Injecting Utility into Anonymized Datasets

Daniel Kifer
Department of Computer Science
Cornell University
dkifer@cs.cornell.edu

Johannes Gehrke
Department of Computer Science
Cornell University
johannes@cs.cornell.edu

ABSTRACT

Limiting disclosure in data publishing requires a careful balance between privacy and utility. Information about individuals must not be revealed, but a dataset should still be useful for studying the characteristics of a population. Privacy requirements such as k-anonymity and ℓ-diversity are designed to thwart attacks that attempt to identify individuals in the data and to discover their sensitive information. On the other hand, the utility of such data has not been well-studied.

In this paper we will discuss the shortcomings of current heuristic approaches to measuring utility and we will introduce a formal approach to measuring utility. Armed with this utility metric, we will show how to inject additional information into k-anonymous and ℓ-diverse tables. This information has an intuitive semantic meaning, it increases the utility beyond what is possible in the original k-anonymity and ℓ-diversity frameworks, and it maintains the privacy guarantees of k-anonymity and ℓ-diversity.

1. INTRODUCTION

Mining data sets that contain information about individuals in a population is a great way of learning about properties of that population. Applications include studying the effects of treatments on disease, tracking disease outbreaks, and building economic models (from census data). Aside from this "good" information, such data sets also contain sensitive information: the disease of an individual, the salary, etc. Because of this, the goal of privacy-preserving data publishing is to maximize the "good utility" while limiting the ability of an adversary to identify specific individuals and learn their sensitive information from the data set.

In terms of privacy, k-anonymity [33] and ℓ-diversity [23] provide strong guarantees on the confidentiality of individuals in the data. Both concepts rely on *generalizations* to preserve privacy: attributes are replaced with less specific information (for example, "state" may be replaced with "region" and "age" may be replaced with "age range"). However, the utility of these "anonymized" data sets has received much less study. Many heuristics for measuring utility have been proposed, but to the best of our knowledge there are no formal measures of the utility of an anonymized dataset.

To make matters worse, the curse of dimensionality that haunts the statistics and machine learning communities [8] also has an adverse effect on anonymized data [2]. Experimental evidence [2] suggests that many attributes in the data need to be suppressed in order to guarantee privacy. This effect was also present in the experiments we conducted for this paper: while generating anonymized data from the Adult dataset in the UCI Machine Learning Repository [29], attributes such as "ethnicity" had to be completely suppressed. Clearly this is bad for utility no matter what measure is used.

One way to ameliorate this curse of dimensionality is to publish additional information, such as a table containing just the ethnicities and their frequencies in the original table. Clearly we can generalize this to publishing marginals (or, equivalently, duplicate-preserving projections, or views) of the original table. The marginals themselves can be anonymized (i.e., generalization can be performed on these marginals) and the generalizations used on the marginals need not be the same. This is precisely the approach that we are proposing.

Consider Figure 1: we begin with a base table, Figure 1(a), and then use an anonymization algorithm to create the $(2, 3)$-diverse table in Figure 1(b) (the precise definition of ℓ-diversity will be given in Section 2). The current approach in the literature is to stop at this point and publish this table. Note that there is additional information we can publish in terms of anonymized marginals: more detailed age information (Figure 1(c)), detailed zip-code/disease information (Figure 1(d)), more detailed joint information about age and zip code (Figure 1(e)), etc. Given so many choices, which anonymized marginals should we publish? Clearly we cannot publish them all, because if an attacker knows that a person who lives in 14850 and is under 40 is in the original table (from Figure 1(e) or from background knowledge), then the attacker can join Figures 1(b) and 1(d) to deduce that the person has measles.

In order to answer the question of which anonymized marginals to publish, we first introduce in Section 3 a new way of quantifying the amount of information (utility) contained in anonymized data, and we discuss how this information should be combined to approximate the original data. Using these ideas, we will take accepted single-table privacy definitions and extend them so that they apply to collections of anonymized marginals in Section 4. The technical challenges are that combining information from marginals and computing the utility require slow iterative algorithms, and checking for privacy is NP-hard. We will deal with those issues in Sections 5, 6, and 7. In Section 5.1 we will review the notion of *decomposability* from the graphical models literature. This will let us impose a structure on anonymized marginals that will simplify utility calculations and will let us develop tractable algorithms for checking for privacy. As a result, a data publisher will have the abil-

SIGMOD 2006, June 27–29, 2006, Chicago, Illinois, USA.
Copyright 2006 ACM 1-59593-256-9/06/0006 ...$5.00.

Zip Code	Condition	Age
14850	Measles	33
14853	Allergy	26
14853	Gout	22
14853	Cancer	32
14850	Flu	48
14850	Heart	47
14850	Flu	46
14853	Cancer	53
14853	Heart	51
13063	Flu	24
13063	Cancer	38
13068	Cancer	38
13068	Heart	30

(a) Original table

Zip Code	Condition	Age
1485*	Allergy	≤ 40
	Gout	≤ 40
	Measles	≤ 40
	Cancer	≤ 40
1485*	Heart	> 40
	Flu	> 40
	Heart	> 40
	Flu	> 40
	Cancer	> 40
1306*	Heart	≤ 40
	Flu	≤ 40
	Cancer	≤ 40
	Cancer	≤ 40

(b) $(2, 3)$-diverse table

Age	Count
$[21-25]$	2
$[26-30]$	2
$[31-35]$	2
$[36-40]$	2
$[41-45]$	0
$[46-50]$	3
$[51-55]$	2

(c) Age Marginal

Zip Code	Condition	Count
14850	Flu	2
	Heart	1
	Measles	1
14853	Cancer	2
	Heart	1
	Allergy	1
	Gout	1
13063	Flu	1
	Cancer	1
13068	Cancer	1
	Heart	1

(d) Zip/Condition Marginal

Zip Code	Age	Count
14850	[41-50]	3
	[31-40]	1
14853	[51-60]	2
	[21-30]	2
	[31-40]	1
13063	[21-30]	1
	[31-40]	1
13068	[31-40]	1
	[21-30]	1

(e) Zip/Age Marginal

Figure 1: Anonymized Marginals

ity to examine many different collections of anonymized marginals before deciding which collection to publish. In Section 5.2 we will briefly review the theory of log-linear models and show that publishing marginals can be viewed in a different way: it can be viewed as selecting the set of conditional independence relations that best describe the original table. The theory in Section 5 is designed for ordinary marginals. In Section 6 we will extend this theory to anonymized marginals and discuss how to traverse the search space of anonymized marginals. Then we will describe our algorithms for checking for privacy in Section 7. In Section 8 we will present our experiments that show that the strategy of publishing a collection of anonymized marginals indeed provides a dramatic improvement in utility over the strategy of publishing a single anonymized table.

In summary, our contributions include formalizing the notion of utility for k-anonymous and ℓ-diverse tables, extending these definitions to anonymized marginals, extending results from graphical and log-linear models to anonymized marginals (including results on search-space traversal), and providing algorithms for checking for privacy.

2. PRELIMINARIES

In this section we will introduce the notation and basic definitions that will be used later on. First we will introduce basic notation and definitions related to privacy, and then we will introduce notation for dealing with tabular data.

2.1 Privacy Basics

Let $D = \{t_1, \dots, t_m\}$ be a database of tuples where each tuple has $d = d_1 + d_2$ attributes: $t_i = \{t_i.R_1, \dots, t_i.R_{d_1}, t_i.S_1, \dots, t_i.S_{d_2}\}$. We will slightly abuse notation and use R_i to also refer to the domain of attribute $t.R_i$ and S_i to refer to the domain of attribute $t.S_i$. The attributes $R_1, \dots, R_{d_1}$ are called the *nonsensitive* attributes. This is because they are either public knowledge or because they are available from some external data set. For example, date of birth, gender, and zip code are available from voter registration records and so are considered nonsensitive. The attributes $S_1, \dots, S_{d_2}$ are the *sensitive* attributes. For example, disease (in a hospital data set) would be considered sensitive. Since the dataset is given in tabular form, we will use the terms “dataset” and “table” interchangeably.

The goal of privacy-preserving data publishing is to make it difficult for an attacker to determine that someone is in the dataset, and to make it difficult to determine the values of the sensitive attributes of individuals that are known to be in the table. One glaring example of these goals not being met is described in [33]: a (supposedly anonymized) medical dataset of Massachusetts state employees was joined to voter registration records using the attributes birth date, gender, and zip code. Since the governor of Massachusetts had a unique combination of those attributes, his medical records were easily identified. Generally, a set of attributes (like the set {birth date, gender, zip code} in the previous example) that acts almost like a key and can be used to uniquely identify some individuals is known as a quasi-identifier [33]:

DEFINITION 2.1 (QUASI-IDENTIFIER). *A set of nonsensitive attributes $\{R_1, \dots, R_n\}$ in a database D is called a* quasi-identifier *if this set can be used to identify at least one individual from a given population by linking those attributes to external data sets.*

Without loss of generality, we assume that all the set of all nonsensitive attributes forms the quasi-identifier. To prevent linking attacks that use the quasi-identifier, it is common to use *generalizations*:

DEFINITION 2.2 (GENERALIZATION). *Let V be the domain of an attribute $t.V$. A* generalization *W of V is a new domain formed by partitioning V into disjoint buckets and identifying all the points in a bucket with one value in W. A* generalization map *is a function $\phi : V \to W$ such that $\phi(v)$ corresponds to the bucket that contains v.*

As an example, consider the integer-valued attribute Age. One generalization of Age is the set of intervals $A' = \{[0-5], [6-10], [11-15], \dots\}$. The generalization map from Age to A' replaces each integer with an interval. A' itself can also be generalized. One such generalization of A' is $A'' = \{[0-10], [11-20], \dots\}$. Note that A'' is also a generalization of Age (generalizations are transitive). Thus we can define a partial order on domains: $A \prec B$ if and only if B is a generalization of A (note that $A \prec A$ is always true). To generalize a table, we choose a generalization for each attribute and apply the appropriate generalization maps to the attributes of all tuples t. We can perform generalizations on the nonsensitive attributes (quasi-identifier) to make linking attacks difficult. This is the goal of k-anonymity [33].

DEFINITION 2.3 (k-ANONYMITY). *A table D satisfies k-anonymity if for every tuple $t \in D$ there exist at least $k-1$ other tuples that have the same values as t for every quasi-identifier.*

Given the quasi-identifier {zip code, age}, the table in Figure 1(b) is 4-anonymous. Generalizations partition the tuples into *anonymized groups*:

DEFINITION 2.4 (ANONYMIZED GROUP). *An anonymized group is a (set-wise) maximal set of tuples that have the same (generalized) value for each nonsensitive attribute.*

Note that k-anonymity says nothing about the sensitive attributes. In particular, it does not prevent all tuples in an anonymized group from having the same value for some sensitive attribute (thus benefiting an attacker who knows some of the individuals that are in that anonymized group). The concept of ℓ-diversity is designed to guard against this. Machanavajjhala et al [23] give several alternative formulations of ℓ-diversity. Any of them can be used here, but for concreteness, we will use the following:

DEFINITION 2.5 ($(c, \ell)-$DIVERSITY). *Let $c > 0$ be a constant and let q be an anonymized group. Let S be a sensitive attribute, let $s_1, \ldots, s_m$ be the values of S that appear in the group q and let $r_1, r_2, \ldots, r_m$ be the their corresponding frequency counts in q. Let $r_{(1)}, r_{(2)}, \ldots, r_{(m)}$ be those counts sorted in non-increasing order. We say that the anonymized group q satisfies (c, ℓ)-diversity with respect to a sensitive attribute S if $r_{(1)} \leq c \sum\limits_{i=\ell}^{m} r_{(i)}$. The set $\{s_{(1)}\}$ is the **head** and the set $\{s_{(\ell)}, \ldots, s_{(m)}\}$ is called the **tail**.*

Intuitively, ℓ-diversity means that an adversary needs $\ell - 1$ pieces of background knowledge to eliminate $\ell - 1$ possible values of a sensitive attribute in order to breach privacy. Like k-anonymity, ℓ-diversity can be achieved through generalizations. Throughout the paper we will assume there is only one sensitive attribute. This is only necessary for clarity. The extension to multiple sensitive attributes is straightforward and is done in the same way as in [23].

Note that there are other ways of sanitizing data: tuple suppression [31], adding random noise [5, 16, 1], and swapping attributes between tuples [12]. We will restrict our attention to generalizations because they are the most *faithful* to the data: any fact in a generalized table is true of the original table. For example, if the value for the Age attribute has been generalized to $[25 - 30]$ then we know for certain that the age is within that interval. Furthermore, we could determine exactly how many individuals in the original table were between 25 and 30 years old. Faithfulness is an important concept because it may be used by future data mining algorithms to give quality guarantees on their results.

2.2 Tabular Data

Recall that our database $D = \{t_1, \ldots, t_m\}$ is a set of tuples where each tuple has $d = d_1 + d_2$ attributes: $t_i = \{t_i.R_1, \ldots, t_i.R_{d_1}, t_i.S_1, \ldots, t_i.S_{d_2}\}$. The domain of D, $\text{Domain}(D)$, is the crossproduct $R_1 \times \cdots \times R_{d_1} \times S_1 \times \cdots \times S_{d_2}$. The nonsensitive domain, $\text{NonSenDomain}(D)$, is the domain of the nonsensitive attributes: $R_1 \times \cdots \times R_{d_1}$; the sensitive domain, $\text{SenDomain}(D)$, is the domain of the sensitive attributes: $S_1 \times \cdots \times S_{d_2}$. When all the attributes are discrete (or have been bucketized), it is convenient to think of the data set as a contingency table $T^{(D)}$: for any $t \in \text{Domain}(D)$, $T^{(D)}(t)$ is the number of times t appears in D. Whenever it is unambiguous, we will drop the explicit notational dependency on D and refer to the corresponding contingency table as T. Let $C \subseteq \text{Domain}(D)$. $T(C)$ is the number of tuples in D that also belong to the subset of the domain represented by C (we will use lower-case letters to denote tuples and upper-case letters to denote sets). Thus we can think of T as a function defined on the powerset of $\text{Domain}(D)$.

We will also be concerned with marginals of the contingency table $T^{(D)}$. Let $\mathcal{A}$ be a set of attributes. Then $T^{(D)}_{\mathcal{A}}$ is the marginal contingency table that we get by projecting out all attributes of D that are not in $\mathcal{A}$ (while preserving duplicates). For example, if $\mathcal{A} = \{R_1, R_2\}$ and $t = (r_1, r_2)$ then $T^{(D)}_{\mathcal{A}}(t)$ is the number of tuples in D whose value for R_1 is r_1 and whose value for R_2 is r_2. Note that the original contingency table T is itself a marginal, and that the marginal $T_\emptyset$ can be thought of as the function that always returns $|D|$ (the number of tuples in the database). We will use the term *anonymized table* to refer to a table that has been altered through the use of generalizations. In particular, when we apply generalizations to a marginal contingency table, the result is an *anonymized marginal*.

3. UTILITY MEASURES

In this section we review current utility measures for anonymized datasets (Section 3.1) and then discuss a more formal measure of utility (Section 3.2) which has connections to maximum entropy and conditional independence, and which will be suitable for anonymized marginals.

3.1 Current Measures of Utility

One of the earliest utility metrics is *generalization height* [30]. Generalization height is the total number of generalization steps that have been performed on the original data set. The intuition behind it is that a generalization step represents a loss of information, so one should use as few generalization steps as possible. As noted in [23], the problem with this approach is that not all generalization steps are created equal: a generalization step on one attribute may put many more tuples into an anonymized group than a generalization step on another attribute.

Two similar metrics that take anonymized group size into account are the average size of anonymized groups [23] and *discernibility* [6]. Discernibility assigns a cost to each tuple based on how many other tuples are indistinguishable from it. If a tuple is suppressed, its cost is $|D|$ (the number of tuples in the original data). If a tuple is not suppressed, its cost is the number of tuples in its anonymized group. Thus the discernibility is the sum of the squares of the anonymized group sizes plus $|D|$ times the number of suppressed tuples. While appealing, neither of these two metrics takes the data distribution into account. An anonymized group where the original attributes were uniformly distributed represents less information loss than an anonymized group whose original attributes were skewed. For example, suppose there are 10 people who have the same value for every attribute except age, and their ages are between 20 and 30. In a sense, our best guess (using maximum entropy, or the principle of indifference) is that each age is equally likely. Intuitively, had the original ages been $21, 22, \ldots, 30$, this would not have been as much an information loss as the case where 5 people were 21 years old and the other 5 were 29. Iyengar [17] presents a related loss metric which considers how many elements in the original domain have been grouped together. Since it also ignores the tuple distribution, it has the same shortcoming.

Two utility metrics that take distribution into account are the *classification metric* [17] and *information-gain-privacy-loss ratio* [34]. The classification metric is appropriate when one wants to train a classifier over the anonymized data. Thus one attribute is treated as a class label. The classification metric assigns a penalty of 1 to every tuple that is suppressed. If a tuple t is not suppressed, we look at the majority class label in its anonymized group. If the class label for t is different than the majority class label, t is assigned a penalty of 1. The classification metric is then the sum of all penalties. The classification metric is an appealing measure of utility because it considers a possible use for the data. However, it is not clear what we should do if we want to build classifiers for several different attributes. The information-gain-privacy-loss ratio is also designed for the purposes of classification. It is a local heuristic in the sense that it is used to determine the next generalization step (it is similar to the way information gain is used to choose the next split point in a decision tree). Here as well it is not clear what we should do if we want to build classifiers for several different attributes. Furthermore, because it is a local heuristic, it is difficult to compare the utilities of two different anonymized tables.

3.2 A Formal Utility Measure

One of the main goals of data mining and statistics is to make statements about the probability distribution that generated the data – this is certainly true of classification, parameter estimation, hypothesis testing, and regression. In this spirit, we view the data as an iid (identically and independently distributed) sample generated from some multidimensional distribution F. Here we shall assume that all attributes are discrete. Note that if we have a continuous domain, we can bucketize it and treat the collection of buckets as a discrete domain; other ways of dealing with continuous attributes are the subject of future work. With this simplification, the data follows a multinomial distribution. Suppose the tuples in our dataset have discrete-valued attributes $U_1, \dots, U_n$. Then we can estimate F with the *empirical distribution* $\hat{F}$. $\hat{F}(u_1, \dots, u_n)$ is an estimate of the probability $P(t.U_1 = u_1, \dots, t.U_n = u_n)$ and is defined as the fraction of tuples in the database that satisfy this constraint (i.e., $t.U_1 = u_1, \dots, t.U_n = u_n$).

Now that we have given a probabilistic interpretation to the original data, we will proceed to do the same for the anonymized data. Suppose we are given a collection of anonymized marginals that were derived from the same table. We can view these marginals as constraints. Figure 1 shows a set of anonymized marginals that correspond to the table in Figure 1(a) that has 13 tuples. These anonymized marginals impose constraints such as: 23% of the tuples have age between 46 and 50 (Figure 1(c)); 38.5% of the tuples are in zip code 14853 (Figure 1(d)); 15.4% of the tuples are in zip code 1306*, have cancer, and have age at most 40 (Figure 1(b)). Thus given anonymized marginals, we can compute the maximum entropy probability distribution that corresponds to these constraints. We will see in Section 5.2 that this maximum entropy distribution is exactly the same as the maximum likelihood distribution for a multinomial model that satisfies certain intuitive conditional independence requirements.

We now have a probability distribution $\hat{F_1}$ associated with the original data, and a probability distribution $\hat{F_2}$ associated with the released anonymized marginals. The next step is to compare them. Let $x_1, \dots, x_N$ be the elements of the multidimensional domain for our data. Let $p_i^{(1)}$ be the probability of x_i according to $\hat{F_1}$ and let $p_i^{(2)}$ be the probability according to $\hat{F_2}$. The Kullback-Leibler divergence (KL-divergence) between $\hat{F_1}$ and $\hat{F_2}$ is defined as:

$$\sum_i p_i^{(1)} \log \frac{p_i^{(1)}}{p_i^{(2)}}$$

The KL-divergence is minimized only when $\hat{F_1} = \hat{F_2}$. In Section 5.2, we shall see that the KL-divergence is equal to the difference in log-likelihood when we estimate the "true" distribution F with the original data and when we estimate the "true" distribution F with the anonymized marginals. Since our goal is to determine which anonymized marginals to release, we will be changing $\hat{F_2}$ but not $\hat{F_1}$. Thus minimizing the KL-divergence will be mathematically equivalent to maximizing $\sum_i p_i^{(1)} \log p_i^{(2)}$ which is -1 times the cross-entropy between $\hat{F_1}$ and $\hat{F_2}$. We will use the standard convention that $0 \log 0 = 0$ so that we only need to compute probabilities for cells that appear in the original table.

Note that this approach is similar in spirit to [4] and [5]: we are trying to reconstruct the original distribution as accurately as possible given anonymized (but unperturbed) data while [4] and [5] try to reconstruct the original distribution given a dataset with additive noise.

4. EXTENDING PRIVACY DEFINITIONS

The next step is to extend the privacy definitions for k-anonymity and ℓ-diversity from single-table anonymized data to collections of anonymized marginals. We can extend k-anonymity in two ways, reflecting the motivation provided by Sweeney [33]. Following [33], we can protect the anonymized marginals from being linked to external data by requiring every anonymized marginal to be k-anonymous.

DEFINITION 4.1 (k-LINK ANONYMITY). *A collection of anonymized marginals $M_1, \dots, M_r$ satisfies k-*link anonymity *if for all $i = 1 \dots r$ and for all $t \in \text{NonSenDomain}(M_i)$ either $M_i(t) = 0$ or $M_i(t) \geq k$.*

We must also be sure that an adversary cannot use combinatorial techniques to determine that a tuple with a certain value for its quasi-identifiers exists in the original table and that the number of such tuples is less than k.

DEFINITION 4.2 (k-COMBINATORIAL ANONYMITY). *Let $\mathcal{D}$ be the domain of the nonsensitive attributes. A collection of anonymized marginals $M_1, \dots, M_r$ satisfies k-*combinatorial anonymity *if for all $t \in \mathcal{D}$* ***one*** *of the following holds:*

1. *For all tables T consistent with the marginals $M_1, \dots, M_r$, $T(t) = 0$*
2. *There exists a table T consistent with the marginals $M_1, \dots, M_r$ such that $T(t) \geq k$.*

Our final privacy requirement is a straightforward generalization of ℓ-diversity. Using the anonymized marginals, the maximum entropy principle, and techniques we will discuss later, we can fill in (fractional) cell counts for the original table. Thus we directly apply the definition of ℓ-diversity [23] to the reconstructed table (i.e., for each combination of nonsensitive attributes, the sensitive attributes must have ℓ well-represented values).

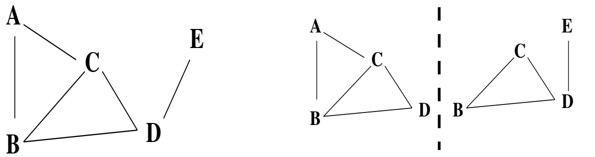
DEFINITION 4.3 (MAXENT ℓ-DIVERSITY). *The anonymized marginals $M_1, \dots, M_r$ satisfy maxent ℓ-diversity if the maximum entropy distribution that is consistent with $M_1, \dots, M_r$ satisfies ℓ-diversity.*

5. STATISTICAL MODELS

In this section we will discuss how to combine information from ordinary marginals to estimate the original data; in Section 6 we will show how to apply this theory to anonymized marginals. The estimator can be viewed as a maximum entropy distribution as well as a maximum likelihood estimator for multinomial models. This will give marginals interpretations as constraints and as statements about conditional independence. In general, computing the maximum entropy distribution requires iterative algorithms [7, 24, 10]. However, with some additional restrictions on the allowable marginals, there is a closed-form solution [20]. This lets the data publisher examine many different collections of marginals before deciding which ones to publish. We will discuss how to compute the maximum entropy distribution in Section 5.1 and we will discuss the connection to maximum likelihood in Section 5.2.

5.1 Decomposable Graphical Models

Given a set of marginals $M_1, \dots, M_r$, build an interaction graph in the following way: the vertices of the graph are the attributes that appear in any marginal. For any two vertices A and B, draw an undirected edge between A and B if attributes A and B appear together in some marginal. Figure 2(a) shows an interaction graph



(a) Interaction Graph (b) Components under a decomposition

Figure 2: Interaction Graph and Decomposition for Marginals ABC**,** BCD**,** DE

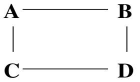


Figure 3: Smallest Non-decomposable graph

for the three marginals whose attributes are ABC, BCD, and DE Our first requirement is that the interaction graph must be *triangulated*:

Definition 5.1 (Triangulated Graph). *An undirected graph is triangulated if for every cycle of length 4 or more, there exists an edge not in the cycle that connects two vertices in the cycle.*

Intuitively this means that every cycle that has more than 3 nodes has a "shortcut". Undirected triangulated graphs are equivalent to undirected *decomposable* graphs [20]. To explain this concept, we need the following definitions:

Definition 5.2 (Separator [20]). *Let $G = (V, E)$ be an undirected graph and let $\mathcal{A}, \mathcal{B} \subset V$ be disjoint sets of vertices. The set $\mathcal{C} \subseteq V \setminus \{\mathcal{A} \cup \mathcal{B}\}$ separates (is a separator of) $\mathcal{A}$ and $\mathcal{B}$ if every path from $\mathcal{A}$ to $\mathcal{B}$ contains a node of $\mathcal{C}$.*

Definition 5.3 (Decomposition [20]). *Let $G = (V, E)$ be an undirected graph and let $\mathcal{A}, \mathcal{B}, \mathcal{C} \subset V$ be disjoint sets of vertices such that $V = \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$. Then $\mathcal{A}, \mathcal{B}, \mathcal{C}$ form a decomposition of G if $\mathcal{C}$ separates $\mathcal{A}$ and $\mathcal{B}$, and $\mathcal{C}$ is complete (every two vertices in $\mathcal{C}$ have an edge between them).*

A decomposition $(\mathcal{A}, \mathcal{B}, \mathcal{C})$ splits a graph $G = (V, E)$ into two components. The first component is the subgraph of G that contains the vertices $\mathcal{A} \cup \mathcal{C}$ and the second component contains the subgraph induced by the vertices $\mathcal{B} \cup \mathcal{C}$. If both $\mathcal{A}$ and $\mathcal{B}$ are nonempty then the decomposition is *proper* (each component is strictly smaller that the original graph G).

Definition 5.4 (Decomposable). *A graph is decomposable if it is complete or if it has a proper decomposition where each component is decomposable.*

Theorem 5.1 (Triangulated Graphs [20]). *An undirected graph is decomposable if and only if it is triangulated.*

Figure 2(b) shows the components of the graph in Figure 2(a) under the decomposition $\{A\}, \{E\}, \{B, C, D\}$ (where $\{B, C, D\}$ is the separator). Because of the equivalence between decomposable and triangulated graphs, it is easy to check whether a graph is decomposable. Figure 3 is the smallest non-decomposable graph (if it had an edge A and B or C and D then it would be triangulated/decomposable).

Definition 5.5 (Graphical Marginals). *Let $M_1, \ldots, M_r$ be a collection of marginals and let G be the corresponding interaction graph. The collection $M_1, \ldots, M_r$ is graphical if the marginals contain all of the maximal cliques of the corresponding interaction graph G.*

In other words, if $\mathcal{E}_i$ is a maximal clique of the interaction graph, there is a marginal M_j such that all of the attributes that correspond to vertices in $\mathcal{E}_i$ are the attributes of M_j.

Definition 5.6 (Decomposable Marginals). *A set of marginals $M_1, \ldots, M_r$ is decomposable if it is graphical and the corresponding interaction graph G is decomposable.*

The interaction graph in Figure 2(a) is generated by the marginals with attributes ABC, BCD, and DE. It is also generated by ABC, BD, CD, and DE. While the interaction graph is decomposable, the maximal cliques correspond only to ABC, BCD, and DE. Thus the set of marginals $\{ABC, BCD, DE\}$ is decomposable while $\{ABC, BD, CD, DE\}$ is not.

Decomposability is important to us because the maximal cliques can be ordered in a *perfect sequence* [20] which can be used to compute the maximum entropy distribution:

Definition 5.7 (Perfect Sequence [20]). *Let G be a graph and let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_p$ be a sequence of complete subgraphs of G that includes all the maximal cliques of G. The sequence $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_p$ is perfect if for $i = 2 \ldots p$, the set $S_i = \mathcal{E}_i \cap (\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_{i-1})$ is complete and there exists a $j < i$ such that $S_i \subseteq \mathcal{E}_j$.*

Returning to our example in Figure 2(a), we see that $\{A, B, C\}$, $\{B, C, D\}$, $\{D, E\}$ is a perfect sequence. It is a fundamental fact that every decomposable graph has a perfect sequence [20]. Also, it is easy to see that S_i separates $(\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_{i-1}) \setminus S_i$ and $\mathcal{E}_i \setminus S_i$. The sets $\mathcal{E}_1 \ldots \mathcal{E}_p$ and $S_2, \ldots, S_p$ (note the separators are numbered starting from 2) can be used to compute the maximum entropy (maxent) distribution. Let $\mathcal{E}_1, \ldots, \mathcal{E}_p$ be sets of attributes that correspond to a decomposable graphical model and that have already been arranged in a perfect sequence. Let $S_2, \ldots, S_p$ be the corresponding separators as in Definition 5.7. For each i, let $T_{\mathcal{E}_i}$ be the marginal of the base table T corresponding to attribute set $\mathcal{E}_i$ and T_{S_i} be the marginal corresponding to the attribute set S_i. For each $t \in \text{Domain}(T)$ let $t_{\mathcal{E}_i}$ be the projection of t onto the attributes in $\mathcal{E}_i$ and similarly for t_{S_i}. The maximum entropy probability associated with t (which is also the maximum likelihood estimate associated with log-linear models, which will be briefly discussed in Section 5.2) is [24]:

$$\frac{1}{|T|} \frac{\prod_{i=1}^{p} T_{\mathcal{E}_i}(t_{\mathcal{E}_i})}{\prod_{j=2}^{p} T_{S_j}(t_{S_j})} \tag{1}$$

and, given a table of size $|T|$, the expected count of tuples in the cell corresponding to t is therefore:

$$\frac{\prod_{i=1}^{p} T_{\mathcal{E}_i}(t_{\mathcal{E}_i})}{\prod_{j=2}^{p} T_{S_j}(t_{S_j})} \tag{2}$$

By definition, for any j, there is an i such that $S_j \subseteq \mathcal{E}_i$ so that T_{S_j} can be computed from $T_{\mathcal{E}_i}$ and so the maximum entropy distribution can be computed from the marginals corresponding to $\mathcal{E}_1, \ldots, \mathcal{E}_p$ with little effort once they are ordered in a perfect sequence (an algorithm for such an ordering is provided in [20]).

5.2 Log-linear models: the connection to maximum likelihood

An interaction graph (e.g., Figure 2(a)) also has an interpretation in terms of conditional independence [10]. Let $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ be disjoint sets of attributes (vertices) in the interaction graph and suppose that every path from a vertex in $\mathcal{A}$ to a vertex in $\mathcal{B}$ contains a vertex in $\mathcal{C}$. Intuitively, the effects of $\mathcal{A}$ and $\mathcal{B}$ on each other are blocked by the separator $\mathcal{C}$ and thus we have the interpretation that $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{C}$. Using the notation that $t_\mathcal{A}$, $t_\mathcal{B}$, $t_{\mathcal{A}\cup\mathcal{B}}$ and $t_\mathcal{C}$ are the projections of t onto the attributes in $\mathcal{A}$, $\mathcal{B}$, $\mathcal{A}\cup\mathcal{B}$, $\mathcal{C}$, respectively, we can write the conditional independence restriction mathematically. For $a \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}$:

$$\begin{aligned} &P(t_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}} = (a, b, c)) \\ &\quad = P(t_{\mathcal{A}\cup\mathcal{B}} = (a, b) | t_\mathcal{C} = c) \cdot P(t_\mathcal{C} = c) \\ &\quad = P(t_\mathcal{A} = a | t_\mathcal{C} = c) \cdot P(t_\mathcal{B} = b | t_\mathcal{C} = c) \cdot P(t_\mathcal{C} = c) \\ &\quad = \frac{P(t_{\mathcal{A}\cup\mathcal{C}} = (a, c)) \cdot P(t_{\mathcal{B}\cup\mathcal{C}} = (b, c))}{P(t_\mathcal{C} = c)} \end{aligned} \tag{3}$$

To see the relation to Equation 1, suppose we were given two marginals $T_{\mathcal{A}\cup\mathcal{C}}$ and $T_{\mathcal{B}\cup\mathcal{C}}$. We can compute $T_\mathcal{C}$ from either $T_{\mathcal{A}\cup\mathcal{C}}$ or $T_{\mathcal{B}\cup\mathcal{C}}$. We can estimate $P(t_{\mathcal{A}\cup\mathcal{C}})$ using the count data $\frac{T_{\mathcal{A}\cup\mathcal{C}}(t_{\mathcal{A}\cup\mathcal{C}})}{|T|}$ (and similarly for $t_\mathcal{C}$ and $t_{\mathcal{B}\cup\mathcal{C}}$). By substituting these estimates in the right side of Equation 3 we get

$$P(t_{\mathcal{A}\cup\mathcal{B}\cup\mathcal{C}} = (a, b, c)) = \frac{(T_{\mathcal{A}\cup\mathcal{C}}(t_{\mathcal{A}\cup\mathcal{C}})/|T|)\,(T_{\mathcal{B}\cup\mathcal{C}}(t_{\mathcal{B}\cup\mathcal{C}})/|T|)}{T_\mathcal{C}(t_\mathcal{C})/|T|}$$

which is the same as the maximum entropy estimate (Equation 1).

To formally relate maximum entropy and maximum likelihood in multinomial models, we must first discuss log-linear models – popular statistical tools for analyzing contingency tables [10].

Let t be a cell in a contingency table, and let $q(t)$ be the expected cell count under a multinomial model. The goal of statistical analysis of contingency tables is to learn about some dependencies between the $q(t)$; in particular, the goal is to determine how $q(t)$ is affected by the various attribute dimensions of the contingency table[1]. In this analysis, one builds a linear model for predicting $\log q(t)$ (there are technical reasons for modeling $\log q(t)$ rather than $q(t)$). To see how this works, suppose our table has 2 attributes A and B. For a cell $t \in A \times B$, let $q(t)$ be the expected cell count for t. For this scenario, the *saturated* log-linear model is:

$$\log q(a, b) = u + u_A(a) + u_B(b) + u_{AB}(a, b) \tag{4}$$

where u represents a baseline cell occupancy based on no interactions, u_A represents the effect of attribute A on cell occupancy (beyond the effects of the baseline), u_B represents the effects of attribute B on cell occupancy (beyond the effects of the baseline) and u_{AB} represents the effects of the interactions between A and B (beyond the individual effects of A, B, and the baseline). In the general case, a saturated model has a term for each subset of the attributes. The saturated model is not very interesting because it overfits the data: the maximum likelihood estimator for $q(t)$ is just the number of times t appears in the data. Thus the saturated model is also called the *unrestricted model*. Because the unrestricted model is too powerful (i.e., it overfits the data), a typical statistical analysis would only look at a subset of the possible interaction terms. For example, if our table had 3 attributes A, B, C, we could try to model it with

$$\log q(a, b, c) = u + u_A(a) + u_B(b) + u_C(c) \tag{5}$$

Since there is no interaction term (i.e., u_{AB}, u_{AC}, u_{BC}, u_{ABC}), this model seems to suggest that attributes A, B, and C are independent. This is not a coincidence: we will see that A, B, and C are indeed independent in the maximum likelihood distribution for that model. In general, the maximum likelihood estimates for a restricted model will not be the same as the cell counts (unlike the case for saturated models). Thus saturated models will always have higher log-likelihoods. A common way of measuring how well a restricted model fits the data, is to look at the difference between the log-likelihood of the saturated model and the log-likelihood of the restricted model [10]. Since these models estimate parameters for a multinomial distribution (i.e., the cell counts or, equivalently, the cell probabilities), the difference in log-likelihoods is exactly the KL-divergence between the two respective maximum-likelihood distributions.

Most log-linear models used in practice have the following property: if an interaction term is included in the model, then its lower order effects are also included. For example, if a log-linear model has a u_{ABCD} term, then there is a term u_X for every $X \subseteq \{A, B, C, D\}$. Log-linear models that have this property are called *hierarchical*, and we shall restrict our attention to these types of models.

Just as with marginals, we can build an interaction graph for hierarchical log-linear models: the vertices are the attributes and there is an edge between two vertices if both are contained in some interaction term (for example, if the model had a u_{ABC} term, there would be edges between A and B, B and C, and A and C). It is common to describe hierarchical log-linear models using only the highest order interaction terms [10]: if a model has a parameter u_{ABCD} and u_{BC}, we would omit u_{BC} because it is implied by the u_{ABCD} term. It is also common to represent interaction terms as $[ABC]$ rather than u_{ABC}. Thus the log-linear model in Equation 4 can be compactly represented as $[AB]$ and the log-linear model in Equation 5 can be compactly represented as $[A][B][C]$. This compact representation is sufficient for constructing the interaction graph.

The log-linear model is *graphical* if the interaction terms are exactly the maximal cliques of the corresponding interaction graph, and it is *decomposable* if it is graphical and if the interaction graph is decomposable as well. Figure 2(a) shows the interaction graph for the log-linear model $[ABC][BCD][DE]$.

The similarity between the interaction terms of a decomposable log-linear model and the marginals (of the base table) with the same attributes is not superficial. The model can be built using only the marginals whose attributes are specified by the interaction terms, and the maximum likelihood estimates of cell probabilities and expected values for a graphical decomposable log-linear model are given by Equations 1 and 2, respectively [20]. Furthermore, any conditional independence relations that we can read off the interaction graph are also true of the maximum likelihood estimator [10]:

THEOREM 5.2. *Given a graphical, decomposable interaction graph for a log-linear model, if the sets of variables/nodes $\mathcal{A}$ and $\mathcal{B}$ are separated by $\mathcal{C}$, then under the maximum likelihood distribution, $P(\mathcal{A}, \mathcal{B}|\mathcal{C}) = P(\mathcal{A}|\mathcal{C})P(\mathcal{B}|\mathcal{C})$.*

Thus we have two complementary goals in releasing marginals: to provide a set of constraints (marginals) that lead to a maximum entropy distribution that is as close as possible to the real data subject to privacy restrictions; and to determine a set of conditional independence relations that best approximates the data subject to privacy restrictions.

[1] In the statistical literature, the term "factor" is used instead of "attributes" and "level" instead of "attribute value"

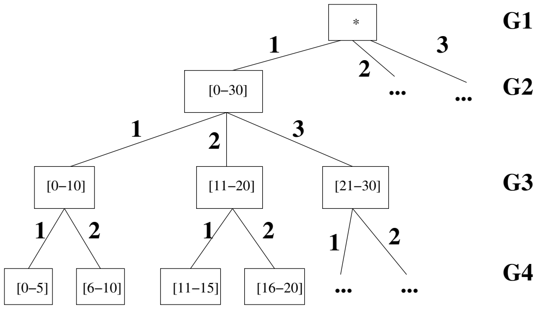


Figure 4: Age Hierarchy

Generalization $G4$			
L_1^{Age}	L_2^{Age}	L_3^{Age}	Count
1	1	1	5
1	1	2	3
1	2	1	2
1	2	2	0
1	3	1	0
1	3	2	0
2	1	1	4
2	1	2	1
...	...	...	...
3	1	1	0
3	1	2	0
...	...	...	...

Generalization $G3$		
L_1^{Age}	L_2^{Age}	Count
1	1	8
1	2	2
1	3	0
2	1	5
...	...	...
3	1	0
...	...	...

Figure 5: Induced Attributes

6. ANONYMIZED MARGINALS

In this section, we will provide a reduction from anonymized marginals to ordinary marginals (Section 6.1) and we will use this reduction to extend the theory of graphical models to anonymized marginals (Section 6.2).

6.1 A Reduction

For each attribute R_i, let $\mathcal{G}_i$ be the set of possible generalizations of R_i which are to be considered. For example, we may specify that we are only interested in three generalizations for the Age attribute: one that partitions Age into the intervals $[0-5]$, $[6-10]$, $\ldots$; one that partitions Age into intervals $[0-10]$, $[11-20]$, $\ldots$; and one that partitions Age into the intervals $[0-30]$, $[31-60]$, $\ldots$. For tractability, and for clarity, we will restrict our attention to the case where, for each R_i, the generalizations in $\mathcal{G}_i$ are totally ordered according to $\prec$ (Section 2.1). This restriction is common in the literature ([30, 21]). Without loss of generality we can assume that the most general generalization simply suppresses the attribute.

First, we require that a collection of anonymized marginals M_1, $M_2, \ldots, M_r$ only use the anonymizations we specify.

DEFINITION 6.1 (VALIDITY). *Let $M_1, M_2, \ldots, M_r$ be a set of marginals and let $R_1, \ldots, R_d$ be the attributes that appear in those marginals. For $i = 1 \ldots d$, let $\mathcal{G}_i$ be a set of generalizations for R_i such that $\mathcal{G}_i$ is totally ordered according to $\prec$. Then $M_1, \ldots, M_r$ are valid with respect to $\mathcal{G}_1, \ldots, \mathcal{G}_d$ if for every marginal M_j and every attribute R_q that appears in M_j, R_q has been generalized according to one of the generalizations in $\mathcal{G}_q$.*

Note that attributes can be generalized differently in different marginals: if Age appears in marginal M_1 and marginal M_2, it could have been generalized using the intervals $\{[0-5], [6-10], \ldots\}$ in M_1 and generalized using the intervals $\{[0-10], [11-20], \ldots\}$ in M_2. The reduction from anonymized to ordinary marginals relies on the fact that a totally ordered set of generalizations induces a natural hierarchy on the base domain. The reduction proceeds as follows. Let R_j be an attribute and $\mathcal{G}_j$ be a set of possible generalizations for R_j. Let h be the number of generalizations in $\mathcal{G}_j$. The first step is to label the generalizations so that the most general generalization is labeled G^1, the second most general generalization is labeled G^2, etc. Figure 4 shows (part) of a hierarchy over the age attribute. Here we have four generalizations $G^4 \prec G^3 \prec G^2 \prec G^1$, where G^1 is equivalent to suppressing the entire attribute. Each generalization represents one level of the hierarchy, and each node in the hierarchy tree has a bounded number of children (since we have finitely many data points). For each node, we order its children (arbitrarily) and number them according to that order. We will use this numbering to create a new set of attributes. For $i = 1, \ldots, h-1$, let c_i be the maximum number of children for any node in generalization G^i (in our example, $c_1 = 3$, $c_2 = 3$, and $c_3 = 2$).

We can now treat the attribute R_j as an $(h-1)$-dimensional vector of induced attributes $L_1^{(R_j)}, \ldots, L_{h-1}^{(R_j)}$ where the i^{th}dimension has c_i points. A point $(x_1, \ldots, x_{h-1})$ in this space represents the path taken from the root G^1 to a leaf node. It is easy to see that any generalization $G^i \in \mathcal{G}_j$ of R_j corresponds to the subspace consisting of the first $i-1$ dimensions $(L_1^{(R_j)}, \ldots, L_{i-1}^{(R_j)})$. Figure 5 shows the induced attributes for two of the generalizations in Figure 4. By applying this reduction to every attribute in every anonymized marginal $M_1, \ldots, M_r$, we get a new set of marginals $M'_1, M'_2, \ldots, M'_r$. The only restriction on these new marginals is that if attribute $L_m^{(R_j)}$ appears in some marginal M'_i then M'_i must also contain the attributes $L_1^{(R_j)}, L_2^{(R_j)}, \ldots, L_{m-1}^{(R_j)}$. Thus we view the set of induced attributes $\{L_1^{(R_j)}, L_2^{(R_j)}, \ldots, L_i^{(R_j)}\}$ as R_j at resolution level i. The higher the resolution, the more information there is about attribute R_j.

6.2 Extensions of the Theory

Given the reduction in Section 6.1, the notions of interaction graph and decomposability carry over directly to anonymized marginals. In this section, we will discuss how this affects the maximum entropy distribution and the statements of conditional independence. We will also discuss how to traverse the search space of anonymized marginals.

For anonymized marginals, the probability computation (Equation 1) needs to be extended to deal with the case where some attribute never appears at its highest level of resolution. Recall that for any attribute B, the induced attributes $L_1^{(B)}, \ldots, L_i^{(B)}$ represents a path from the root to an interior node of the generalization tree for B. Let $\text{child}(t, L_i^{(B)})$ be the function that first projects t onto the induced attributes $L_1^{(B)}, \ldots, L_i^{(B)}$ to get a node in the generalization tree and then returns the number of leaves in the subtree rooted at that node. For example, let us consider the left-most table in Figure 5. If t is a tuple in the first row, then $\text{child}(t, L_1^{Age}) = 6$ because t projected onto L_1^{Age} gives us the left-most child of the root, and it has 6 leaf descendants.

THEOREM 6.1. *Let $\mathcal{E}_1, \ldots, \mathcal{E}_p$ be the maximal cliques of a decomposable graphical model G arranged in a perfect sequence, and let $S_2, \ldots, S_p$ be the separators (as defined in Definition 5.7). Let $B_1, \ldots, B_q$ be the original attributes. For each original attribute B_j, let $L_{max}^{(B_j)}$ be the induced attribute of B_j that appears in one of the marginals (that correspond to the $\mathcal{E}_i$) such that $L_{max+1}^{(B_j)}$*

does not appear. Then the maximum entropy probability of a tuple $t \in \text{Domain}(T)$ *is:*

$$\frac{1}{|T|}\frac{\prod\limits_{i=1}^{p} T_{\mathcal{E}_i}(t_{\mathcal{E}_i})}{\prod\limits_{j=2}^{p} T_{S_j}(t_{S_j})}\cdot\prod_{j=1}^{q}\frac{1}{\text{child}\left(t, L_{max}^{(B_j)}\right)} \tag{6}$$

This says that the probability for a tuple t (that is not at the highest level of resolution) is spread uniformly across all possible completions of t.

The interpretation of conditional independence can also be extended to anonymized marginals. Let $L_1^{(A)}, \dots, L_p^{(A)}$ be the induced attributes for original attribute A. Recall that $L_1^{(A)}, \dots, L_{p-1}^{(A)}$ represent A at a lower level of resolution, $L_1^{(A)}, \dots, L_{p-2}^{(A)}$ represents A at an even lower level of resolution, etc (note if $L_i^{(A)}$ appears in a marginal then we have the requirement that $L_j^{(A)}$ appears in the same marginal for all $j < i$). Thus generalizing a marginal involves suppressing (marginalizing) the induced attribute $L_i^{(A)}$ with the largest index (i.e., the marginal does not contain an induced attribute $L_j^{(A)}$ with $j > i$). Equivalently, generalizing a marginal can be seen as reducing the level resolution for (original) attribute A. The conditional independence interpretation relies on the following theorem:

THEOREM 6.2. *Let* X*,* Y*, and* $\mathcal{C}$ *be disjoint sets of vertices of an interaction graph* G *for graphical decomposable anonymized marginals* $M_1, \dots, M_r$*. If* X *and* Y *are complete subgraphs of* G *and* $\mathcal{C}$ *is a set-wise minimal (i.e., no subset of* $\mathcal{C}$ *has this property) separator of* X *and* Y *then the following is true*

1. $\mathcal{C}$ *is complete (and therefore the vertices corresponding to* $\mathcal{C}$ *correspond to attributes that are contained in some marginal* M_j*)*
2. *If* $L_i^{(A)} \in X$ *then for all* j*,* $L_j^{(A)} \notin Y$*.*
3. *If* $L_i^{(A)} \in X$ *and* $L_j^{(A)} \in \mathcal{C}$ *then* $i > j$ *(i.e.,* $\mathcal{C}$ *has lower resolution information about original attribute* A *than does* X*).*

Theorem 6.2 says that if X is a set of attributes that appear in some anonymized marginal, and Y is a set of attributes that appear in some anonymized marginal, the minimal separator C between X and Y is also a set of attributes in some anonymized marginal. Furthermore, X and Y do not contain any of the same original attributes (even at different levels of resolution) for it does not make sense to talk about independence between the age ranges $\{[0-2], [3-10], [11-20], [21-30][30, \infty]\}$ and $\{[0-10], [11-20], [20, \infty]\}$. Additionally, if X and $\mathcal{C}$ (or Y and $\mathcal{C}$) have information about the same original attributes, then the information in X is an incremental gain in resolution over $\mathcal{C}$. Thus in addition to statements about conditional independence of attributes, we also have statements about independence of resolution: "given some the level of resolution in $\mathcal{C}$, the extra precision in X and Y is independent." For example, suppose we have a table of flu patients categorized by geographical region and age range ($[0-10]$, $[11-20]$, etc). Given a marginal that consists of age ranges ($[0-5]$, $[6-10]$, etc) and a marginal that consists of states (instead of just geographical regions), the maximum entropy distribution would be consistent with the assumption that given the first table, increased resolution in age is independent of the increased resolution in location. In other words, once we know that a flu patient is in the Northeast and is between 11 and 20 years old, knowing the exact state of a region would not help narrow the age range (assuming the maximum entropy distribution is correct).

The previous example raises an important issue – how correct is the maximum entropy distribution? Intuitively, adding additional marginals or merging marginals together (releasing $ABCD$ instead of AB and CD) gives us additional information and should help us better approximate the original distribution. In fact, this is also true mathematically.

THEOREM 6.3. *Let* G *and* H *be the interaction graphs of two decomposable graphical models. If the vertices of* H *are a subset of the vertices of* G *and the edges of* H *are a subset of the edges of* G*, then the maximum entropy distribution for* G *approximates the original table at least as well as the maximum entropy distribution for* H *(in terms of the KL-divergence).*

Note that the case where the vertices of G and the vertices of H are the same is proved in [24]. Since generalization may completely remove some induced attributes from all of the marginals, this removal will result in a model with less nodes as well as edges. Thus we need the result that adding edges and vertices to H (when the vertices of H are a subset of the vertices of G) never hurts utility.

The only case where adding marginals or merging them would not increase utility is when the tuple distribution of original table is exactly the maximum entropy distribution for that set of marginals. Since that is unlikely in practice, even the following simple technique is almost certainly guaranteed to improve the utility of a single anonymized table T' that was derived from a base table T: take the marginal M of T that has all attributes but the sensitive ones. Create a k-anonymous version M' of M. Then releasing M' and T' gives more utility than releasing T' alone (as is the standard practice in the literature). In Section 7 we will discuss how to make sure that privacy guarantees still hold.

We conclude this section with a discussion of how to select a set of anonymized marginals to publish. It is known that model selection for decomposable graphical models requires an exhaustive search [28] and that even finding an optimal k-anonymous table is NP hard [25, 3]. Therefore a search algorithm such as a genetic algorithm or a random walk on the space of models is needed. We will briefly discuss how to extend results on stepwise edge/vertex selection [35, 13] that will allow us to go from one graphical model to another.

The following three conditions need to be simultaneously satisfied in order to remove an edge connecting $L_i^{(A)}$ and $L_j^{(B)}$:

1. $L_i^{(A)}$ and $L_j^{(B)}$ cannot appear together in 2 or more marginals (equivalently, they do not both appear in a minimal separator between two nodes). This rule, due to Wermuth, ensures that the resulting model is decomposable [35].
2. $A \neq B$.
3. There is no edge connecting $L_{i'}^{(A)}$ and $L_j^{(B)}$ with $i' > i$ (and similarly for j).

The last two conditions ensure that the induced attributes in every marginal describe a path from the root to an interior node of the generalization tree (instead of only a subset of a path) and therefore correspond to an actual generalization. A node $L_i^{(A)}$ can be removed if

- There is no node $L_j^{(A)}$ with $j > i$.

This also ensures that the induced attributes in every marginal describe a path from the root to an interior node of the generalization tree. Note that removing a node from a decomposable graph results in a graph that is decomposable [20]. A node $L_i^{(A)}$ can be added to a graph if one of the following is true

- $i = 1$ (in which case the node is added with no edges) OR
- $L_j^{(A)}$ is already in the graph for every $j \leq i - 1$. In this case, $L_i^{(A)}$ is added with an edge to every $L_j^{(A)}$.

And edge between $L_i^{(A)}$ and $L_j^{(B)}$ can be added if the following conditions hold:

1. There exists a minimal separator S between $L_i^{(A)}$ and $L_j^{(B)}$ such that every node in S has an edge to both $L_i^{(A)}$ and $L_j^{(B)}$. This rule, due to Deshpande et al, ensures that the resulting model is decomposable [13].
2. $i = j = 1$ or for all $i' < i$ there is an edge from $L_{i'}^{(A)}$ to $L_j^{(B)}$ (and similarly for j)

7. ALGORITHMS

In this section we discuss procedures for checking a set of anonymized marginals for privacy. The first criterion, from Definition 4.1, is that an attempt to link any marginal to external data will give either 0 or at least k tuples. Thus it is sufficient to check that each marginal satisfies k-anonymity by itself.

The next requirement, k-combinatorial anonymity, is more stringent. An adversary should not be able to use combinatorial tools (such as the inclusion-exclusion principle) to determine that for all tables consistent with a set of given marginals, a particular cell must have between 1 and $k - 1$ tuples (for then this cell can be linked back to external data). In general, checking for privacy by computing upper and lower bounds for a cell is NP-hard [22]. However, when the marginals correspond to a decomposable graphical model, exact bounds can be computed in closed form. Dobra's bounds [15] extend to anonymized marginals: the cell count $T(t)$ is bounded by

$$T(t) \leq \min(T_{\mathcal{E}_1}(t_{\mathcal{E}_1}), \ldots, T_{\mathcal{E}_p}(t_{\mathcal{E}_p})) \qquad (7)$$

(note the similarity to Equation 1) and this bound is tight in the sense that for each upper bound, there exists a table that achieves it. Checking for k-combinatorial anonymity relies on Equation 7 and a variant of the maxent ℓ-diversity algorithm that is described below. Details are omitted due to lack of space.

Checking maxent ℓ-diversity for all points in $\mathrm{NonSenDomain}(T)$ is a harder task. First, there are several simplifications we can perform:

Proposition 7.1. *Let $M_1, \ldots, M_p$ be a set of anonymized marginals in a graphical model. Let V be the set of induced attributes that appear in at least one of the M_i. Then checking for maxent ℓ-diversity in $\mathrm{Domain}(T)$ is equivalent to checking for maxent ℓ-diversity in $\mathrm{Domain}(T_V)$.*

Proposition 7.1 tells us that the we do not need to worry about any level of resolution that does not appear in the marginals. Thus we can use Equation 1 instead of Equation 6 for our computations.

Proposition 7.2. *Let $\mathcal{E}_1, \ldots, \mathcal{E}_p$ be sets of (induced) attributes arranged in a perfect sequence (Definition 5.7) and let $S_2, \ldots, S_p$ be the corresponding separators. If any of these two conditions hold, then the set of anonymized marginals $\{T_{\mathcal{E}_1}, \ldots, T_{\mathcal{E}_p}\}$ does not satisfy maxent ℓ-diversity:*

1. *The marginal $T_{\mathcal{E}_i}$ is not ℓ-diverse (individually) for some i.*
2. *There exists a j such that the intermediate product $T_{\mathcal{E}_1}(t_{\mathcal{E}_1}) \prod\limits_{j=2}^{p} \frac{T_{\mathcal{E}_j}(t_{\mathcal{E}_j})}{T_{S_j}(t_{S_j})}$ corresponds to a probability distribution that is not ℓ-diverse.*

Proposition 7.2 tells us that sometimes intermediate results (rather than the complete maximum entropy distribution) can be used to determine if a set of marginals does not satisfy entropy ℓ-diversity.

Proposition 7.3. *Let C be the sensitive attribute. Let $T_{\mathcal{E}_1}$, ..., $T_{\mathcal{E}_p}$ be decomposable and graphical (anonymized) marginals, let $\mathcal{E}_1, \ldots, \mathcal{E}_p$ be sets of (induced) attributes arranged in a perfect sequence (Definition 5.7), and let $S_2, \ldots, S_p$ be the corresponding separators. Let $\mathcal{E}'_1, \ldots, \mathcal{E}'_{p_1}$ be the subsequence of the $\mathcal{E}_i$ consisting of precisely the $\mathcal{E}_i$ that contain C. Let $S'_2, \ldots, S'_{p_2}$ be the subsequence of the S_i consisting of precisely the S_i that contain C. Then the following is true:*

1. *$T_{\mathcal{E}'_1}, \ldots, T_{\mathcal{E}'_{p_1}}$ are graphical and decomposable marginals.*
2. *$\mathcal{E}'_1, \ldots, \mathcal{E}'_{p_1}$ is a perfect sequence and $S'_1, \ldots, S'_{p_2}$ are the corresponding separators to the perfect sequence (and hence $p_1 = p_2$).*
3. *Checking $T_{\mathcal{E}'_1}, \ldots, T_{\mathcal{E}'_{p_1}}$ for maxent ℓ-diversity in C is equivalent to checking $T_{\mathcal{E}_1}, \ldots, T_{\mathcal{E}_p}$ for maxent ℓ-diversity in C.*

Proposition 7.3 tells us two things. First, marginals that do not contain the sensitive attribute do not affect maxent ℓ-diversity at all. Thus if we publish a table that is both ℓ-diverse and k-anonymous, we can approximate the original table better (while preserving privacy) just by releasing additional k-anonymous marginals that do not contain the sensitive attribute. This is already an improvement over the standard technique of just releasing one table. Second, by ignoring marginals without sensitive attributes, we get a smaller decomposable graphical model to which we can apply Propositions 7.1 and 7.2. This lets us cut down on the size of the domain that must be checked.

It is clear that checking for maxent ℓ-diversity can be done in time that is linear in the size of the join of all the marginals containing the sensitive attribute. In cases where the sensitive attribute has a small domain, we can use the decomposable property of the interaction graph to reduce the complexity even further: we will prune away tuples that do not need to be joined. We will also present a variant of this pruning algorithm for the case where $|C| = \ell$. In this case, the running time will be $O(|C|^2 p|J|)$ where p is the number of marginals, $|J|$ is the size of the largest join between 2 marginals (not counting duplicates), C is the domain of the sensitive attribute and $|C|$ is its size. For other cases (when the overall join size is too large and when $|C|$ is large), we will present an algorithm that relaxes the ℓ-diversity conditions.

To discuss the algorithms, we need to introduce the following definition:

Definition 7.1 (Junction Tree). *Let $\mathcal{V} = \{V_1, \ldots, V_p\}$ be a collection of sets. A junction tree is a graph $(\mathcal{V}, E)$ that is a tree with the following property: for any $V_i, V_j \in \mathcal{V}$ and for any $V' \in \mathcal{V}$ in the path between V_i and V_j, we have $V_i \cap V_j \subseteq V'$.*

Figure 6 shows a junction tree for the interaction graph in Figure 2(a). If we let the V_i be the maximal cliques of a connected decomposable graph, then there always exists a junction tree that contains all of the V_i [20]. Junction trees can be created from scratch in time that is quadratic in the number of cliques [18] or maintained incrementally as in [13]. Because of Proposition 7.3, we can assume

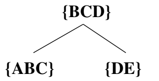


Figure 6: Junction Tree for Figure 2(a)

Algorithm 1 : Diversity_Check(node: v)

```
Require: Each node $v$ of a junction tree is a set of attributes
$I_v \leftarrow T_v$
for all $x \in \text{children}(v)$ do
    Diversity_Check($x$)
    Tmp$\leftarrow I_v \bowtie I_x$
    for all $t \in$Tmp do
        Tmp($t$) $\leftarrow I_v(t_v) \cdot I_x(t_x)/T_{v\cap x}(t_{v\cap x})$
    end for
    Prune(Tmp,$v$)
    $I_v \leftarrow$ Tmp
end for
if $v$ =root then
    Check if $I_v$ satisfies $(c, \ell)$-diversity
else
    Prune(Tmp,$v \cap \text{parent}(v)$)
    $I_v \leftarrow$ Tmp
end if
```

that all marginals contain the sensitive attribute C, that the decomposable graph is therefore connected, and therefore that a junction tree exists. In our case, each node of the junction tree is a set of attributes and corresponds to a marginal. It is not hard to see that any topological sort of a junction tree results in a perfect sequence (Definition 5.7), and that the intersection between a parent and child is the separator for the child in the perfect sequence. Thus we can compute the expected cell counts (in Equation 2) by multiplying the marginal counts $T_v(t_v)$ corresponding to each node v, and dividing by the separators $T_{v\cap\text{parent}(v)}(t_{v\cap\text{parent}(v)})$.

We will perform all joins in the junction tree from the bottom up. For a node v, let $\mathcal{A}_v$ be the set of attributes that appear in the subtree rooted at v. Note that attributes in $\mathcal{A}_v$ not involved in a join between v and $\text{parent}(v)$ will never be used later on because, by definition of the junction tree, those attributes are separated from the rest of the tree by the attributes that are involved in the join. For example, in Figure 6, $\{ABC\}$ can be joined with its parent $\{BCD\}$ using the attributes B and C. A is not involved in the join and so does not appear anywhere except in the subtree rooted at $\{ABC\}$. The attributes of v that appear only in the subtree rooted at v will be denoted by $\text{irrel}(v)$ (because they are *irrelevant* for the join between v and its parent and any other join that will be performed afterwards) and the rest of the nonsensitive attributes of v will be denoted by $\text{rel}(v)$ (relevant). After each join, we will group tuples into *relevant blocks* where all tuples with the same values for the attributes in $\text{rel}(v)$ are in the same relevant block. Within each relevant block we will do the pruning.

To do pruning, first note that each relevant block is composed of anonymized groups (recall that an anonymized group consists of all tuples with the same values for the nonsensitive attributes; in this case they are the attributes in $\text{rel}(v)$ and $\text{irrel}(v)$). For pruning, we will treat each anonymized group as a vector of length $|C|$ where the i$^{\text{th}}$ component is the frequency of sensitive value s_i in the anonymized group. In the pruning step, we remove all anonymized groups that are not in the convex hull in their respective relevant blocks.

The pruning algorithm runs from the bottom up. For each node v whose children $d_1, \ldots, d_j$ are all leaves, it sequentially joins the marginals corresponding to v and its children to get an intermediate result $\mathcal{I}_v$. For each tuple $t \in \mathcal{I}_v$, the pruning algorithm computes the expected count by multiplying the marginal counts and dividing by the separators: $T_v(t_v)\prod_i(T_{d_i}(t_{d_i})/T_{v\cap d_i}(t_{v\cap d_i}))$. We can think of $\mathcal{I}_v$ as a new marginal where the count of each cell is the expected count that we computed. $\mathcal{I}_v$ will be treated as the "new" marginal for v and so $\mathcal{I}_v$ will itself be joined with the parent of v and v's siblings. After each join, a pruning step is performed. The pseudo-code is shown in Algorithm 1. Note that Algorithm 1 calls a procedure called "Prune" which takes two arguments. The first is a marginal and the second is the set of attributes $\text{rel}(v)$. In the basic pruning algorithm, "Prune" removes anonymized groups that are not part of the convex hull of their relevant blocks.

THEOREM 7.1 (CORRECTNESS OF PRUNING). *If there exists any $t \in \text{NonSenDomain}(T)$ that is not maxent ℓ-diverse then at least one such t will belong to an unpruned anonymized group of $\mathcal{I}_{root}$ at the end of the algorithm.*

In the case where $\ell = |C|$, we can efficiently check for (c, ℓ)-diversity while avoiding the computation of convex hulls. To accomplish this, we only need to modify the "Prune" procedure. Let $s_1, \ldots, s_\ell$ be the sensitive values of $|C|$. Within each relevant block B_i we do the following. For each ordered pair $(s_j, s_{j'})$ of sensitive values, we find and retain the anonymized group in B_i where the ratio of the frequencies of s_j to $s_{j'}$ is maximal. The anonymized groups that are not retained are discarded. Thus for each combination of values that will participate in a join between two marginals, we have at most $|C|^2$ tuples. With this pruning procedure, once we get to the root, we look at the ratio of frequencies of s_j to $s_{j'}$ (for all j, j') in each anonymized group. If all of the ratios are $\leq c$ then the marginals satisfy maxent (c, ℓ)-diversity.

In the case where the size of the join of all marginals containing the sensitive value is large (the worst case occurs when all marginals contain the same sensitive attribute and one additional attribute), and when $|C|$ is large, there are several ways we can speed up the checking for maxent ℓ-diversity. The first approach is to reduce the join size by imposing additional restrictions on the structure of anonymized marginals. When searching through the space of collections of anonymized marginals, we can restrict our attention to collections where at most m (a user-defined parameter) anonymized marginals contain the same sensitive attribute. Another approach is to take a base table and to first apply any of the existing algorithms that can be used to generate minimal ℓ-diverse tables (see [23, 21, 6]); an ℓ-diverse table T' is minimal if there is no ℓ-diverse table that can be transformed into T' by using generalizations. Afterwards, when we search for collections of anonymized marginals, we only consider collections that include that particular ℓ-diverse version of the base table as one of the marginals. This type of search is equivalent to starting out with an ℓ-diverse table and searching for which additional anonymized marginals can be published as well (thus these marginals are *injecting* utility into the original anonymized table). The inclusion of such an ℓ-diverse table naturally limits the join size, and our experiments indicate empirically that this approach also yields good utility.

The other approach is to relax the ℓ-diversity requirements. Given a relaxation parameter ϵ, we can guarantee that for each $t \in \text{NonSenDomain}(T)$, there are at least ℓ sensitive values that are at least $(1-\epsilon)^p$ times as frequent as the most frequent sensitive value for t. First we have a preprocessing step where for each non-root node v and each tuple $t \in T_v$, we set $T^*_v(t) = T_v(t)/T_{v\cap\text{parent}(v)}(t)$ (i.e., we perform the division by the sepa-

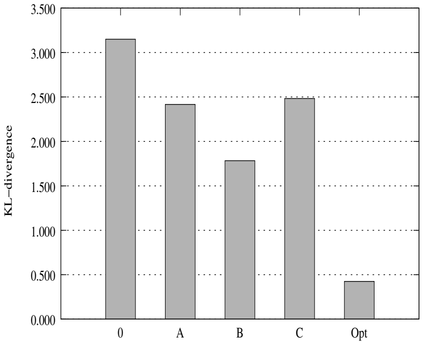


Figure 7: Anonymized Tables

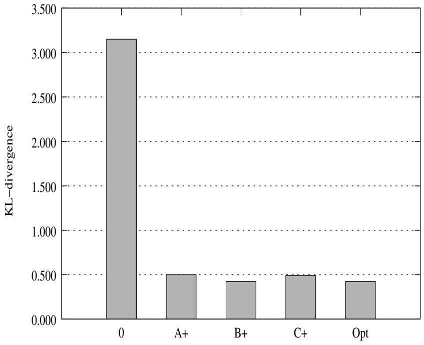


Figure 8: Anonymized Marginals

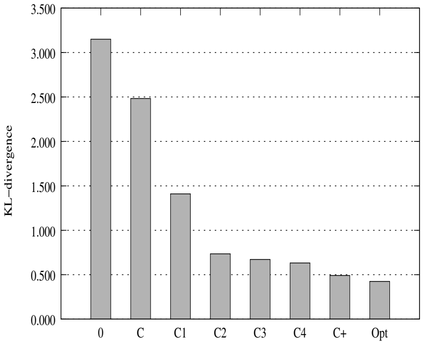


Figure 9: Incremental Utility

rators in advance). When v is the root, we set $T^*_v = T_v$. We then form the marginals I_v as follows. For each $t \in T^*_v$, we examine the anonymized group to which t belongs. Let t_{max} be the tuple in t's anonymized group such that $T^*_v(t_{max})$ is maximized. If $T^*_v(t) \geq (1-\epsilon)T^*_v(t_{max})$ then $I_v(t) = 1$ and otherwise $I_v(t) = 0$. These are now the marginals that would appear in Line 1 of Algorithm 1. Line 6 is now replaced by $Tmp(t) \leftarrow I_v(t_v) * I_x(t_x)$. Each anonymized group can now be treated as a vector of length $|C|$ where the i^{th} component is 1 if and only if $I_v(t_i) = 1$ (where t_i is the tuple in the anonymized group such that $t.C = s_i$). The the pruning step removes redundant anonymized groups in each relevant block. It also removes anonymized groups which *dominate* another anonymized group: if $\vec{g_1}$ and $\vec{g_2}$ are vectors corresponding to anonymized groups g_1 and g_2, respectively, then we say g_1 dominates g_2 if every component of $\vec{g_1}$ is greater than or equal to every component of $\vec{g_2}$. At the root, we say that there is maxent ℓ-diversity with ϵ relaxation if each anonymized group, when treated as a vector, has at least ℓ components equal to 1.

8. EXPERIMENTS

We performed our experiments on the Adult dataset in the UCI Machine Learning Repository [29]. We removed all tuples with missing values and were left with a table containing 45222 tuples. We used the attributes *race*, *gender*, *age*, and *marital status* as the nonsensitive attributes, and *occupation* as the sensitive attribute. Using the same generalization hierarchies as in [21] and [23], we generated three tables that were simultaneously 6-diverse and 6-anonymous. These tables were minimal in the sense that any other 6-diverse, 6-anonymous table can be generated from one of these three by using generalizations.

We measured utility in terms of KL-divergence; the smaller the number, the better it approximates the original un-anonymized table. Figure 7 shows the utilities of the three minimal 6-diverse, 6-anonymous tables. The bar labeled *0* corresponds to the KL-divergence to the table where all nonsensitive attributes were completely suppressed (i.e., they were generalized to a single value). The bar labeled *opt* corresponds to the KL-divergence to the best set of anonymized marginals that satisfy 6-anonymity and maxent 6-diversity. The three 6-diverse, 6-anonymous tables were labeled *A*, *B*, and *C*. As we can see, the anonymized tables do not approximate the original table particularly well.

One way to speed up the search for a good collection of anonymized marginals, and to make checking for ℓ-diversity more efficient, is to start with an anonymized table and to only consider collections of anonymized marginals such that the given anonymized table is one of them (i.e., start with an anonymized table and *inject* utility by adding additional anonymized marginals). For each of the anonymized tables *A*, *B*, and *C*, we found the best sets of anonymized marginals that contained each table. The results are shown in in Figure 8. Here the bar labeled $A+$ is the KL-divergence to the best collection of marginals that contain the anonymized table *A* (similarly for *B*$+$, and *C*$+$). In our experiments it turned out that the best collection of anonymized marginals containing *B* was also the overall best collection of anonymized marginals (whose utility is labeled *opt* in Figures 7 and 8).

Finally, in Figure 9 we show that even a very simple search for anonymized marginals can yield dramatic results when compared to the utility of just a single anonymized table. To illustrate this effect, we used table *C*, although our results were qualitatively similar for tables *A* and *B* as well. We measured how the KL-divergence decreased as we added marginals that contained only one attribute each. The marginals were added in order of greatest improvement in utility. Starting out with table *C*, we first added a marginal on *race* (bar labeled *C1*), to this we then added a marginal on *marital status* (bar *C2*), then *gender* (bar *C3*), and finally we also added marginal on age (bar *C4*). The marginal on age was bucketized into ranges of size 5 ($[0-4], [5-9], \dots$) in order to meet the k-anonymity requirements. Note that there is still a noticeable difference in utility between this collection of anonymized marginals and the best collection that contains table *C* (as well as the overall best collection of anonymized marginals); however, this simple collection of marginals still created an enormous improvement in utility over a single anonymized table.

9. RELATED WORK

The utility of data that has been altered to preserve privacy has often been studied in contexts where the future use of the data is known. For example, [16] studies how to reconstruct association rules after noise has been added; [5] and [4] study how to reconstruct the distribution of a continuous variable after noise with a known distribution has been added; [9] studies how to perturb the values of continuous numeric attributes so that data clusters can be reconstructed (note that [9] also proposes publishing perturbed data in addition to a histogram, but this method does not handle non-numeric attributes and the privacy guarantees use the assumption that the data is generated from a uniform distribution); and [17] and [34] anonymize data while trying to maximize decision tree accuracy. There have also been some negative results for utility. In addition to the curse of dimensionality for k-anonymity [2], there is work showing that an ideal privacy criterion places extremely strong restrictions on the types of queries that can be answered [26] (in particular, aggregate statistics cannot be computed). k-

Anonymity [33] and ℓ-diversity [23] are weaker privacy definitions (they do not protect against adversaries with arbitrary amounts of background knowledge) but they provide considerably more utility.

There are several approaches to sanitizing a dataset to ensure privacy. These include generalizations [31], tuple suppression [11, 31], adding noise [1, 5, 16, 9], publishing marginals that satisfy a safety range [15], and data swapping [12] – a technique where attributes are swapped between tuples in such a way that certain marginal totals are preserved. Queries can also be posed online and the answers audited [19] or perturbed [14].

Log-linear models [10, 20] and logistic regression are popular techniques for analyzing tabular data, and graphical models [20, 28, 27] provide a compact and interpretable representation of high-dimensional probability distributions.

The maximum entropy distribution that satisfies given constraints has also been studied in the database literature. For example, this has been applied to the exploration of OLAP data cubes [32].

10. CONCLUSIONS AND FUTURE WORK

Anonymized marginals can be thought of as statements about the original data set that are guaranteed to be true. The maximum entropy distribution is then our best guess about the rest of the data. Another way to think of this is that anonymized marginals are a compact representation of a statistical model (a density estimate of the original table). A promising direction of future work is releasing a set of models in addition to the data, studying the utility of such an ensemble, providing guarantees about the resulting privacy of information, and constructing data mining algorithms that use all of this information as the input.

11. REFERENCES

[1] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, 1989.

[2] Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[3] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology (JOPT)*, 2005.

[4] D. Agrawal and C. C. Aggarwal. On the design and quantifiaction of privacy preserving data mining algorithms. In *PODS*, May 2001.

[5] R. Agrawal and R. Srikant. Privacy preserving data mining. In *SIGMOD*, May 2000.

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

[7] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, 1996.

[8] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *ICDT*, pages 217–235, 1999.

[9] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In *Theory of Cryptography Conference*, 2005.

[10] Ronald Christensen. *Log-Linear Models and Logistic Regression*. Springer-Verlag, 1997.

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

[12] T. Dalenius and S. Reiss. Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.

[13] Amol Deshpande, Minos N. Garofalakis, and Michael I. Jordan. Efficient stepwise selection in decomposable models. In *UAI*, pages 128–135, 2001.

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

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

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

[17] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.

[18] Finn Verner Jensen and Frank Jensen. Optimal junction trees. In *UAI*, pages 360–366, 1994.

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

[20] S. L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

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

[22] Jesús A. De Loera and Shmuel Onn. The complexity of three-way statistical tables. *SIAM J. Comput.*, 33(4):819–836, 2004.

[23] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *ICDE*, 2006.

[24] Francesco M. Malvestuto. Approximating discrete probability distributions with decomposable models. *IEEE Transactions on systems, Man and Cybernetics*, 21(5):1287–1294, 1991.

[25] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.

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

[27] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, December 2000.

[28] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

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

[30] P. Samarati. Protecting respondents' identities in microdata release. In *TKDE*, pages 1010 – 1027, 2001.

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

[32] Sunita Sarawagi. User-adaptive exploration of multidimensional data. In *VLDB*, 2000.

[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] K. Wang, B. C. M. Fung, and P. S. Yu. Template-based privacy preservation in classification problems. In *ICDM*, November 2005.

[35] Nanny Wermuth. Model search among multiplicative models. *Biometrics*, 1976.