ℓ-Diversity: Privacy Beyond k-Anonymity

Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer
Muthuramakrishnan Venkitasubramaniam
Cornell University

Abstract

Publishing data about individuals without revealing sensitive information about them is an important problem. In recent years, a new definition of privacy called k-anonymity has gained popularity. In a k-anonymized dataset, each record is indistinguishable from at least $k-1$ other records with respect to certain "identifying" attributes.

In this paper we show with two simple attacks that a k-anonymized dataset has some subtle, but severe privacy problems. First, we show that an attacker can discover the values of sensitive attributes when there is little diversity in the sensitive attributes. Second, real attackers often have background knowledge and we show that k-anonymity does not guarantee privacy against attackers with background knowledge. We give a detailed analysis of these two attacks and we propose a novel powerful privacy definition called ℓ-diversity. In addition to building a formal foundation for ℓ-diversity, we show in an experimental evaluation that ℓ-diversity is practical and can be implemented efficiently.

1. Introduction

Many organizations are increasingly publishing microdata – tables that contain unaggregated information about individuals. These tables include medical, voter registration, census, and customer data. Microdata is a valuable source of information for the allocation of public funds, medical research, and trend analysis. However, if individuals can be uniquely identified in the microdata then their private information (such as medical condition) would be disclosed, and this is unacceptable.

To avoid the identification of records in microdata, uniquely identifying information like names and social security numbers are removed from the table. However, this first sanitization still does not ensure the privacy of individuals in the data. A recent study estimated that 87% of the population of the United States can be uniquely identified using the seemingly innocuous attributes gender, date of birth, and 5-digit zip code [23]. In fact, those three attributes were used to link Massachusetts voter registration records (which included the name, gender, zip code, and date of birth) to supposedly anonymized medical data from GIC (which included gender, zip code, date of birth and diagnosis)[1]. This "linking attack" managed to uniquely identify the medical records of the governor of Massachusetts in the medical data [24].

Sets of attributes (like gender, date of birth, and zip code in the example above) that can be linked with external data to uniquely identify individuals in the population are called *quasi-identifiers*. To counter linking attacks using quasi-identifiers, Samarati and Sweeney proposed a definition of privacy called *k-anonymity* [21, 24]. A table satisfies k-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes; such a table is called a *k-anonymous* table. Hence, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks.

An Example. Figure 1 shows medical records from a fictitious hospital located in upstate New York. Note that the table contains no uniquely identifying attributes like name, social security number, etc. In this example, we divide the attributes into two groups: the *sensitive* attributes (consisting only of medical condition) and the *non-sensitive* attributes (zip code, age, and nationality). An attribute is marked sensitive if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset. Attributes not marked sensitive are non-sensitive. Furthermore, let the collection of attributes $\{$zip code, age, nationality$\}$ be the quasi-identifier for this dataset. Figure 2 shows a 4-anonymous table derived from the table in Figure 1 (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range $[14850-14859]$ and "age=3*" means the age is in the range $[30-39]$). Note that in the 4-anonymous table, each tuple has the same values for the quasi-identifier as at least three other tuples in the table.

[1] Group Insurance Company (GIC) is responsible for purchasing health insurance for Massachusetts state employees.

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	$*$	Heart Disease
2	130**	< 30	$*$	Heart Disease
3	130**	< 30	$*$	Viral Infection
4	130**	< 30	$*$	Viral Infection
5	1485*	≥ 40	$*$	Cancer
6	1485*	≥ 40	$*$	Heart Disease
7	1485*	≥ 40	$*$	Viral Infection
8	1485*	≥ 40	$*$	Viral Infection
9	130**	$3*$	$*$	Cancer
10	130**	$3*$	$*$	Cancer
11	130**	$3*$	$*$	Cancer
12	130**	$3*$	$*$	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Because of its conceptual simplicity, k-anonymity has become synonymous with privacy, and due to algorithmic advances in creating k-anonymous versions of a dataset [21, 24, 3, 18, 6, 16, 25], k-anonymity has grown in popularity. However, does k-anonymity really guarantee privacy? In the next section, we will show that the answer to this question is interestingly *no*. We give examples of two simple, yet subtle, attacks on a k-anonymous dataset that allow an attacker to identify individual records. Defending against these attacks requires a stronger notion of privacy called ℓ-diversity, the focus of this paper. But we are jumping ahead in our story. Let us first show the two attacks to give the intuition behind the problems with k-anonymity.

1.1. Attacks On k-Anonymity

In this section we present two attacks, the *homogeneity attack* and the *background knowledge attack*, and we show how they can be used to compromise a k-anonymous dataset.

Homogeneity Attack: Alice and Bob are antagonistic neighbors. One day Bob falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Alice sets out to discover what disease Bob is suffering from. Alice discovers the 4-anonymous table of current inpatient records published by the hospital (Figure 2), and so she knows that one of the records in this table corresponds to Bob. Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053. Therefore, Alice knows that Bob's record number is 9,10,11, or 12. Now, all of those patients have the same medical condition (cancer), and so Alice concludes that Bob has cancer.

Observation 1 *k-Anonymity can create groups that due to lack of diversity in the the sensitive attribute leak information.*

Note that such a situation is not uncommon. As a back-of-the-envelope calculation, suppose we have dataset containing 60,000 distinct tuples where the sensitive attribute can take 3 distinct values and is not correlated with the non-sensitive attributes. A 5-anonymization of this table will have around 12,000 groups[2] and, on average, 1 out of every 81 groups will have no diversity (the values for the sensitive attribute will all be the same). Thus we should expect about 148 groups with no diversity; multiplying that by 5 we see that information about 740 people can be compromised by a homogeneity attack. This suggests that in addition to k-anonymity, the sanitized table should also ensure "diversity" – all tuples that share the same values of their quasi-identifiers should have diverse values for their sensitive attributes.

Our next observation is that an adversary could use "background" knowledge to discover sensitive information.

Background Knowledge Attack: Alice has a pen-friend named Umeko who is admitted to the same hospital as Bob, and whose patient records also appear in the table shown in Figure 2. Alice knows that Umeko is a 21 year-old Japanese female who currently lives in zip code 13068. Based on this information, Alice learns that Umeko's information is contained in record number 1,2,3, or 4. Without additional information, Alice is not sure whether Umeko caught a virus or has heart disease. However, it is well-known that Japanese have an extremely low incidence of heart disease. Therefore Alice concludes with near certainty that Umeko has a viral infection.

Observation 2 *k-Anonymity does not protect against attacks based on background knowledge.*

We have demonstrated (using the homogeneity and background knowledge attacks) that a k-anonymous table may

[2] In reality data is very skewed and a 5-anonymous table might not have so many groups. See section 6.

disclose sensitive information. Since both of these attacks are plausible in real life, we need a stronger definition of privacy that takes into account diversity and background knowledge. This paper addresses this very issue.

1.2. Contributions and Paper Outline

In the previous section, we showed that k-anonymity is susceptible to the homogeneity and the background knowledge attack; thus a stronger definition of privacy is needed. In the remainder of the paper, we start by introducing an ideal notion of privacy called *Bayes-optimal* for the case that both data publisher and the adversary have full background knowledge (Section 3). Unfortunately, the data publisher is unlikely to possess this information, and in addition, the adversary may have more specific background knowledge than the data publisher. Hence, while Bayes-optimal privacy sounds great in theory, it is unlikely that it can be guaranteed in practice. To address this problem, we show that the notion of Bayes-optimal privacy naturally leads itself to a novel *practical* definition that we call ℓ-diversity that provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main idea behind ℓ-diversity is the requirement that the values of the sensitive attributes are well-represented in each group (Section 4).

We show that existing algorithms for k-anonymity can be adapted to compute ℓ-diverse tables (Section 5), and in an experimental evaluation we show that ℓ-diversity is practical and can be implemented efficiently (Section 6). We discuss related work in Section 7, and we conclude in Section 8. Before jumping into the contributions of this paper starting with Section 3, we introduce the notation needed to formally discuss data privacy in the next section.

2. Model and Notation

In this section we will introduce some basic notation that will be used in later sections. We will also discuss how a table can be anonymized and what kind of background knowledge an adversary may possess.

Basic Notation. Let $T = \{t_1, t_2, \ldots, t_n\}$ be a table with attributes $A_1, \ldots, A_m$. We assume that T is a subset of some larger population Ω where each tuple represents an individual from the population. For example, if T is a medical dataset then Ω could be the population of the United States. Let $\mathcal{A}$ denote the set of all attributes $\{A_1, A_2, \ldots, A_m\}$ and $t[A_i]$ denote the value of attribute A_i for tuple t. If $\mathcal{C} = \{C_1, C_2, \ldots, C_p\} \subseteq \mathcal{A}$ then we use the notation $t[\mathcal{C}]$ to denote the tuple $(t[C_1], \ldots, t[C_p])$, which is the projection of t onto the attributes in $\mathcal{C}$.

In privacy-preserving data publishing, there exist several important subsets of the set of all attributes $\mathcal{A}$. A *sensitive attribute* is an attribute whose value for any particular individual must be kept secret from people who have no direct access to the original data. Let $\mathcal{S}$ denote the set of all sensitive attributes. An example of a sensitive attribute is *Medical Condition* from Figure 1. The association between individuals and *Medical Condition* should be kept secret; thus we should not disclose which particular patients have cancer, but it is permissible to disclose the information that there exist cancer patients in the hospital. We assume that the data publisher knows which attributes are sensitive. All attributes that are not sensitive are called *nonsensitive* attributes. Let $\mathcal{N}$ denote the set of nonsensitive attributes. Let us now formally define the notion of a quasi-identifier.

Definition 2.1 (Quasi-identifier) *A* quasi-identifier *is a setwise minimal collection of nonsensitive attributes $\{Q_1, \ldots, Q_w\}$ that uniquely identifies at least one individual in the general population Ω.*

One example of a quasi-identifier is a primary key. Another example is the set {Gender, Age, Zip Code} in the GIC dataset that was used to identify the governor of Massachusetts as described in the introduction. We denote the set of all quasi-identifiers by $\mathcal{QI}$. We are now ready to formally define when a table satisfies k-anonymity.

Definition 2.2 (k-Anonymity) *A table T satisfies k-Anonymity if for every tuple $t \in T$ there exist $k-1$ other tuples $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ such that $t[\mathcal{C}] = t_{i_1}[\mathcal{C}] = t_{i_2}[\mathcal{C}] = \cdots = t_{i_{k-1}}[\mathcal{C}]$ for all $\mathcal{C} \in \mathcal{QI}$.*

The Anonymized Table $T^\star$. Since the quasi-identifiers might uniquely identify tuples in T, the table T is not published; it is subjected to an *anonymization procedure* and the resulting table $T^\star$ is published instead.

There has been a lot of research on techniques for anonymization (see Section 7 for a discussion of related work). These techniques can be broadly classified into *generalization* techniques [16, 3], *generalization with tuple suppression* techniques [22, 6], and *data swapping and randomization* techniques [1, 13]. In this paper we limit our discussion only to generalization techniques.

Definition 2.3 (Domain Generalization) *A domain $D^\star = \{P_1, P_2, \ldots\}$ is a* generalization *(partition) of a domain D if $\bigcup P_i = D$ and $P_i \cap P_j = \emptyset$ whenever $i \neq j$. For $x \in D$ we let $\phi_{D^\star}(x)$ denote the element $P \in D^\star$ that contains x.*

Note that we can create a partial order $<_G$ on domains by requiring $D <_G D^\star$ if and only if $D^\star$ is a generalization of D. Given a table $T = \{t_1, \ldots, t_n\}$ with the set of nonsensitive attributes $\mathcal{N}$ and a generalization $D^\star_N$ of domain($\mathcal{N}$), we can construct a table $T^\star = \{t^\star_1, \ldots, t^\star_n\}$ by replacing the value of $t_i[N]$ with the generalized value $\phi_{D^\star_N}(t_i[N])$ to get a new tuple $t^\star_i$. The table $T^\star$ is called a *generalization* of T

and we use the notation $T \to^\star T^\star$ to mean "$T^\star$ is a generalization of T". Typically, ordered attributes are partitioned into intervals, and categorical attributes are partitioned according to a user-defined hierarchy (for example, cities are generalized to counties, counties to states, and states to regions).

Example 1 (Continued). The table in Figure 2 is a generalization of the table in Figure 1. We generalized on the *Zip Code* attribute by partitioning it into two sets: "1485*" (representing all zip codes whose first four digits are 1485) and "130*" (representing all zip codes whose first three digits are 130). Then we partitioned *Age* into three groups: "< 30", "3*" (representing all ages between 30 and 39), and "≥ 40". Finally, we partitioned *Nationality* into just one set "*" representing all nationalities.

The Adversary's Background Knowledge. Since the background knowledge attack was enabled due to additional knowledge about the table that the adversary had, let us shortly discuss the type of background knowledge of an attacker that we consider.

First, the adversary has of course access to the published table $T^\star$, and she also knows that $T^\star$ is a generalization of some base table T. The adversary also knows the domain of each attribute of T.

Second, the adversary could know that some individuals are in the table. This knowledge is often easy to acquire. For example, GIC published medical data about Massachusetts state employees. If the adversary Alice knows that her neighbor Bob is a state employee of Massachusetts then Alice is almost certain that Bob's information is contained in that table. In this case, we assume that Alice knows all of Bob's nonsensitive attributes. In addition, the adversary could have knowledge about the sensitive attributes of specific individuals in the population and/or the table. For example, the adversary Alice might know that neighbor Bob does not have pneumonia since Bob does not show any of the symptoms of pneumonia. We call such knowledge "instance-level background knowledge," since it is associated with specific instances in the table.

Third, the adversary could have partial knowledge about the distribution of sensitive and nonsensitive attributes in the population. We call this "demographic background knowledge". For example, the adversary may know $P\left(t[\text{Condition}] = \text{``cancer''} \middle| \, t[\text{Age}] \geq 40\right)$, and may use it to make additional inferences about records in the table.

Now armed with the right notation, let us start looking into principles and definitions of privacy that leak little information.

3. Bayes-Optimal Privacy

In this section we analyze an ideal notion of privacy called *Bayes-Optimal Privacy* since it involves modeling background knowledge as a probability distribution over the attributes and uses bayesian inference techniques to reason about privacy. We introduce tools for reasoning about privacy (Section 3.1), we use them to discuss theoretical principles of privacy (Section 3.2), and then we point out the difficulties that need to be overcome to arrive at a practical definition of privacy (Section 3.3).

3.1. Changes in Belief Due to Data Publishing

For simplicity of discussion, we will combine all the nonsensitive attributes into a single, multi-dimensional quasi-identifier attribute Q whose values of Q are generalized to create the anonymized table $T^\star$ from the base table T. Since Bayes-optimal privacy is only used to motivate a practical definition, we make the following other two simplifying assumptions: We assume that there is a single sensitive attribute S, and that T is a simple random sample[3] from some larger population Ω.[4]

Recall that in our attack model, the adversary Alice has partial knowledge of the distribution of the sensitive and non-sensitive attributes. Let us assume a worst case scenario where Alice knows the complete joint distribution f of Q and S. She knows that Bob corresponds to a record $t \in T$ that has been generalized to a record t^* in the published table $T^\star$, and she also knows the value of Bob's non-sensitive attributes (i.e., she knows that $t[Q] = q$). Alice's goal is to use her background knowledge to discover Bob's sensitive information — the value of $t[S]$. We gauge her success using two quantities: Alice's *prior belief*, and her *observed belief*.

Alice's prior belief, $\alpha_{(q,s)}$, that Bob's sensitive attribute is s given that his nonsensitive attribute is q, is just her background knowledge:

$$\alpha_{(q,s)} = P_f\left(t[S] = s \middle| \, t[Q] = q\right).$$

After Alice observes the table $T^\star$, her belief about Bob's sensitive attribute changes. This new belief, $\beta_{(q,s,T^\star)}$, is her *observed belief*:

$$\beta_{(q,s,T^\star)} = P_f\left(t[S] = s \middle| \, t[Q] = q \wedge t \in T^\star\right).$$

Given f and $T^\star$, we can derive using bayesian reasoning a formula for $\beta_{(q,s,T^\star)}$. This is a key technical result of this paper since in this formula lies the foundation for this section and the next. Deriving the expression for the observed belief is quite tricky. The main idea is to find a set of equally likely disjoint random worlds (like in [5]) such that the conditional probability $P[A|B]$ is the number of worlds satisfying the condition $A \wedge B$ divided by the number of worlds

[3] A sample of size n drawn without replacement is called a *simple random sample* if every sample of size n is equally likely.

[4] We would like to emphasize that both these assumptions will be dropped in Section 4 when we introduce a practical definition of privacy.

satisfying the condition B. We avoid double-counting since the random worlds are disjoint. In our case, every permutation of a simple random sample of size n of the population that is *compatible* with the published table $T^\star$ is a random world.[5]

Theorem 3.1 *Let q be the value of the nonsensitive attribute in the base table T; $q^\star$ be the generalized value of q in the published table $T^\star$; s be a possible value of the sensitive attribute; $n_{(q^\star,s')}$ be the number of tuples $t^\star \in T^\star$ in the published table where $t^\star[Q] = q^\star$ and $t^\star[S] = s'$; and let $f(s' \mid q^\star)$ be the conditional probability of the sensitive attribute conditioned on the fact that the nonsensitive attribute Q belongs to the part of its domain that is represented by $q^\star$ (for example, if $q^\star = 1485*$ then $q^\star$ represents the zip codes $\{14850, 14851, \ldots, 14859\}$). Then the following relationship holds:*

$$\beta_{(q^\star,s,T^\star)} = \frac{n_{(q^\star,s)} \frac{f(s|q)}{f(s|q^\star)}}{\sum_{s' \in S} n_{(q^\star,s')} \frac{f(s'|q)}{f(s'|q^\star)}} \tag{1}$$

Armed with a way of calculating Alice's belief about Bob's private data after seeing T^*, let us now examine some principle for building definitions of privacy.

3.2. Privacy Principles

Given the adversary's background knowledge, a published table $T^\star$ might disclose information in two important ways: *positive disclosure* and *negative disclosure*.

Definition 3.1 (Positive disclosure) *Publishing the table $T^\star$ that was derived from T results in a positive disclosure if the adversary can correctly identify the value of a sensitive attribute with high probability; i.e., given a $\delta > 0$, there is a positive disclosure if $\beta_{(q,s,T^\star)} > 1 - \delta$ and there exists $t \in T$ such that $t[Q] = q$ and $t[S] = s$.*

Definition 3.2 (Negative disclosure) *Publishing the table $T^\star$ that was derived from T results in a negative disclosure if the adversary can correctly eliminate some possible values of the sensitive attribute (with high probability); i.e., given an $\epsilon > 0$, there is a negative disclosure if $\beta_{(q,s,T^\star)} < \epsilon$ and there exists a $t \in T$ such that $t[Q] = q$ but $t[S] \neq s$.*

The homogeneity attack in Section 1.1 where Alice determined that Bob has cancer is an example of a positive disclosure. Similarly, in the example from Section 1.1, Alice can deduce that Umeko does not have cancer, and this is an example of a negative disclosure.

[5]Due to space constraints we had to omit the proof of the following theorem; see [17] for the derivation of Equation 1.

Note that not all positive disclosures are disastrous. If the prior belief was that $\alpha_{(q,s)} > 1 - \delta$, the adversary would not have learned anything new. Similarly, negative disclosures are not always bad: discovering that Bob does not have Ebola might not be very serious because the prior belief of this event was small. Hence, the ideal definition of privacy can be based on the following principle:

Principle 3.1 (Uninformative Principle) *The published table should provide the adversary with little additional information beyond the background knowledge. In other words, there should not be a large difference between the prior and observed beliefs.*

The uninformative principle can be instantiated in several ways, for example with the (ρ_1, ρ_2)-privacy breach definition [14]. Under this definition, privacy is breached either when $\alpha_{(q,s)} < \rho_1 \wedge \beta_{(q,s,T^\star)} > \rho_2$ or when $\alpha_{(q,s)} > 1 - \rho_1 \wedge \beta_{(q,s,T^\star)} < 1 - \rho_2$. An alternative privacy definition based on the uninformative principle would bound the maximum difference between $\alpha_{(q,s)}$ and $\beta_{(q,s,T^\star)}$ using any of the functions commonly used to measure the difference between probability distributions. Any privacy definition that is based on the uninformative principle, and instantiated either by a (ρ_1, ρ_2)-privacy breach definition or by bounding the difference between $\alpha_{(q,s)}$ and $\beta_{(q,s,T^\star)}$ is a Bayes-optimal privacy definition. The specific choice of definition depends on the application.

Note that any Bayes-optimal privacy definition captures diversity as well as background knowledge. To see how it captures diversity, suppose that all the tuples whose nonsensitive attribute Q have been generalized to $q^\star$ have the same value s for their sensitive attribute. Then $n_{(q^\star,s')} = 0$ for all $s' \neq s$ and hence the value of the observed belief becomes 1 in Equation 1. This will be flagged as a breach whenever the prior belief is not close to 1.

3.3. Limitations of the Bayes-Optimal Privacy

For the purposes of our discussion, we are more interested in the properties of Bayes-optimal privacy rather than its exact instantiation. In particular, Bayes-optimal privacy has several drawbacks that make it hard to use in practice.

Insufficient Knowledge. The data publisher is unlikely to know the full distribution f of sensitive and nonsensitive attributes over the general population Ω from which T is a sample.

The Adversary's Knowledge is Unknown. It is also unlikely that the adversary has knowledge of the complete joint distribution between the non-sensitive and sensitive attributes. However, the data publisher does not know how much the adversary knows. In fact, the adversary may have more information than the data publisher.

Instance-Level Knowledge. The theoretical definition does not protect against knowledge that cannot be modeled by a probabilistic setting. For example, suppose Bob's son tells Alice that Bob does not have diabetes. The theoretical definition of privacy will not be able to protect against such adversaries.

Multiple Adversaries. There will likely be multiple adversaries with different levels of knowledge. Suppose Bob has a disease that is rare for his age. An adversary who knows the interaction of age and illness will think that Bob is unlikely to have that disease. An adversary who does not know how age and illness interact is more likely to think that Bob does have that disease. Thus, although additional knowledge can yield better inferences on average, there are specific instances where it doesn't. Thus the data publisher must take into account all possible levels of background knowledge.

Let us thus turn to the next section to a definition that eliminates these drawbacks.

4. ℓ-Diversity: A Practical Definition

In this section we discuss how to overcome the difficulties outlined at the end of the previous section. We derive the ℓ-diversity principle (Section 4.1), show how to instantiate it with specific definitions of privacy (Section 4.2), outline how to handle multiple sensitive attributes (Section 4.3), and finally discuss how ℓ-diversity addresses the issues raised in the previous section.

4.1. The ℓ-Diversity Principle

Recall Theorem 3.1 that allows us to calculate the observed belief of the adversary. Let us define a $q^\star$*-block* to be the set of tuples in $T^\star$ whose nonsensitive attribute values generalize to $q^\star$. Consider the case of positive disclosures; i.e., Alice wants to determine that Bob has $t[S] = s$ with very high probability, i.e., $\beta_{(q,s,T^\star)} \approx 1$. From Theorem 3.1, this can happen only when:

$$\forall s' \neq s, \quad n_{(q^\star,s')} \frac{f(s'|q)}{f(s'|q^\star)} \ll n_{(q^\star,s)} \frac{f(s|q)}{f(s|q^\star)} \tag{2}$$

The condition in Equation (2) could occur due to a combination of two factors: (i) a lack of diversity in the sensitive attributes in the $q^\star$-block, and/or (ii) strong background knowledge. Let us discuss these in turn.

Lack of Diversity. Lack of diversity in the sensitive attribute manifests itself as follows:

$$\forall s' \neq s, \quad n_{(q^\star,s')} \ll n_{(q^\star,s)} \tag{3}$$

In this case, almost all tuples have the same value s for the sensitive attribute S, and thus $\beta_{(q,s,T^\star)} \approx 1$. Note that such insufficient diversity can be easily checked since it only involves counting the values of S in the published table $T^\star$. It can also be easily ensured by requiring that *all* the possible values $s' \in domain(S)$ occur in the $q^\star$-block with roughly equal proportions. This, however, is likely to cause significant loss of information: if $domain(S)$ is large then the $q^\star$-blocks will necessarily be large and so the data will be partitioned into only a small number of $q^\star$-blocks. Another way to ensure diversity and to disallow Equation 3 to hold is to require that a $q^\star$-block has at least $\ell \geq 2$ different sensitive values such that the ℓ most frequent values (in the $q^\star$-block) have roughly the same frequency. We say that such a $q^\star$-block is *well-represented by ℓ sensitive values*.

Strong Background Knowledge. The other factor that could lead to a positive disclosure (Equation 2) is strong background knowledge. Even though a $q^\star$-block may have ℓ "well-represented" sensitive values, Alice may still be able to eliminate all but one sensitive value using her background knowledge by forcing the following to happen:

$$\frac{f(s'|q)}{f(s'|q^\star)} \approx 0 \tag{4}$$

This equation states that Bob with quasi-identifier $t[Q] = q$ is much less likely to have sensitive value s' than any other individual in the $q^\star$-block. For example, Alice may know that Bob never travels, and thus he is extremely unlikely to have Ebola. It is not possible for a data publisher to guard against attacks employing arbitrary amounts of background knowledge. However, the data publisher can still guard against many attacks even without having access to Alice's background knowledge. In our model, Alice might know the distribution $f(q, s)$ over the sensitive and non-sensitive attributes, as well as the conditional distribution $f(s|q)$. The most damaging type of such information has the form $f(s|q) \approx 0$, e.g., "men do not have breast cancer", or of the form of equation 4, e.g., "among Asians, Japanese have a very low incidence of heart disease".[6] Alice also could have damaging instance-level knowledge such as "Bob does not have diabetes".

In spite of such background knowledge, if there are ℓ "well represented" sensitive values in a $q^\star$-block, then Alice needs $\ell - 1$ damaging pieces of background knowledge to eliminate $\ell - 1$ possible sensitive values and infer a positive disclosure! Thus, by setting the parameter ℓ, the data publisher can determine how much protection is provided against background knowledge — even if this background knowledge is unknown to the publisher.

Putting these two arguments together, we arrive at the following principle.

Principle 4.1 (ℓ-Diversity Principle) *A $q^\star$-block is ℓ-diverse if contains at least ℓ "well-represented" values for*

[6] Note that *a priori* information of the form $f(s|q) = 1$ is not included since this positive disclosure is independent of the published table $T^\star$

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	$*$	Heart Disease
4	1305*	≤ 40	$*$	Viral Infection
9	1305*	≤ 40	$*$	Cancer
10	1305*	≤ 40	$*$	Cancer
5	1485*	> 40	$*$	Cancer
6	1485*	> 40	$*$	Heart Disease
7	1485*	> 40	$*$	Viral Infection
8	1485*	> 40	$*$	Viral Infection
2	1306*	≤ 40	$*$	Heart Disease
3	1306*	≤ 40	$*$	Viral Infection
11	1306*	≤ 40	$*$	Cancer
12	1306*	≤ 40	$*$	Cancer

Figure 3. 3-Diverse Inpatient Microdata

the sensitive attribute S. A table is ℓ-diverse if every $q^\star$-block is ℓ-diverse.

Returning to our example, consider the inpatient records shown in Figure 1. We present a 3-diverse version of the table in Figure 3. Comparing it with the 4-anonymous table in Figure 2 we see that the attacks against the 4-anonymous table are prevented by the 3-diverse table. For example, Alice cannot infer from the 3-diverse table that Bob (a 31 year old American from zip code 13053) has cancer. Even though Umeko (a 21 year old Japanese from zip code 13068) is extremely unlikely to have heart disease, Alice is still unsure whether Umeko has a viral infection or cancer.

The ℓ-diversity principle advocates ensuring ℓ "well represented" values for the sensitive attribute in every $q^\star$-block, but does not clearly state what "well represented" means.[7] Let us now give two concrete instantiations of the ℓ-diversity principle and discuss their relative trade-offs.

4.2. ℓ-Diversity: Instantiations

Let us follow a path where along the way we encounter different privacy definitions, all motivated by the ℓ-diversity principle. One instantiation of the ℓ-diversity principle is using the information-theoretic notion of entropy:

Definition 4.1 (Entropy ℓ-Diversity) *A table is* Entropy ℓ-Diverse *if for every $q^\star$-block*

$$-\sum_{s \in S} p_{(q^\star,s)} \log(p_{(q^\star,s')}) \geq \log(\ell)$$

where $p_{(q^\star,s)} = \frac{n_{(q^\star,s)}}{\sum\limits_{s' \in S} n_{(q^\star,s')}}$ is the fraction of tuples in the $q^\star$-block with sensitive attribute value equal to s.

[7]Note that we called it a "principle" instead of a theorem — we will use it to give formal definitions of privacy in the next section.

This condition guarantees that every $q^\star$-block has at least ℓ distinct values for the sensitive attribute. Using this definition, Figure 3 is actually 2.8-diverse.

Since $-x\log(x)$ is a concave function, it can be shown that if we split a $q^\star$-block into two sub-blocks $q^\star_a$ and $q^\star_b$ then $\mathrm{entropy}(q^\star) \geq \min(\mathrm{entropy}(q^\star_a, q^\star_b))$. This implies that in order for entropy ℓ-diversity to be possible, the entropy of the entire table must be $\geq \log(\ell)$. This might not be the case, especially if one value of the sensitive attribute is very common – for example, if 90% of the patients have "no serious problems" as the value for the "Medical Condition" attribute.

Thus entropy ℓ-diversity may sometimes be too restrictive. If the data publisher has some background knowledge that some positive disclosures are allowable (for example, we are allowed to disclose that a patient has "no serious problems" because it appears so often in the table) then we can do better. This reasoning follows the uninformative principle and leads us to develop a less conservative instantiation of the ℓ-diversity principle called *recursive ℓ-diversity*.

Let $s_1, \ldots, s_m$ be the possible values of the sensitive attribute S in a $q^\star$-block. Assume that we sort the counts $n_{(q^\star,s_1)}, \ldots, n_{(q^\star,s_m)}$ in descending order and name the elements of the resulting sequence $r_1, \ldots, r_m$. One way to think about ℓ-diversity is the following: the adversary needs to eliminate at least $\ell - 1$ possible values of S in order to infer a positive disclosure. This means that, for example, in a 2-diverse table, none of the sensitive values should appear too frequently. However, eliminating one of the sensitive values might cause one or more of the remaining sensitive values to appear too frequently. In particular, the most frequent sensitive value in the $q^\star$-block (the one that appears r_1 times) might become too frequent. So if the $q^\star$-block satisfies 2-diversity then $r_1 < c(r_2 + \cdots + r_m)$ for some user-specified constant c. Generalizing this to $\ell > 2$, a $q^\star$-block satisfies ℓ-diversity if we can remove $\ell - 2$ sensitive values and still have a 2-diverse group. This is what we call *recursive* (c,ℓ)-diversity, and we define it formally as follows.

Definition 4.2 (Recursive (c,ℓ)-Diversity) *A $q^\star$-block satisfies* recursive (c,ℓ)-diversity *if $r_1 < c(r_\ell + r_{\ell+1} + \cdots + r_m)$ (the most frequent sensitive value is not too frequent after we have removed the next $\ell - 2$ most frequent sensitive values). A table $T^\star$ satisfies recursive (c,ℓ)-diversity if every $q^\star$-block satisfies recursive ℓ-diversity. For pedantic reasons, we say that 1-diversity is always trivially satisfied.*

Let us continue our path. Now suppose that Y is the set of sensitive values for which positive disclosure is allowed (for example, they are extremely frequent, or they may not be an invasion of privacy – like "Medical Condition"="Healthy"). Since we are not worried about those values being too frequent, let s_y be the most frequent sensitive value in the $q^\star$-block that is *not* in Y and let r_y be the associated frequency. Then the $q^\star$-block satisfies ℓ-diversity if we can eliminate the $\ell - 2$ most frequent values of S *not including* r_y without making s_y too frequent in the resulting set. This is the same as saying that after we remove the sensitive values with counts $r_1, \ldots, r_{y-1}$, then the result is $(\ell - y + 1)$-diverse. This brings us to the following definition.

Definition 4.3 (Positive Disclosure-Recursive (c, ℓ)-Diversity) *A $q^\star$-group satisfies* pd-recursive (c, ℓ)-diversity *if one of the following hold:*

- $y \leq \ell - 1$ *and* $r_y < c \sum_{j=\ell}^{m} r_j$
- $y > \ell - 1$ *and* $r_y < c \sum_{j=\ell-1}^{y-1} r_j + c \sum_{j=y+1}^{m} r_j$

We denote the summations on the right hand side of the both conditions by $tail_{q^\star}(s_y)$.

Now, note that if $r_y = 0$ then the $q^\star$-block only has sensitive values that can be disclosed and so both conditions in Definition 4.3 are trivially satisfied. Second, note that if $c > 1$ then the second condition clearly reduces to just the condition $y > \ell - 1$ because $r_y \leq r_{\ell-1}$. The second condition states that even though the $\ell - 1$ most frequent values can be disclosed, we still don't want r_y to be too frequent if $\ell - 2$ of them have been eliminated (i.e. we want the result to be 2-diverse).

Until now we have treated negative disclosure as relatively unimportant compared to positive disclosure. However, negative disclosure may also be important. If W is the set of values for the sensitive attribute for which negative disclosure is not allowed then, given a user-specified constant $c_2 < 100$, we require that each $s \in W$ appear in at least c_2-percent of the tuples in every $q^\star$-block, resulting in the following definition.

Definition 4.4 (Negative/Positive Disclosure) -Recursive (c_1, c_2, ℓ)-Diversity *Let W be the set of sensitive values for which negative disclosure is not allowed. A table satisfies* npd-recursive (c_1, c_2, ℓ)-diversity *if it satisfies pd-recursive (c_1, ℓ)-diversity and if every $s \in W$ occurs in at least c_2 percent of the tuples in every $q^\star$-block.*

4.3. Multiple Sensitive Attributes

Having multiple sensitive attributes presents additional challenges. Suppose there are two sensitive attributes S and V and a $q^\star$-block with the following tuples: $\{(q^\star, s_1, v_1), (q^\star, s_1, v_2), (q^\star, s_2, v_3), (q^\star, s_3, v_3)\}$. This $q^\star$-block is 3-diverse (actually recursive (2,3)-diverse) with respect to S (ignoring V) and 3-diverse with respect to V (ignoring S). However, if we know that Bob is in this group and his value for S is not s_1 then we have also determined that his value for V cannot be v_1 or v_2 and so it must be v_3. Thus a q^*-block that is ℓ-diverse in each sensitive attribute separately may still violate the principle of ℓ-diversity. Note that k-anonymity also suffers from these drawbacks. Furthermore, multiple sensitive attributes increase the chance that a k-anonymous group will lack diversity in at least one sensitive attribute.

One solution is to create a table that is individually diverse in S and V and then to randomly permute the values for V within each $q^\star$-block to break the correlation between S and V. However, this may be unacceptable for researchers who need to study the interaction between S and V. Instead, we present a novel method that allows trade-off between information about V and information about the interactions between S and V.

We can only describe here the intuition behind our algorithm due to space constraints. Let $\phi = 1 - 1/\ell$ and let S, V_1, V_2, …, V_m be the sensitive attributes. First, ignore $V_1, \ldots, V_m$ and generate a table that is ℓ-diverse in S. Within each q^*-block, let t_s be the number of tuples whose value for S is s. For each $s' \in S$, choose $\lceil \phi t_{s'} \rceil$ tuples whose value for S is s' and suppress the values of $V_1, \ldots, V_n$ for those tuples. Intuitively, our suppression scheme works by creating uncertainty through missing values of $V_1, \ldots, V_n$ for each value of S.

Note that while simple, the method has several advantages:

- It preserves information about S. If we use this approach with k-anonymity, we still need a way to maintain diversity in S.
- It preserves partial information about the other sensitive attributes $V_1, \ldots, V_m$
- It preserves partial information about the interactions of all the sensitive attributes.
- The more data there is in the base table, the less information is removed due to suppression. The reason is that given enough data, we will still be able to make strong statistical statements about the distributions of the sensitive attributes.

4.4. Discussion

Recall that we started our journey into Section 4 motivated by the weaknesses of Bayes-optimal privacy. Let us now revisit these issues one by one.

- ℓ-Diversity no longer requires knowledge of the full distribution of the sensitive and nonsensitive attributes.
- ℓ-Diversity does not even require the data publisher to have as much information as the adversary. The parameter ℓ protects against more knowledgeable adversaries; the larger the value of ℓ, the more information is needed to rule out possible values of the sensitive attribute.
- Instance-level knowledge (Bob's son tells Alice that Bob does not have diabetes) is automatically covered. It is treated as just another way of ruling out possible values of the sensitive attribute.
- Different adversaries can have different background knowledge leading to different inferences. ℓ-Diversity simultaneously protects against all of them without the need for checking which inferences can be made with which levels of background knowledge.
- Jumping ahead: Unlike Bayes-optimal privacy, ℓ-diversity possesses a property called *monotonicity*. We will define this concept in Section 5, and we show how this property can be used to efficiently generate ℓ-diverse tables.

Overall, we believe that ℓ-diversity is practical, easy to understand, and addresses the shortcomings of background and homogeneity attacks on k-anonymity. Let us now see whether we can give efficient algorithms to implement ℓ-diversity.

5. Implementing Privacy Preserving Data Publishing

In this section we discuss how to build algorithms for privacy-preserving data publishing using domain generalization. Let us first review the search space for privacy-preserving data publishing using domain generalization [6, 16]. For ease of explanation, we will combine all the nonsensitive attributes into a single multi-dimensional attribute Q. On attribute Q, there is a user-defined generalization lattice. Formally, we define a generalization lattice to be a set of domains partially ordered by a generalization relation $<_G$ (as described in Section 2). The bottom element of this lattice is $\mathrm{domain}(Q)$ and the top element is the domain where each dimension of Q is generalized to a single value. Given a base table T, each domain $D_Q^\star$ in the lattice defines an anonymized table $T^\star$ which is constructed by replacing each tuple $t \in T$ by the tuple $t^\star \in T^\star$, such that the value $t^\star[Q] \in D_Q^\star$ is the generalization of the value $t[Q] \in \mathrm{domain}(Q)$. An algorithm for data publishing should find a point on the lattice such that the corresponding generalized table $T^\star$ preserves privacy and retains as much utility as possible. In the literature, the utility of a generalized table is usually defined as a distance metric on the lattice – the closer the lattice point is to the bottom, the larger the utility of the corresponding table $T^\star$. Hence, finding a a suitable anonymized table $T^\star$ is essentially a lattice search problem. There has been work on search strategies for k-anonymous tables that explore the lattice top-down [6] or bottom-up [16].

In general, searching the entire lattice is computationally intractable. However, lattice searches can be made efficient if there is a stopping condition of the form: if $T^\star$ preserves privacy then every generalization of $T^\star$ also preserves privacy [22, 16]. This is called the *monotonicity property*, and it has been used extensively in frequent itemset mining algorithms [4]. k-Anonymity satisfies the monotonicity property, and it is this property which guarantees the correctness of many algorithms [6, 16]. Thus, if we show that ℓ-diversity also possesses the monotonicity property, then we can re-use these efficient lattice search algorithms to find the ℓ-diverse table with optimal utility. Although more of theoretical interest, we can prove the following theorem that gives a computational reason why Bayes-optimal privacy does not lend itself to efficient algorithmic implementations.

Theorem 5.1 *Bayes-optimal privacy does not satisfy the monotonicity property.*

Due to space constraints, we refer the reader to [17] for a proof. However, we can prove that all variants of ℓ-diversity satisfy monotonicity.

Theorem 5.2 (Monotonicity of Entropy ℓ-diversity) *Entropy ℓ-diversity satisfies the monotonicity property: if a table $T^\star$ satisfies entropy ℓ-diversity, then any generalization $T^{\star\star}$ of $T^\star$ also satisfies entropy ℓ-diversity.*

Theorem 5.2 follows from the fact that entropy is a concave function. Thus if the $q^\star$-blocks $q_1^\star, \ldots, q_d^\star$ from table $T^\star$ are merged to form the $q^\star$-block $q^{\star\star}$ of table $T^{\star\star}$, then the $\mathrm{entropy}(q^{\star\star}) \geq \min_i(\mathrm{entropy}(q_i^\star))$. We can also prove the following theorem [17].

Theorem 5.3 (Monotonicity of NPD Recursive ℓ-diversity) *npd recursive (c_1, c_2, ℓ)-diversity satisfies the monotonicity property: if a table $T^\star$ satisfies npd recursive (c_1, c_2, ℓ)-diversity, then any generalization $T^{\star\star}$ of $T^\star$ also satisfies npd recursive (c_1, c_2, ℓ)-diversity.*

The main intuition behind the proof is as follows. Suppose the $q^\star$-blocks $q_1^\star, \ldots, q_d^\star$ from table $T^\star$ are merged to form the $q^\star$-block $q^{\star\star}$ of table $T^{\star\star}$. Let s_{y_i} is the most frequent sensitive value in the $q^\star$-block $q_i^\star$ of table $T^\star$. Let r_{y_i} be the frequency of s_{y_i} and $tail_{q_i^\star}(s_{y_i})$ the corresponding tail frequency (see definition 4.3). Similarly, define s_y, r_y

and $tail_{q^{\star\star}}(s_y)$ for the $q^\star$-block $q^{\star\star}$ of table $T^{\star\star}$. Then the result follows since

$$r_y \leq \sum_i r_{y_i}$$

$$\text{and } tail_{q^{\star\star}}(s_y) \geq \sum_i tail_{q_i^\star}(s_{y_i})$$

.

Thus to create an algorithm for ℓ-diversity, we simply take any algorithm for k-anonymity and make the following change: every time a table $T^\star$ is tested for k-anonymity, we check for ℓ-diversity instead. Since ℓ-diversity is a property that is local to each $q^\star$-block and since all ℓ-diversity tests are solely based on the counts of the sensitive values, this test can be performed very efficiently.

6. Experiments

For our experiments, we used an implementation based on Incognito [16] to generate k-anonymous and ℓ-diverse tables from the Adult Database at the UCIrvine Machine Learning Repository [20] and the Lands End Database. The Adult Database contains 45,222 tuples and contains US Census data. The Lands End Database has 4,591,581 tuples and consists of point-of-sale information. As in [16], we removed tuples with null values and adopted the same generalizations for the attributes. For immediate reference we give the details of the generalizations in Figure 4. The experiments were run on a machine with a 3GHz Intel Pentium 4 processor and 1 GB RAM. The operating system was Fedora Core 3 Linux distribution and the database manager was IBM DB2 v8.1. Due to lack of space, we only report a small subset of all the experiments we ran. [8]

Homogeneity Attack. To illustrate the *homogeneity attack* on a k-anonymized dataset, we used the Lands End Database with the first 5 attributes in Figure 4 as the quasi-identifier. We partitioned the Cost attribute into 147 buckets of size 100 and used it as the sensitive attribute. We produced all 3 minimal[9] 3-anonymous tables and found that 2 of them were vulnerable to the homogeneity attack, with more than 1,000 tuples being affected. Surprisingly, the average size of the affected groups was larger than 100 in both tables even though the sensitive attribute had a large domain. The table that was not vulnerable to the attack was entropy 2.61-diverse.

For the Adult Database, we used the first 5 attributes in Figure 4 as the quasi-identifier. With Occupation as the sensitive attribute, there were 12 minimal 6-anonymous tables and one of them was vulnerable to the homogeneity attack. With Salary Class as the sensitive attribute there were 9 minimal 6-anonymous tables and 8 of them were vulnerable. The other table was recursive (6,2)-diverse. The reason for these results is that Salary Class is a binary attribute with one value occurring 4 times as frequently as the other.

Performance. We compare the running times of entropy ℓ-diversity and k-anonymity in Figures 5 and 6. The sensitive attributes are Occupation for the Adults Database and Cost for Lands End. We measured the time taken to return all 6-anonymous and 6-diverse tables as the quasi-identifier size[10] was varied from 3 up to 8. The running times were comparable for both datasets, but we found that sometimes the running time for ℓ-diversity was faster. This happens since the algorithm prunes part of the lattice earlier than k-anonymity.

Utility. We measured the utility of ℓ-diverse and k-anonymous tables using three metrics: the generalization height metric [21, 16], average $q^\star$-block size, and the *discernibility* metric [6], which equals the sum of squares of the $q^\star$-block sizes when no tuples are suppressed.

We first looked at generalization height (the leftmost graph in Figure 7). This graph shows the minimum generalization height over all k-anonymous and ℓ-diverse tables for different values of k and $l = k$. As the graph shows, diversity does not require us to go much higher in the lattice; the minimum generalization heights for the same value of k and l are identical or off by one. However, we found that the generalization height metric [21] was not good for utility since tables at small generalization height can still have very large group sizes. For example, when using 5 quasi-identifiers, full-domain generalization on the Adults database returned 4-anonymous tables with average group sizes larger than 1000. The reason for the large group sizes was data skew. For example, there were only 114 tuples having age within (81-90), while there were 12,291 tuples having age within (31-40). So generalizing age from ranges of length 5 to ranges of length 10 would create very large groups. Generalization hierarchies which are aware of data skew may yield more useful data, a promising avenue for future work, especially since there are recent algorithms that allow such adaptations [6].

So to study utility, we used a 5% Bernoulli subsample of the Adults database such that much of the skew was removed from the age attribute[11]. The two rightmost graphs in Figure 7 plot the cost metrics versus the parameters k and ℓ with Occupation as the sensitive attribute. In each of the tables, smaller cost represents higher utility. As expected

[8] Full experimental results can be found in our technical report [17], which contains full experiments with different sizes of quasi-identifiers on both datasets; the results are qualitatively similar.

[9] Minimal in the generalization lattice

[10] Quasi-identifier size j consisted of the first j attributes from Figure 4

[11] Most of the age values appeared in around 20 tuples each while a few values appeared in less than 10 tuples each.

Adults

	Attribute	Domain size	Generalizations type	Ht.
1	Age	74	ranges-5,10,20	4
2	Gender	2	Suppression	1
3	Race	5	Suppression	1
4	Marital Status	7	Taxonomy tree	2
5	Education	16	Taxonomy tree	3
6	Native Country	41	Taxonomy tree	2
7	Work Class	7	Taxonomy tree	2
8	Salary class	2	*Sensitive att.*	
9	Occupation	41	*Sensitive att.*	

Lands End

	Attribute	Domain size	Generalizations type	Ht.
1	Zipcode	31953	Round each digit	5
2	Order date	320	Taxonomy tree	3
3	Gender	2	Suppression	1
4	Style	1509	Suppression	1
5	Price	346	Round each digit	4
6	Quantity	1	Suppression	1
7	Shipment	2	Suppression	1
8	Cost	147	*Sensitive att.*	

Figure 4. Description of Adults and Lands End Databases

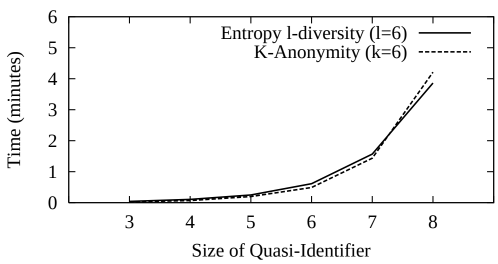

Figure 5. Adults Database

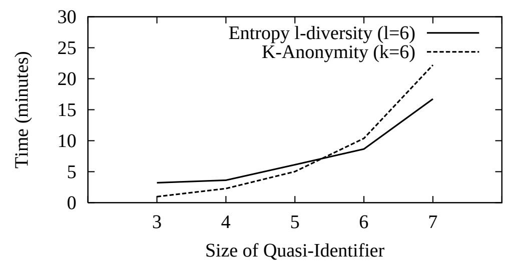

Figure 6. Lands End Database

the utility of the best ℓ-diverse table is worse than that of the best k-anonymous table since utility has to be traded off for privacy. However, in most cases the utility was not very different. It is interesting to note that Recursive $(3, \ell)$-diversity permits tables which have higher utility than Entropy ℓ-diversity – the discernibility costs are the same but the average group sizes are smaller for the recursive definition. The residual skew in the data set causes the entropy definition to perform worse.

7. Related Work

The problem of publishing public-use microdata has been extensively studied in both the statistics and computer science communities. The statistics literature, motivated by the need to publish census data, focuses on identifying and protecting the privacy of sensitive entries in contingency tables, or tables of counts which represent the complete cross-classification of the data. Two main approaches have been proposed for protecting the privacy of sensitive cells: *data swapping* and *data suppression*. The data swapping approach involves moving data entries from one cell to another in the contingency table in a manner that is consistent with the set of published marginals [9, 10, 13]. In the data suppression approach [8], cells with low counts are simply deleted, which in turn might lead to the deletion of additional cells. An alternate approach is to determine a *safety range* or *protection interval* for each cell [12], and publish only those marginals which ensure that the feasibility intervals (i.e. upper and lower bounds on the values a cell may take) contain the protection intervals for all the cell entries. The above techniques, however, do not provide a strong theoretical guarantee of the privacy ensured.

Computer science research also has tried to solve the data publishing problem. [24] propose a technique called k-anonymity which guarantees that every individual is hidden in a group of k with respect to the non-sensitive attributes. [18] show that the problem of k-anonymization by suppressing cells in the table is NP-hard. [3] propose approximation algorithms for the above. There has been a lot of study into creating efficient algorithms for k-anonymity using generalization and tuple suppression techniques [22, 2, 6, 16]. [7] propose a formal definition of privacy for published data based on the notion of *blending in a crowd*. However, since it is an inter-tuple distance centric measure of privacy, the privacy definition fails to capture scenarios where identification of even a single attribute may constitute a privacy breach.

Query answering techniques are very related to the data publishing approach, where instead of publishing the data, the database answers queries as long as the answers do not breach privacy. [19] attempt to characterize the set of views that can be published while keeping some query answer information-theoretically secret. The paper shows that the privacy required is too strong and most interesting queries like aggregates are not allowed to be published. Related techniques in the statistical database literature (see [1] for a survey), especially auditing [15] and output perturbation

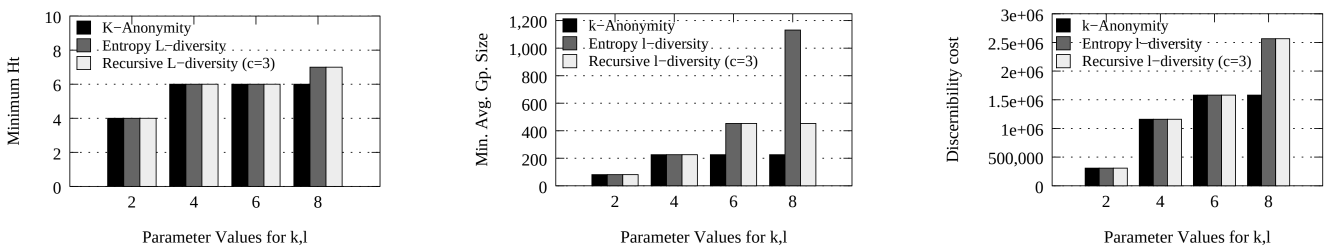

Figure 7. Adults Database. Q = $\{$age, gender, race, marital_status$\}$

[11], require maintaining state about the previous queries, while data publishing does not need to maintain any state of the queries asked. The reader is referred to [17] for a more extensive survey of related work.

8. Conclusions and Future Work

In this paper we have shown that a k-anonymized dataset permits strong attacks due to lack of diversity in the sensitive attributes. We have introduced ℓ-diversity, a framework that gives stronger privacy guarantees.

There are several avenues for future work. First, we want to extend our initial ideas for handling multiple sensitive attributes, and we want to develop methods for continuous sensitive attributes. Second, although privacy and utility are duals of each other, privacy has received much more attention than the utility of a published table. As a result, the concept of utility is not well-understood.

Acknowledgments. We thank Joe Halpern for an insightful discussion on the proposed privacy model and Kristen LeFevre for the Incognito source code.

References

[1] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 1989.

[2] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In *EDBT*, 2004.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity: Algorithms and hardness. Technical Report, Stanford University, 2004.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *VLDB*, 1994.

[5] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. *A. I.*, 1996.

[6] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.

[7] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In *TCC*, 2005.

[8] L. H. Cox. Suppression, methodology and statistical disclosure control. *J. of American Statistical Association*, 75, 1980.

[9] T. Dalenius and S. Reiss. Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 1982.

[10] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 1, 1998.

[11] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, 2003.

[12] A. Dobra. *Statistical Tools for Disclosure Limitation in Multiway Contingency Tables*. PhD thesis, CMU, 2002.

[13] G. T. Duncan and S. E. Feinberg. Obtaining information while preserving privacy: A markov perturbation method for tabular data. In *Joint Statistical Meetings*, 1997.

[14] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, 2003.

[15] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. In *PODS*, 2005.

[16] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. In *SIGMOD*, 2005.

[17] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. Technical Report, Cornell University, 2005.

[18] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.

[19] G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. In *SIGMOD*, 2004.

[20] U.C.Irvine Machine Learning Repository. http://www.ics.uci.edu/ mlearn/mlrepository.html.

[21] P. Samarati. Protecting respondents' identities in microdata release. In *IEEE TKDE*, 2001.

[22] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. T R, CMU, SRI, 1998.

[23] L. Sweeney. Uniqueness of simple demographics in the u.s. population. Technical Report, CMU, 2000.

[24] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002.

[25] S. Zhong, Z. Yang, and R. N. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS*, 2005.