Composition Attacks and Auxiliary Information in Data Privacy

Srivatsava Ranjit Ganta
Pennsylvania State University
University Park, PA 16802

ranjit@cse.psu.edu

Shiva Prasad
Kasiviswanathan
Pennsylvania State University
University Park, PA 16802

kasivisw@cse.psu.edu

Adam Smith*
Pennsylvania State University
University Park, PA 16802

asmith@cse.psu.edu

ABSTRACT

Privacy is an increasingly important aspect of data publishing. Reasoning about privacy, however, is fraught with pitfalls. One of the most significant is the auxiliary information (also called external knowledge, background knowledge, or side information) that an adversary gleans from other channels such as the web, public records, or domain knowledge. This paper explores how one can reason about privacy in the face of rich, realistic sources of auxiliary information. Specifically, we investigate the effectiveness of current anonymization schemes in preserving privacy when multiple organizations *independently* release anonymized data about overlapping populations.

1. We investigate *composition attacks*, in which an adversary uses independently anonymized releases to breach privacy. We explain why recently proposed models of limited auxiliary information fail to capture composition attacks. Our experiments demonstrate that even a simple instance of a composition attack can breach privacy in practice, for a large class of currently proposed techniques. The class includes k-anonymity and several recent variants.

2. On a more positive note, certain randomization-based notions of privacy (such as differential privacy) provably resist composition attacks and, in fact, the use of arbitrary side information. This resistance enables "stand-alone" design of anonymization schemes, without the need for explicitly keeping track of other releases.

We provide a precise formulation of this property, and prove that an important class of relaxations of differential privacy also satisfy the property. This significantly enlarges the class of protocols known to enable modular design.

Categories and Subject Descriptors

H.2.0 [**Database Management**]: General

General Terms

Management, Security, Theory

Keywords

Privacy, Anonymization, Adversarial Attacks

1. INTRODUCTION

Privacy is an increasingly important aspect of data publishing. The potential social benefits of analyzing large collections of personal information (census data, medical records, social networks) are significant. At the same time, the release of information from such repositories can be devastating to the privacy of individuals or organizations [5]. The challenge is therefore to discover and release the global characteristics of these databases without compromising the privacy of the individuals whose data they contain.

Reasoning about privacy, however, is fraught with pitfalls. One of the most significant difficulties is the auxiliary information (also called external knowledge, background knowledge, or side information) that an adversary gleans from other channels such as the web or public records. For example, simply removing obviously identifying information such as names and address does not suffice to protect privacy since the remaining information (such as zip code, gender and date of birth [33]) may still identify a person uniquely when combined with auxiliary information (such as voter registration records). Schemes that resist such linkage have been the focus of extensive investigation, starting with work on publishing contingency tables [1], and more recently, in a line of techniques based on "k-anonymity" [33].

This paper explores how one can reason about privacy in the face of rich, realistic sources of auxiliary information. This follows lines of work in both the data mining [28, 29, 10] and cryptography [11, 13] communities that have sought principled ways to incorporate *unknown* auxiliary information into anonymization schemes.

Specifically, we investigate the effectiveness of current anonymization schemes in preserving privacy when multiple organizations *independently* release anonymized data about overlapping populations. We show new attacks on some schemes and also deepen the current understanding of schemes known to resist such attacks. Our results and their relation to previous work are discussed below.

Schemes that retain privacy guarantees in the presence of independent releases are said to *compose securely*. The terminology, borrowed from cryptography (which borrowed, in turn, from software engineering), stems from the fact that schemes which compose securely can be designed in a stand-alone fashion without explicitly taking other releases into account. Thus, understanding independent releases is essential for enabling modular design. In fact, one would like schemes that compose securely not only with

independent instances of themselves, but with *arbitrary external knowledge*. We discuss both types of composition in this paper.

The dual problem to designing schemes with good composition properties is the design of attacks that exploit independent releases. We call these *composition attacks*. A simple example of such an attack, in which two hospitals with overlapping patient populations publish anonymized medical data, is presented below. Composition attacks highlight a realistic and important class of vulnerabilities. As privacy preserving data publishing becomes more commonly deployed, it is increasingly difficult to keep track of all the organizations that publish anonymized summaries involving a given individual or entity. Schemes that are vulnerable to composition attacks will be consequently difficult to use safely.

1.1 Contributions

Our contributions are summarized briefly in the abstract, above, and discussed in more detail in the following subsections.

1.1.1 Composition Attacks on Partition-based Schemes

We introduce composition attacks and study their effect on a popular class of partitioning-based anonymization schemes. Very roughly, computer scientists have worked on two broad classes of anonymization techniques. *Randomization-based* schemes introduce uncertainty either by randomly perturbing the raw data (a technique called *input perturbation*, *randomized response*, e.g., [37, 2, 17]), or *post-randomization*, e.g., [35]), or by injecting randomness into the algorithm used to analyze the data (e.g., [6, 30]). *Partition-based schemes* cluster the individuals in the database into disjoint groups satisfying certain criteria (for example, in k-anonymity [33], each group must have size at least k). For each group, certain exact statistics are calculated and published. Partition-based schemes include k-anonymity [33] as well as several recent variants, e.g., [28, 39, 40, 25, 41, 29, 10, 23].

Because they release exact information, partition-based schemes seem especially vulnerable to composition attacks. In the first part of this paper we study a simple instance of a composition attack called an *intersection attack*. We observe that the specific properties of current anonymization schemes make this attack possible, and we evaluate its success empirically.

Example. Suppose two hospitals H_1 and H_2 in the same city release anonymized patient-discharge information. Because they are in the same city, some patients may visit both hospitals with similar ailments. Tables 1(a) and 1(b) represent (hypothetical) independent k-anonymizations of the discharge data from H_1 and H_2 using $k = 4$ and $k = 6$, respectively. The sensitive attribute here is the patient's medical condition. It is left untouched. The other attributes, deemed non-sensitive, are generalized (that is, replaced with aggregate values), so that within each group of rows, the vectors of non-sensitive attributes are identical. If Alice's employer knows that she is 28 years old, lives in zip code 13012 and recently visited both hospitals, then he can attempt to locate her in both anonymized tables. Alice matches four potential records in H_1's data, and six potential records in H_2's. However, the only disease that appears in both matching lists is AIDS, and so Alice's employer learns the reason for her visit.

Intersection Attacks. The above example relies on two properties of the partition-based anonymization schemes: **(i) Exact sensitive value disclosure:** the "sensitive" value corresponding to each member of the group is published exactly; and **(ii) Locatability:** given any individual's non-sensitive values (non-sensitive values are exactly those that are assumed to be obtainable from other, public information sources) one can locate the group in which individual has been put in. Based on these properties, an adversary can

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

(b)

Table 1: A simple example of a composition attack. Tables (a) and (b) are 4-anonymous (respectively, 6-anonymous) patient data from two hypothetical hospitals. If an Alice's employer knows that she is 28, lives in zip code 13012 and visits both hospitals, he learns that she has AIDS.

narrow down the set of possible sensitive values for an individual by intersecting the sets of sensitive values present in his/her groups from multiple anonymized releases.

Properties (i) and (ii) turn out to be widespread. The exact disclosure of sensitive value lists is a design feature common to all the schemes based on k-anonymity: preserving the exact distribution of sensitive values is important, and so no recoding is usually applied. Locatability is less universal, since it depends on the exact choice of partitioning algorithm (used to form groups) and the recoding applied to the non-sensitive attributes. However, some schemes always satisfy locatability by virtue of their structure (e.g., schemes that recursively partition the data set along the lines of a hierarchy that is subsequently used for generalization [23, 24], or schemes that release the exact set of non-sensitive attribute vectors for each group [39]). For other schemes, locatability is not perfect but our experiments suggest that using simple heuristics one can locate a individual's group with high probability.

Even with these properties, it is difficult to come up with a theoretical model for intersection attacks because the partitioning techniques generally create dependencies that are hard to model analytically. (If the sensitive values of the members of a group could be assumed to be statistically independent of their non-sensitive attribute values, then an analytic solution would be tractable; however, the assumption is does not fit the data we considered).

Experimental Results. Instead, we evaluated the success of intersection attacks empirically. We ran the intersection attack on two popular census databases anonymized using partition-based schemes. We evaluated the severity of the attack by measuring the number of individuals who had their sensitive value revealed. Our experimental results confirm that partitioning-based anonymization schemes including k-anonymity and its recent variants, ℓ-diversity and t-closeness, are indeed vulnerable to intersection attacks. Section 3 elaborates our methodology and results.

Related Work on Modeling Background Knowledge. It is important to point out that the partition-based schemes in the literature were not designed to be used in contexts where independent releases are available. Thus, we do not view our results as pointing out a flaw in these schemes, but rather as directing the community's attention to an important direction for future work.

It is equally important to highlight the progress that has already been made on modeling sophisticated background knowledge in partition-based schemes. One line has focused on taking into account other, *known* releases, such as previous publications by the same organization (called "sequential" or "incremental" releases, [36, 8, 41, 19, 32]) and multiple views of the same data set [42]. In this paper, we deal with the case when the publisher is not aware of other anonymized releases. Another line has considered incorporating knowledge of the partitioning algorithm used to group individuals [38]. Most relevant to this paper are works that have sought to model *unknown* background knowledge. Martin *et al.* [29] and Chen *et al.* [10] provide complexity measures for an adversary's side information (roughly, they measure the size of the smallest formula within a CNF-like class that can encode the side information). Both works design schemes that provably resist attacks based on side information whose complexity is below a given threshold.

Independent releases (and hence composition attacks) fall outside the models proposed by these works. The sequential release models do not fit because they assume that other releases are known to the anonymization algorithm. The complexity-based measures do not fit because independent releases appear to have complexity that is linear in the size of the data set.

1.1.2 Composing Randomization-based Schemes

Composition attacks appear to be difficult to reason about, and it is not initially clear whether it is possible at all to design schemes that resist such attacks. Even *defining* composition properties precisely is tricky in the presence of malicious behavior (for example, see [26] for a recent survey about composability of cryptographic protocols). Nevertheless, a significant family of anonymization definitions do provide guarantees against composition attacks, namely schemes that satisfy *differential privacy* [15]. Recent work has greatly expanded the applicability of differential privacy and its relaxations, both in the theoretical [16, 6, 15, 4, 30, 7] and applied [18, 3, 27] literature. However, certain recently developed techniques such as sampling [9], instance-based noise addition [31] and data synthesis [27] appear to require relaxations of the definition.

Both the strict and relaxed variants of differential privacy compose well (see Lemma 12). Less trivially, however, one can prove that strictly differentially-private algorithms also provide meaningful privacy in the presence of *arbitrary side information* (Dwork and McSherry, [13]). In particular, these schemes compose well even with completely different anonymization schemes.

It is natural to ask if there are weaker definitions which provide similar guarantees. Certainly not all of them do: one natural relaxation of differential privacy, which replaces the multiplicative distance used in differential privacy with total variation distance, fails completely to protect privacy (see example 2 in [15]).

In this paper, we prove that two important relaxations of differential privacy do, indeed, resist arbitrary side information. First, we provide a Bayesian formulation of differential privacy which makes its resistance to arbitrary side information explicit. Second, we prove that the relaxed definitions of [14, 27] still imply the Bayesian formulation. The proof is non-trivial, and relies on the "continuity" of Bayes' rule with respect to certain distance measures on probability distributions. Our result means that the recent techniques mentioned above [14, 9, 31, 27] can be used mod-

ularly with the same sort of assurances as in the case of strictly differentially-private algorithms.

In this paper, we won't be dealing with the utility aspects of differential privacy. Interested readers are referred to recent papers [16, 6, 15, 4, 30, 31, 7, 27] that explore the relation between utility and differential privacy.

2. PARTITION-BASED SCHEMES

Let D be a multiset of tuples where each tuple corresponds to an individual in the database. Let R be an anonymized version of D. From this point on, we use the terms tuple and individual interchangeably, unless the context leads to ambiguity. Let $A = A_1, A_2, \ldots, A_r$ be a collection of attributes and t be a tuple in R; we use the notation $t[A]$ to denote $(t[A_1], \ldots, t[A_r])$ where each $t[A_i]$ denotes the value of attribute A_i in table R for t.

In partitioning-based anonymization approaches, there exists a division of data attributes into two classes, *sensitive attributes* and *non-sensitive attributes*. A sensitive attribute is one whose value and an individual's association with that value should not be disclosed. All attributes other than the sensitive attributes are non-sensitive attributes.

DEFINITION 1 (QUASI-IDENTIFIER). *A set of non-sensitive attributes $\{Q_1, \ldots, Q_r\}$ is called a **quasi-identifier** if there is at least one individual in the original sensitive database D who can be uniquely identified by linking these attributes with auxiliary data.*

Previous work in this line typically assumed that all the attributes in the database other than the sensitive attribute form the quasi-identifier.

DEFINITION 2 (EQUIVALENCE CLASS). *An **equivalence class** for a table R with respect to attributes in A is the set of all tuples $t_1, t_2, \ldots, t_i \in R$ for which the projection of each tuple onto attributes in A is the same, i.e., $t_1[A] = t_2[A] \ldots = t_i[A]$.*

Partition-based schemes cluster individuals into groups, and then recode (*i.e.,* generalize or change) the non-sensitive values so that each group forms an equivalence class with respect to the quasi-identifiers. Sensitive values are not recoded. Different criteria are used to decide how, exactly, the groups should be structured. The most common rule is k-anonymity, which requires that each equivalence class contain at least k individuals.

DEFINITION 3 (k-ANONYMITY [33]). *A release R is k-**anonymous** if for every tuple $t \in R$, there exist at least $k-1$ other tuples $t_1, t_2, \ldots, t_{k-1} \in R$ such that $t[A] = t_1[A] = \ldots = t_{k-1}[A]$ for every collection A of attributes in quasi-identifier.*

In our experiments we also consider two extensions to k-anonymity.

DEFINITION 4 (ENTROPY ℓ-DIVERSITY [28]). *For an equivalence class E, let S denote the domain of the sensitive attributes, and $p(E, s)$ is the fraction of records in E that have sensitive value s, then E is ℓ-**diverse** if:*

$$-\sum_{s \in S} p(E, s) \log(p(E, s)) \geq \log l \,.$$

A table is ℓ-diverse if all its equivalence classes are ℓ-diverse.

DEFINITION 5 (t-CLOSENESS [25]). *An equivalence class E is t-**close** if the distance between the distribution of a sensitive attribute in this class and distribution of the attribute in the whole table is no more than a threshold t. A table is t-close if all its equivalence classes are t-close.*

Locatability. As mentioned in the introduction, many anonymization algorithms satisfy *locatability*, that is, they output tables in which one can locate an individual's group based only on his or her *non-sensitive* values.

DEFINITION 6 (LOCATABILITY). *Let Q be the set of quasi-identifier values of an individual in the original database D. Given the k-anonymized release R of D, the locatability property allows an adversary to identify the set of tuples $\{t_1, \ldots, t_K\}$ in R (where $K \geq k$) that correspond to Q.*

Locatability does not necessarily hold for all partition-based schemes, since it depends on the exact choice of partitioning algorithm (used to form groups) and the recoding applied to the non-sensitive attributes. However it is widespread. Some schemes *always* satisfy locatability by virtue of their structure (*e.g.*, schemes that recursively partition the data set along the lines of a hierarchy always provide locatability if the attributes are then generalized using the same hierarchy, or if (min,max) summaries are used [23, 24]). For other schemes, locatability is not perfect but our experiments suggest that using simple heuristics can locate a person's group with good probability. For example, microaggregation [12] clusters individuals based on Euclidean distance. The vectors of non-sensitive values in each group are replaced by the centroid (i.e., average) of the vectors. The simplest heuristic for locating an individual's group is to choose the group with the closest centroid vector. In experiments on census data, this correctly located approximately 70% of individuals. In our attacks, we always assume locatability. This assumption was also made in previous studies [33, 29].

2.1 Intersection Attack

Armed with these basic definitions, we now proceed to formalize the intersection attack (Algorithm 1).

Algorithm 1 Intersection attack

1: $R_1, \ldots, R_n \leftarrow n$ independent anonymized releases
2: $P \leftarrow$ set of overlapping population
3: **for** each individual i in P **do**
4:     **for** $j = 1$ to n **do**
5:         $e_{ij} \leftarrow$ Get_equivalence_class(R_j, i)
6:         $s_{ij} \leftarrow$ Sensitive_value_set(e_{ij})
7:     **end for**
8:     $S_i \leftarrow s_{i1} \cap s_{i2} \cap \ldots \cap s_{in}$
9: **end for**
10: **return** $S_1, \ldots, S_{|P|}$

Let $R_1, \ldots, R_n$ be n independent anonymized releases with minimum partition-sizes of $k_1, \ldots, k_n$, respectively. Let P be the overlapping population occurring in all the releases. The function Get_equivalence_class returns the equivalence class into which an individual falls in a given anonymized release. The function Sensitive_value_set returns the set of (distinct) sensitive values for the members in a given equivalence class.

DEFINITION 7 (ANONYMITY). *For each individual i in P, the anonymity factor promised by each release R_j is equal to the corresponding minimum partition-size k_j.*

However, as pointed out in [28], the actual anonymity offered is less than this ideal value and is equal to number of distinct values in each equivalence class. We call this as the *effective anonymity*

DEFINITION 8 (EFFECTIVE ANONYMITY). *For an individual i in P, the effective anonymity offered by a release R_j is equal to*

the number of distinct sensitive values of the partition into which the individual falls into. Let e_{ij} be the equivalence class or partition into which i falls into with respect to the release R_j, and let s_{ij} denote the sensitive value set for e_{ij}. The effective anonymity for i with respect to the release R_j is: $EA_{ij} = |s_{ij}|$.

For each target individual i, EA_{ij} is the *effective prior anonymity* with respect to R_j (anonymity before the intersection attack). In the intersection attack, the list of possible sensitive values associated to the target is equal to intersection of all sensitive value sets s_{ij}, $j = 1, \ldots, n$. So the *effective posterior anonymity* ($\widehat{EA}_i$) for i is:

$$\widehat{EA}_i = |\{\cap s_{ij}\}|, j = 1, \ldots, n \, .$$

The difference between the effective prior anonymity and effective posterior anonymity quantifies the drop in effective anonymity.

$$Anon\_Drop_i = \min_{j=1,\ldots,n} \{EA_{ij}\} - \widehat{EA}_i \, .$$

The *vulnerable population* (VP) is the number of individuals (among the overlapping population) for whom the intersection attack leads to a positive drop in the effective anonymity.

$$VP = \{i \in P \, : \, Anon\_Drop_i > 0\} \, .$$

After performing the sensitive value set intersection, the adversary knows only a possible set of values that each individual's sensitive attribute can take. So, the adversary deduces that with equal probability (under the assumption that the adversary does not have any further auxiliary information) the individual's actual sensitive value is one of the values in the set $\{\cap s_{ij}\}, j = 1, \ldots, n$. So, the adversaries *confidence level* for an individual i can be defined as:

DEFINITION 9 (CONFIDENCE LEVEL C_i). *For each individual i, the confidence level C_i of the adversary in identifying the individual's true sensitive value through the intersection attack is defined as $C_i = \frac{1}{\widehat{EA}_i}$.*

Now, given some confidence level C, we denote by VP_C and PVP_C the set and the percentage of overlapping individuals for whom the adversary can deduce the sensitive attribute value with a confidence level of at least C.

$$VP_C = \{i \in P \, : \, C_i \geq C\} \, ,$$
$$PVP_C = \frac{|VP_C| \cdot 100}{|P|} \, .$$

3. EXPERIMENTAL RESULTS

In this section we describe our experimental study[1]. The primary goal is to quantify the severity of such an attack on existing schemes. Although the earlier works address problems with k-anonymization and adversarial background knowledge, to the best of our knowledge, none of these studies deal with attacks resulting from auxiliary independent releases. Furthermore, none of the studies so far have quantified the severity of such an attack.

3.1 Setup

We use three different partitioning-based anonymization techniques to demonstrate the intersection attack: k-anonymity, ℓ-diversity, and t-closeness. For k-anonymity, we use the Mondrian multidimensional approach proposed in [23] and the microaggregation

---

Attribute	Domain Size	Class
Age	74	Quasi ID
Work Class	7	Quasi ID
Education	16	Quasi ID
Marital Status	7	Quasi ID
Race	5	Quasi ID
Gender	2	Quasi ID
Native Country	41	Quasi ID
Occupation	14	Sensitive

Table 2: Description of the Adult census database.

Attribute	Domain Size	Class
Age	100	Quasi ID
Work Class	5	Quasi ID
Education	10	Quasi ID
Marital Status	6	Quasi ID
Race	7	Quasi ID
Sex	2	Quasi ID
Birth Place	113	Quasi ID
Occupation	247	Sensitive

Table 3: Description of the IPUMS census database.

technique proposed in [12]. For ℓ-diversity and t-closeness, we use the definitions of entropy ℓ-diversity and t-closeness proposed in [28] and [25], respectively.

We use two census-based databases from the UCI Machine Learning repository [34]. The first one is the Adult database that has been used extensively in the k-anonymity based studies. The database was prepared in a similar manner to previous studies [23, 28] (also explained in Table 2). The resulting database contained individual records corresponding to 30162 people. The second database is the IPUMS database that contains individual information from the 1997 census studies. We only use a subset of the attributes that are similar to the attributes present in the Adult database to maintain uniformity and to maintain quasi-identifiers. The IPUMS database contains individual records corresponding to a total of 70187 people. This data set was prepared as explained in Table 3.

From both Adult and IPUMS databases, we generate two overlapping subsets (Subset 1 and Subset 2) by randomly sampling individuals without replacement from the total population. We fixed the overlap size to $P = 5000$. For each of the databases, the two subsets are anonymized independently and the intersection attack is run on the anonymization results. All the experiments were run on a Pentium 4 system running Windows XP with 1GB RAM.

3.2 Severity of the Attack

Our first goal is to quantify the extent of damage possible through the intersection attack. For this, we consider two possible situations: (i) Perfect breach and (ii) Partial breach.

3.2.1 Perfect Breach

A perfect breach occurs when the adversary can deduce the exact sensitive value of an individual. In other words, a perfect breach is when the adversary has a confidence level of 100% about the individual's sensitive data. To estimate the probability of a perfect breach, we compute the percentage of overlapping population for whom the intersection attack leads to a final sensitive value set of size 1. Figure 1 ((a) and (b)) plots this result.

We consider three scenarios for anonymizing the two overlapping subsets: (i) Mondrian on both the data subsets, (ii) Microaggregation on both the data subsets, and (iii) Mondrian on the first subset and microaggregation on the second subset. (k_1, k_2) represents the pair of k values used to anonymize the first and the second

subset, respectively. In the experiments, we use the same k values for both the subsets ($k_1 = k_2$). Note that for simplicity, from now on we will be defining confidence level in terms of percentages.

In the case of Adult database we found that around 12% of the population is vulnerable to a perfect breach for $k_1 = k_2 = 5$. For the IPUMS database, this value is much more severe around 60%. As the degree of anonymization increases or in other words, as the value of k increases, the percentage of vulnerable population goes down. The reason for that is that as the value of k increases, the partition sizes in each subset increases. This leads to a larger intersection set and thus lesser probability of obtaining an intersection set of size 1.

3.2.2 Partial Breach

Our next experiment aims to compute a more practical quantification of the severity of the intersection attack. In most cases, to inflict a privacy breach, all that the adversary needs to do is to boil down the possible sensitive values to a *few* values which itself could reveal a lot of information. For example, for a hospital discharge database, by boiling down the sensitive values of the disease/diagnosis to a few values, say, "Flu", "Fever", or "Cold", it could be concluded that the individual is suffering from a viral infection. In this case, the adversary's confidence level is $1/3 = 33\%$. Figure 1 ((c) and (d)) plots the percentage of vulnerable population for whom the intersection attack leads to a partial breach for the Adult and IPUMS databases.

Here, we only use the first anonymization scenario described earlier in which both the overlapping subsets of the database are anonymized using Mondrian multidimensional technique. Observe that the severity of the attack increases alarmingly for slight relaxation on the required confidence level. For example, in the case of IPUMS database, around 95% of the population was vulnerable for a confidence level of 25% for $k_1 = k_2 = 5$. For the Adult database, although this value is not as alarming, more than 60% of the population was affected.

3.3 Drop in Anonymity

Our next goal is to measure the drop in anonymity occurring due to the intersection attack. To achieve this, we first take a closer look at the way these schemes work. As described in the earlier sections, the basic paradigm in partitioning-based anonymization schemes is to partition the data such that each partition size is at least k. The methodology behind partitioning and then summarizing varies from scheme to scheme. The minimum partition-size (k) is thus used as a measure of the anonymity offered by these solutions. However, the effective (or true) anonymity supported by these solutions is far less than the presumed anonymity k (refer to the discussion in Section 2.1).

Figure 2 plots the average partition sizes and the average effective anonymities for the overlapping population. Here again, we only consider the scenario where both the overlapping subsets are anonymized using Mondrian multidimensional technique. Observe that the effective anonymity is much less than the partition size for both the data subsets. Also, note that these techniques result in partition sizes that are much larger than the minimum required of k. For example, the average partition size observed in the IPUMS database for $k = 5$ is close to 40. To satisfy the k-anonymity definition, there is no need for any partition to be larger than $2k + 1$. The reasoning for this is straightforward as splitting the partition of size greater than $2k + 1$ into two we get partitions of size at least k. Additionally, splitting any partition of size $2k + 1$ or more only results in preserving more information. The culprit behind the larger average partition sizes is generalization-based on user-
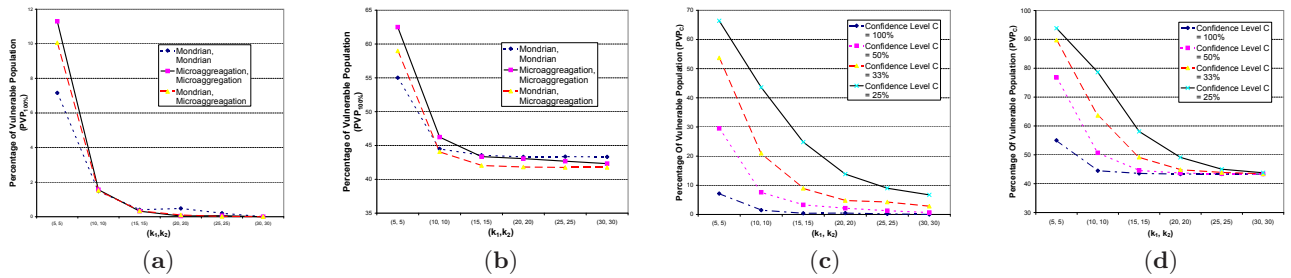
Figure 1: Severity of the intersection attack - perfect breach (a) Adult database and (b) IPUMS database. Severity of the intersection attack - partial breach (c) Adult database and (d) IPUMS database.

defined hierarchies. Since generalization-based partitioning cannot be controlled at finer levels, the resulting partition sizes tend to be much larger than the minimum required value.

For each individual in the overlapping population, the effective prior anonymity is equal to the effective anonymity. We define the average effective prior anonymity with respect to a release as effective prior anonymities averaged over the individuals in the overlapping population. Similarly, the average effective posterior anonymity is the effective posterior anonymities averaged over the individuals in the overlapping population. The difference between the average effective prior anonymity and average effective posterior anonymity gives the average drop in effective anonymity occurring due to the intersection attack. Figure 3 plots the average effective prior anonymities and the average effective posterior anonymities for the overlapping population. Observe that the average effective posterior anonymity is much less than the average effective prior anonymity for both subsets. Also note that we measure drop in anonymities by using effective anonymities instead of presumed anonymities. The situation only gets worse (drops get larger) when presumed anonymities are used.

3.4 ℓ-diversity and t-closeness

We now consider the ℓ-diversity and t-closeness extensions to the original k-anonymity definition. The goal again is to quantify the severity of the intersection attack by measuring the extent to which a partial breach occurs with varying levels of adversary confidence levels. Figure 4 plots the percentage of vulnerable population for whom the intersection attack leads to a partial breach for the IPUMS databases. Here, we anonymize both the subsets of the database with the same definition of privacy. We use the mondrian multidimensional k-anonymity with the additional constraints as defined by ℓ-diversity and t-closeness. Figure 4(a) plots the result for the ℓ-diversity using the same ℓ value for both the subsets ($\ell_1 = \ell_2$) and with $k = 10$. Figure 4(b) plots the same for t-closeness. Even though these extended definitions seem to perform better than the original k-anonymity definition, they still lead to considerable breach in case of an intersection attack. This result is fairly intuitive in the case of ℓ-diversity. Consider the definition of ℓ-diversity: the sensitive value set corresponding to each partition should be "well" (ℓ) diverse. However, there is no guarantee that the intersection of two well diverse sets leads to a well diverse set. t-closeness fares similarly. Also, both these definitions tend to force larger partition sizes, thus resulting in heavy information loss. Figure 4(c) plots the average partition sizes of the individuals corresponding to the overlapping population. It compares the partition sizes observed for k-anonymity, ℓ-diversity, and t-closeness. For the IPUMS database, with a value of $k = 10$, k-anonymity produces partitions with an average partition size of 45. While, for the

same value of $k = 10$, with a value of $l = 5$, the average partition size obtained was close to 450. The partition sizes for t-closeness get even worse, where a combination of $k = 10$ and $t = 0.4$ yield partitions of average size close to 1300.

4. DIFFERENTIAL PRIVACY

In this section we give a precise formulation of "resistance to arbitrary side information" and show that several relaxations of differential privacy imply it. The formulation follows the ideas originally due to Dwork and McSherry, stated implicitly in [13]. This is, to our knowledge, the first place such a formulation appears explicitly. The proof that relaxed definitions (and hence the schemes of [14, 31, 27]) satisfy the Bayesian formulation is new. These results are explained in a greater detail in a separate technical report [22]. In this paper we just reproduce the relevant parts from [22].

We represent databases as vectors in $\mathcal{D}^n$ for some domain $\mathcal{D}$ (for example, in the case of the relational databases above, $\mathcal{D}$ is the product of the attribute domains). There is no distinction between "sensitive" and "insensitive" information. Given a randomized algorithm $\mathcal{A}$, we let $\mathcal{A}(D)$ be the random variable (or, probability distribution on outputs) corresponding to input D.

DEFINITION 10 (DIFFERENTIAL PRIVACY [15]). *A randomized algorithm $\mathcal{A}$ is ϵ-**differentially private** if for all databases $D_1, D_2 \in \mathcal{D}^n$ that differ in one individual, and for all subsets S of outputs, $\Pr[\mathcal{A}(D_1) \in S] \leq e^\epsilon \Pr[\mathcal{A}(D_2) \in S]$.*

This definition states that changing a single individual's data in the database leads to a small change in the *distribution* on outputs. Unlike more standard measures of distance such as total variation (also called statistical difference) or Kullback-Leibler divergence, the metric here is multiplicative and so even very unlikely events must have approximately the same probability under the distributions $\mathcal{A}(D_1)$ and $\mathcal{A}(D_2)$. This condition was relaxed somewhat in other papers [11, 16, 6, 14, 9, 31, 27]. The schemes in all those papers, however, satisfy the following relaxation:

DEFINITION 11 ((ϵ, δ)-DIFFERENTIAL PRIVACY [14]). *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-**differentially private** if for all data-bases $D_1, D_2 \in \mathcal{D}^n$ that differ in one individual, and for all subsets S of outputs, $\Pr[\mathcal{A}(D_1) \in S] \leq e^\epsilon \Pr[\mathcal{A}(D_2) \in S] + \delta$.*

The relaxations used in [16, 6, 27] were in fact stronger (i.e., less relaxed) than Definition 11. One consequence of the results below is that all the definitions are equivalent up to polynomial changes in the parameters, and so given the space constraints we work only with the simplest notion of Definition Definition 11.[2]

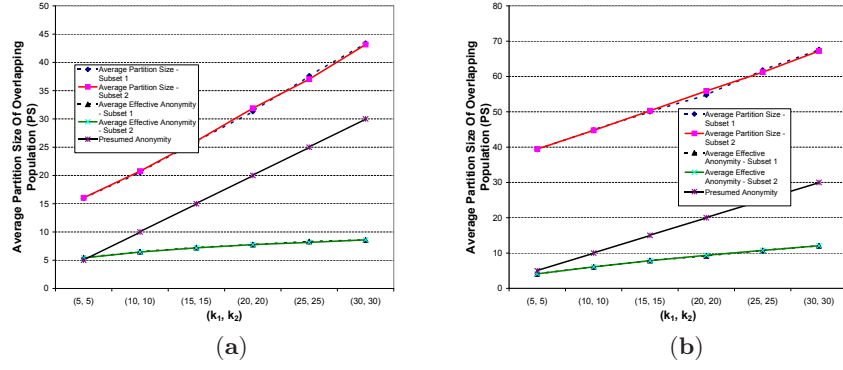[2]That said, some of the other relaxations, such as probabilistic dif-

Figure 2: Comparison of presumed anonymity, actual partition sizes, and effective anonymity **(a)** Adult database and **(b)** IPUMS database.
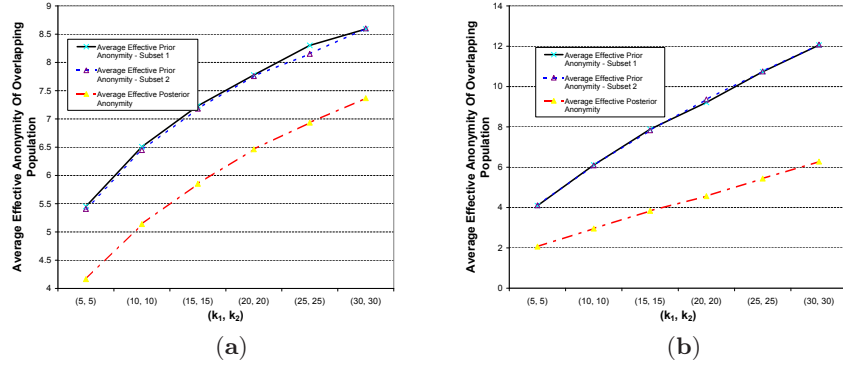

Figure 3: Average drop in effective anonymity due to the intersection attack **(a)** Adult database and **(b)** IPUMS database.
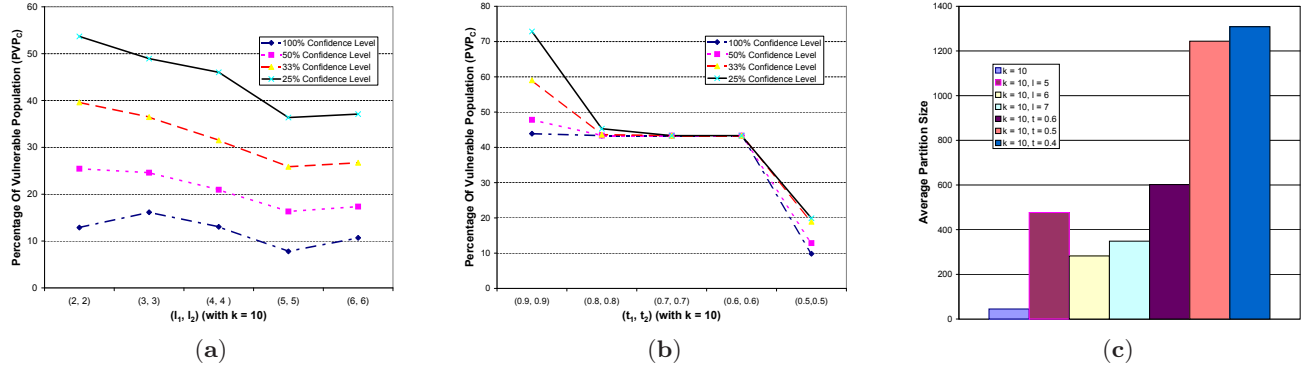

Figure 4: Severity of the intersection attack for the IPUMS database (a) ℓ-diversity and (b) t-closeness. Part (c) depicts the average partition sizes for ℓ-diversity and t-closeness for the IPUMS database

The following well-known lemma shows that (ϵ, δ)-differential privacy composes well (cf. [14, 31, 30]).

LEMMA 12 (COMPOSITION). *If a randomized algorithm $\mathcal{A}$ runs k algorithms $\mathcal{A}_1, ..., \mathcal{A}_k$, where $\mathcal{A}_i$ (ϵ_i, δ_i)-differentially private, and outputs a function of the results (that is, $\mathcal{A}(x) = g(\mathcal{A}_1(x), \mathcal{A}_2(x), ..., \mathcal{A}_k(x))$ for some probabilistic algorithm g), then $\mathcal{A}$ is $(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i)$-differentially private.*

Here, we are interested in stronger guarantees. In the remainder of

ferential privacy of [27], could lead to better parameters in Theorem 16.

this section, we show differential privacy resists arbitrary amount of side information (a stronger property than composition).

4.1 Semantics of Differential Privacy

There is a crisp, semantically-flavored interpretation of differential privacy, due to Dwork and McSherry, and explained in [13]: *Regardless of external knowledge, an adversary with access to the sanitized database draws the same conclusions whether or not my data is included in the original data.* (the use of the term "semantic" for such definitions dates back to semantic security of encryption [21]).

We require a mathematical formulation of "arbitrary external

knowledge", and of "drawing conclusions". The first is captured via a *prior* probability distribution b on $\mathcal{D}^n$ (b is a mnemonic for "beliefs"). Conclusions are modeled by the corresponding posterior distribution: given a transcript t, the adversary updates his belief about the database D using Bayes' rule to obtain a posterior $\widehat{b}$:

$$\widehat{b}[D|t] \stackrel{\text{def}}{=} \frac{\Pr[\mathcal{A}(D) = t]b[D]}{\sum_{D'} \Pr[\mathcal{A}(D') = t]b[D']} . \tag{1}$$

In an interactive scheme, the definition of $\mathcal{A}$ depends on the adversary's choices; for simplicity we omit the dependence on the adversary in the notation. Also, for simplicity, we discuss only discrete probability distributions. Our results extend directly to the interactive, continuous case.

For a database D, define D_{-i} to be the vector obtained by replacing position i by some default value in D (any value in D will do). This corresponds to "removing" person i's data. We consider $n+1$ related scenarios ("games", in the language of cryptography), numbered 0 through n. In Game 0, the adversary interacts with $\mathcal{A}(D)$. This is the interaction that takes place in the real world. In Game i (for $1 \le i \le n$), the adversary interacts with $\mathcal{A}(D_{-i})$. Game i describes the hypothetical scenario where person i's data is not included.

For a particular belief distribution b and transcript t, we consider the $n+1$ corresponding posterior distributions $\widehat{b}_0, \ldots, \widehat{b}_n$. The posterior $\widehat{b}_0$ is the same as $\widehat{b}$ (defined in Eq. (1)). For larger i, the i-th posterior distribution $\widehat{b}_i$ represents the conclusions drawn in Game i, that is

$$\widehat{b}_i[D|t] \stackrel{\text{def}}{=} \frac{\Pr[\mathcal{A}(D_{-i}) = t]b[D]}{\sum_{D'} \Pr[\mathcal{A}(D'_{-i}) = t]b[D']} .$$

Given a particular transcript t, privacy has been breached if there exists an index i such that the adversary would draw different conclusions depending on whether or not i's data was used. It turns out that the exact measure of "different" here does not matter much. We chose the weakest notion that applies, namely statistical difference. If $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on the set $\mathcal{X}$, the statistical difference between $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$\mathbf{SD}\,(\mathbb{P}, \mathbb{Q}) = \max_{S \subset \mathcal{X}} |\,\mathbb{P}\,[S] - \mathbb{Q}[S]|.$$

DEFINITION 13. *An algorithm $\mathcal{A}$ is ϵ-semantically private if for all prior distributions b on $\mathcal{D}^n$, for all databases $D \in \mathcal{D}^n$, for all possible transcripts t, and for all $i = 1, \ldots, n$,*

$$\mathbf{SD}\left(\widehat{b}_0[D|t] \,,\, \widehat{b}_i[D|t]\right) \le \epsilon.$$

This can be relaxed to allow a probability δ of failure.

DEFINITION 14. *An algorithm is (ϵ, δ)-semantically private if, for all prior distributions b, with probability at least $1 - \delta$ over pairs (D, t), where the database $D \leftarrow b$ (D is drawn according to b) and the transcript $t \leftarrow \mathcal{A}(D)$ (t is drawn according to $\mathcal{A}(D)$), for all $i = 1, \ldots, n$:* $\mathbf{SD}\left(\widehat{b}_0[D|t] \,,\, \widehat{b}_i[D|t]\right) \le \epsilon.$

Dwork and McSherry proposed the notion of semantic privacy, informally, and observed that it is equivalent to differential privacy.

PROPOSITION 15 (DWORK-MCSHERRY). *ϵ-differential privacy implies $\widehat{\epsilon}$-semantic privacy, where $\widehat{\epsilon} = e^\epsilon - 1$.*

We show that this implication holds much more generally:

THEOREM 16 (MAIN RESULT). *(ϵ, δ)-differential privacy implies (ϵ', δ')-semantic privacy where $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$.*

Theorem 16 states that the relaxations notions of differential privacy used in some previous work still imply privacy in the face of arbitrary side information. This is *not* the case for *all* possible relaxations, even very natural ones. For example, if one replaced the multiplicative notion of distance used in differential privacy with total variation distance, then the following "sanitizer" would be deemed private: choose an index $i \in \{1, \ldots, n\}$ uniformly at random and publish the entire record of individual i together with his or her identity (example 2 in [15]). Such a "sanitizer" would not be meaningful at all, regardless of side information.

Finally, the techniques used to prove Theorem 16 can also be used to analyze schemes which do not provide privacy for *all* pairs of neighboring databases D_1 and D_2, but rather only for *most* such pairs (neighboring databases are the ones that differ in one individual). Specifically, it is sufficient that those databases where the "indistinguishability" condition fails occur with small probability.

DEFINITION 17 ((ϵ, δ)-INDISTINGUISHABILITY). *Two random variables X, Y taking values in a set $\mathcal{X}$ are (ϵ, δ)-indistinguishable if for all sets $S \subseteq \mathcal{X}$, $\Pr[X \in S] \le e^\epsilon \Pr[Y \in S] + \delta$ and $\Pr[Y \in S] \le e^\epsilon \Pr[X \in S] + \delta$.*

THEOREM 18. *Let $\mathcal{A}$ be a randomized algorithm. Let $\mathcal{E} = \{D_1 \in \mathcal{D}^n : \forall$ neighbors D_2 of D_1, $\mathcal{A}(D_1)$ and $\mathcal{A}(D_2)$ are (ϵ, δ)-indistinguishable$\}$. Then $\mathcal{A}$ satisfies (ϵ', δ')-semantic privacy for any prior distribution b such that $b[\mathcal{E}] = \Pr_{D_3 \leftarrow b}[D_3 \in \mathcal{E}] \ge 1 - \delta$ with $\epsilon' = e^{3\epsilon} - 1 + 2\sqrt{\delta}$ and $\delta' = O(n\sqrt{\delta})$.*

4.2 Proof Sketch for Main Results

The complete proofs are described in [22]. Here we sketch the main ideas behind both the proofs. Let $Y|_{X=a}$ denote the conditional distribution of Y given that $X = a$ for jointly distributed random variables X and Y. The following lemma (proof omitted) plays an important role in our proofs.

LEMMA 19 (MAIN LEMMA). *Suppose two pairs of random variables $(X, \mathcal{A}(X))$ and $(Y, \mathcal{A}'(Y))$ are (ϵ, δ)-indistinguishable (for some randomized algorithms $\mathcal{A}$ and $\mathcal{A}'$). Then with probability at least $1 - \delta''$ over $t \leftarrow \mathcal{A}(X)$ (equivalently $t \leftarrow \mathcal{A}'(Y)$), the random variables $X|_{\mathcal{A}(X)=t}$ and $Y|_{\mathcal{A}'(Y)=t}$ are $(\hat{\epsilon}, \hat{\delta})$-indistinguishable with $\hat{\epsilon} = 3\epsilon$, $\hat{\delta} = 2\sqrt{\delta}$, and $\delta'' = \sqrt{\delta} + \frac{2\delta}{\epsilon e^\epsilon} = O(\sqrt{\delta})$.*

Let $\mathcal{A}$ be a randomized algorithm (in the setting of Theorem 16, $\mathcal{A}$ is a (ϵ, δ)-differentially private algorithm). Let b be a belief distribution (in the setting of Proposition 18, b is a belief with $b(\mathcal{E}) \ge 1 - \delta$). The main idea behind both the proofs is to use Lemma 19 to show that with probability at least $1 - O(\sqrt{\delta})$ over pairs (D, t) where $D \leftarrow b$ and $t \leftarrow \mathcal{A}(D)$, $\mathbf{SD}\left(b|_{\mathcal{A}(D)=t}, b|_{\mathcal{A}(D_{-i})=t}\right) \le \epsilon'$. Taking a union bound over all coordinates i, implies that with probability at least $1 - O(n\sqrt{\delta})$ over pairs (D, t) where $D \leftarrow b$ and $t \leftarrow \mathcal{A}(D)$, for all $i = 1, \ldots, n$, we have $\mathbf{SD}\left(b|_{\mathcal{A}(D)=t}, b|_{\mathcal{A}(D_{-i})=t}\right) \le \epsilon'$. For Proposition 18, it shows that $\mathcal{A}$ satisfies (ϵ', δ')-semantic privacy for b. In the Theorem 16 setting where $\mathcal{A}$ is (ϵ, δ)-differentially private and b is arbitrary, it shows that (ϵ, δ)-differential privacy implies (ϵ', δ')-semantic privacy.

Acknowledgements

5. REFERENCES

[1] *J. Official Statistics*, special issue on disclosure limitation methods, Vol. 14, No. 4, 1998.

[2] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, pages 439–450. ACM Press, 2000.

[3] S. Agrawal and J. R. Haritsa. A framework for high-accuracy privacy-preserving mining. In *ICDE*, pages 193–204. IEEE Computer Society, 2005.

[4] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282. ACM Press, 2007.

[5] M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. *The New York Times*, Aug. 2006.

[6] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The SuLQ framework. In *PODS*, pages 128–138. ACM Press, 2005.

[7] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *STOC*, pages 609–618. ACM Press, 2008.

[8] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li. Secure anonymization for incremental datasets. In *Secure Data Management*, pages 48–63. Springer, 2006.

[9] K. Chaudhuri and N. Mishra. When random sampling preserves privacy. In *CRYPTO*, pages 198–213. Springer, 2006.

[10] B.-C. Chen, R. Ramakrishnan, and K. LeFevre. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB*, pages 770–781. VLDB Endowment, 2007.

[11] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM Press, 2003.

[12] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[13] C. Dwork. Differential privacy. In *ICALP*, pages 1–12. Springer, 2006.

[14] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, pages 486–503. Springer, 2006.

[15] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.

[16] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *CRYPTO*, pages 528–544. Springer, 2004.

[17] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *KDD*, pages 217–228. ACM Press, 2002.

[18] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222. ACM Press, 2003.

[19] B. C. M. Fung, K. Wang, A. W.-C. Fu, and J. Pei. Anonymity for continuous data publishing. In *EDBT*, pages 264–275. ACM Press, 2008.

[20] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. *CoRR*, arXiv:0803.0032v1 [cs.DB], 2008.

[21] S. Goldwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270–299, 1984.

[22] S. P. Kasiviswanathan and A. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, arXiv:0803.39461 [cs.CR], 2008.

[23] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, page 25. IEEE Computer Society, 2006.

[24] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *KDD*, pages 277–286. ACM Press, 2006.

[25] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115. IEEE Computer Society, 2007.

[26] Y. Lindell. *Composition of Secure Multi-Party Protocols: A Comprehensive Study*. Springer-Verlag, 2003.

[27] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: From theory to practice on the map. In *ICDE*, pages 277–286. IEEE Computer Society, 2008.

[28] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[29] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, pages 126–135. IEEE Computer Society, 2007.

[30] F. McSherry and K. Talwar. Differential privacy in mechanism design. In *FOCS*, pages 94–103. IEEE Computer Society, 2007.

[31] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM Press, 2007.

[32] J. Pei, J. Xu, Z. Wang, W. Wang, and K. Wang. Maintaining k-anonymity against incremental updates. In *SSDBM*, page 5, 2007.

[33] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[34] UCI machine learning repository, http://www.ics.uci.edu/ mlearn/databases/.

[35] A. van den Hout and P. van der Heijden. Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, 70:269–288, 2002.

[36] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *KDD*, pages 414–423. ACM Press, 2006.

[37] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[38] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, pages 543–554. VLDB Endowment, 2007.

[39] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.

[40] X. Xiao and Y. Tao. Personalized Privacy Preservation In *SIGMOD*, pages 229–240, 2006.

[41] X. Xiao and Y. Tao. M-invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, pages 689–700. ACM Press, 2007.

[42] C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. In *VLDB*, p. 910–921, 2005.