [7], several specialized techniques have already been developed for differentially private data mining [2,10,16,23]. We have shown that the general Laplace noise addition approach can be used to satisfy differential identifiability; we expect that analogous methods to other differentially private mechanisms can be developed to support differential identifiability as well.

# 7. REFERENCES

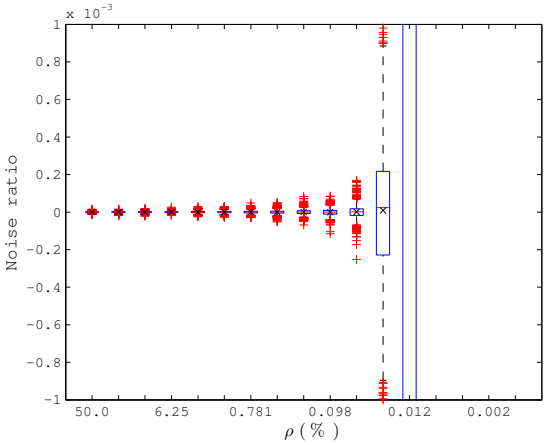

[1] C. C. Aggarwal and P. S. Yu, Eds., *Privacy-Preserving Data Mining: Models and Algorithms*, ser. Advances in Database Systems. Springer, 2008, vol. 34.

[2] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'10)*. New York, NY, USA: ACM, 2010, pp. 503–512. http://doi.acm.org/10.1145/1835804.1835869

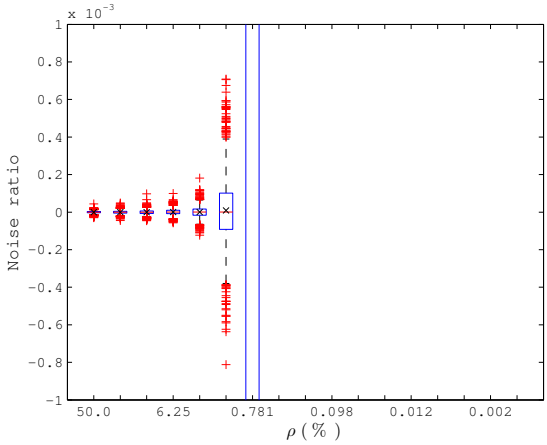
[3] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proceeding of the 22nd Annual*



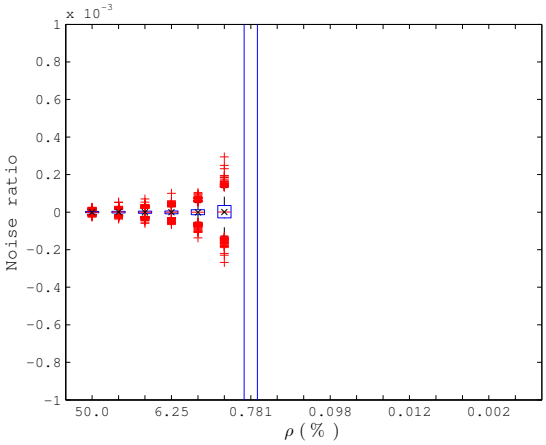

(a) Capital-gain



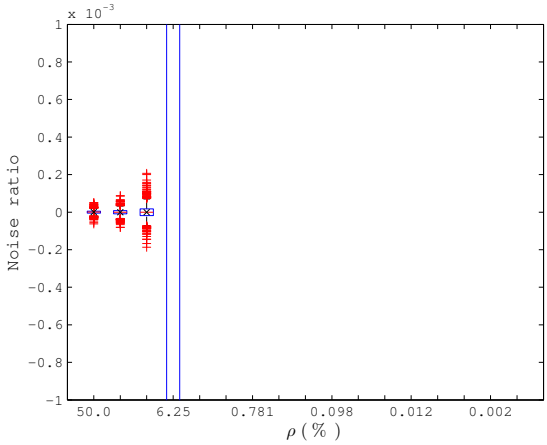

(b) Capital-loss



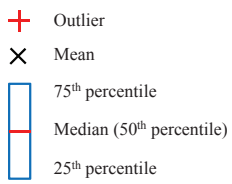

(c) Age




(d) Hours-per-week




(e) Education Number

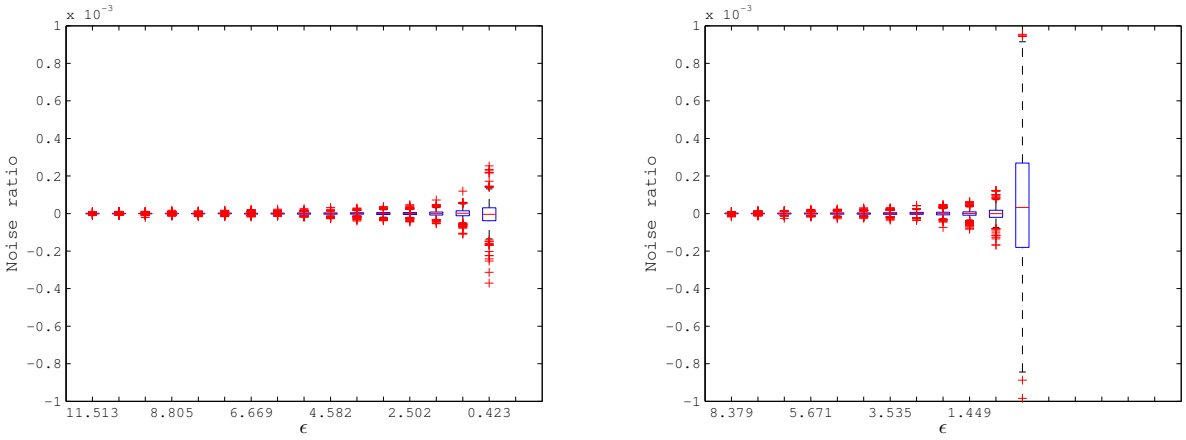



(f) Legend

**Figure 2: Noise Ratio**

**Table 4: Example of Three Query Responses for Mean and Standard Deviation**

| $f$ | Attr. | $\mathcal{S}(f)$ | $f(D)$ | $\rho = 0.1$ | | | | $\rho = 0.001$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $\lambda$ | $R_1$ | $R_2$ | $R_3$ | $\lambda$ | $R_1$ | $R_2$ | $R_3$ |
| Mean | AG | 0.0015 | 38.6435 | 0.0007 | 38.6431 | 38.6438 | 38.6438 | $\infty$ | 6410.178 | 36984.106 | -20388.622 |
| | EN | 0.0003 | 10.0780 | 0.0006 | 10.0784 | 10.0776 | 10.0780 | $\infty$ | 276944.55 | 6481.767 | -43022.968 |
| | CG | 2.0474 | 1079.067 | 0.2198 | 1079.027 | 1079.320 | 1079.308 | 0.4445 | 1080.4969 | 1078.2953 | 1079.1254 |
| | CL | 0.0892 | 87.5023 | 0.0144 | 87.5077 | 87.5179 | 87.5050 | 0.0606 | 87.4522 | 87.5130 | 87.4442 |
| | HW | 0.0020 | 40.4223 | 0.0008 | 40.4207 | 40.4204 | 40.4224 | $\infty$ | 3717.595 | 218292.256 | 32256.161 |
| Std. Dev. | AG | 0.0019 | 13.7105 | 0.0009 | 13.7114 | 13.7120 | 13.7093 | $\infty$ | 66620 | -24260 | 37263.7 |
| | EN | 0.0251 | 2.5709 | 0.0491 | 2.5970 | 2.4357 | 2.5419 | $\infty$ | -594100 | 445487 | 371312 |
| | CG | 26.8833 | 7452.019 | 2.8858 | 7454.0 | 7450.1 | 7452.1 | 5.8364 | 7458.7 | 7448.1 | 7457.8 |
| | CL | 0.4623 | 403.0045 | 0.0748 | 402.9782 | 403.0856 | 403.0185 | 0.3139 | 402.4125 | 402.3056 | 403.0245 |
| | HW | 0.0056 | 12.3914 | 0.0023 | 12.3937 | 12.3952 | 12.3882 | $\infty$ | -115431 | 283352 | 245743 |




(a) Capital-gain

(b) Capital-loss

**Figure 3: Noise Ratio for Differential Privacy**


*Conference on Neural Information Processing Systems (NIPS)*, 2008, pp. 289–296.

[4] G. Cormode, "Personal privacy vs population privacy: learning to attack anonymization," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'11)*. New York, NY, USA: ACM, 2011, pp. 1253–1261. http://doi.acm.org/10.1145/2020408.2020598

[5] B. Ding, M. Winslett, J. Han, and Z. Li, "Differentially private data cubes: optimizing noise sources and consistency," in *Proceedings of the 2011 international conference on Management of data (SIGMOD'11)*. New York, NY, USA: ACM, 2011, pp. 217–228. http://doi.acm.org/10.1145/1989323.1989347

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. of the 3rd Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[7] C. Dwork, "Differential privacy," in *33rd International Colloquium on Automata, Languages and Programming (ICALP 2006)*, Venice, Italy, Jul. 9-16 2006, pp. 1–12. http://dx.doi.org/10.1007/11787006_1

[8] "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Communities*, vol. No I., no. 281, pp. 31–50, Oct. 24 1995. http://ec.europa.eu/justice_home/fsj/privacy/law/index_en.htm

[9] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. http://archive.ics.uci.edu/ml

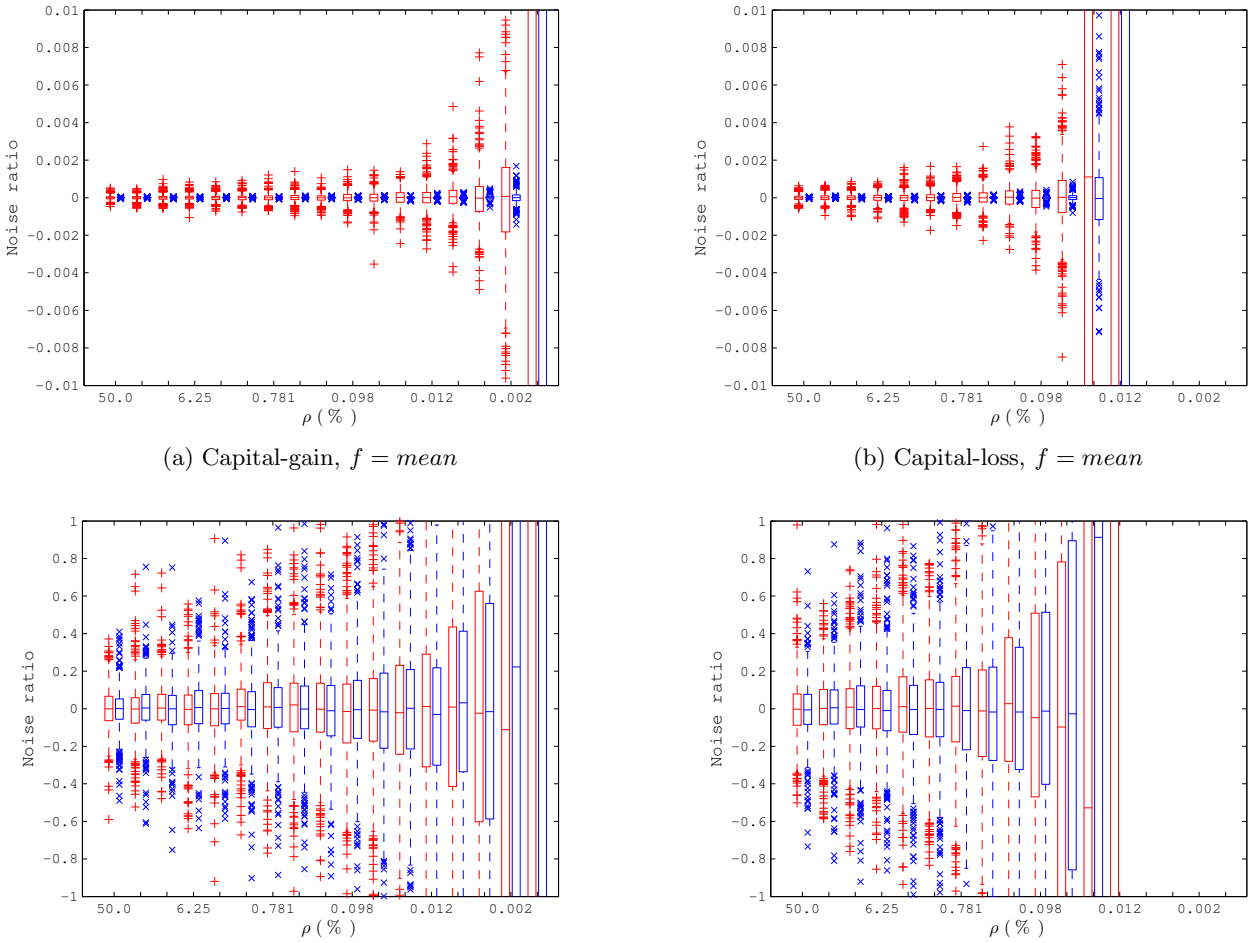
[10] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'11)*. New York, NY, USA: ACM, 2010, pp. 493–502. http://doi.acm.org/10.1145/1835804.1835868

[11] "Standard for privacy of individually identifiable health information," *Federal Register*, vol. 67, no. 157, pp. 53 181–53 273, Aug. 14 2002. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.h%tml

[12] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 Intl. Conf. on Management of data*. 2011, pp. 193–204.

[13] J. Lee and C. Clifton, "How much is enough? choosing $\epsilon$ for differential privacy," in *Information Security*, ser. Lecture Notes in Computer Science, X. Lai, J. Zhou, and H. Li, Eds. Springer Berlin / Heidelberg, 2011, vol. 7001, pp. 325–340.

[14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond




(a) Capital-gain, $f = mean$

(b) Capital-loss, $f = mean$

(c) Capital-gain, $f = sum$

(d) Capital-loss, $f = sum$

**Figure 4: Noise Ratio with small (1000) and large (10,000) item databases**


k-anonymity," *ACM Trans. on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.

[15] F. McSherry, "Privacy integrated queries: an extensible platform for privacy-preserving data analysis," *Commun. ACM*, vol. 53, pp. 89–97, Sep. 2010. http://doi.acm.org/10.1145/1810891.1810916

[16] F. McSherry and I. Mironov, "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, Jun. 28-Jul. 1 2009.

[17] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'11)*. ACM, 2011, pp. 493–501. http://doi.acm.org/10.1145/2020408.2020487

[18] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. 2007, pp. 665–676. http://doi.acm.org/10.1145/1247480.1247554

[19] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. on Knowl. and Data Eng.*, vol. 13, pp. 1010–1027, Nov. 2001. http://dx.doi.org/10.1109/69.971193

[20] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, pp. 557–570, Oct. 2002. http://dl.acm.org/citation.cfm?id=774544.774552

[21] J. Vaidya, C. Clifton, and M. Zhu, *Privacy Preserving Data Mining*, ser. Advances in Information Security. Springer, 2006, vol. 19. http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-0-387-25886-7

[22] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," *IEEE Trans. on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1200–1214, Aug. 2011.

[23] N. Zhang, M. Li, and W. Lou, "Distributed data mining with differential privacy," in *2011 IEEE International Conference on Communications (ICC)*, Jun. 2011, pp. 1–5.