Published in Journal of Computer Security (JCS), Volume 20, Issue 4, Pag 393-436, 2012.

Modeling and Preventing Inferences from Sensitive Value Distributions in Data Release*

Michele Bezzi,[1] Sabrina De Capitani di Vimercati,[2] Sara Foresti,[2]
Giovanni Livraga,[2] Pierangela Samarati,[2] Roberto Sassi[2]

[1]SAP Research, Sophia-Antipolis, France
michele.bezzi@sap.com
[2]DTI - Università degli Studi di Milano - 26013 Crema, Italy
{sabrina.decapitani, sara.foresti, giovanni.livraga, pierangela.samarati}.@unimi.it

Corresponding author: Pierangela Samarati
DTI - Università degli Studi di Milano
Via Bramante 65 - 26013 Crema, Italy
pierangela.samarati@unimi.it
phone: +39-0373-898061, *fax:* +39-0373-898010

Abstract

Data sharing and dissemination are becoming increasingly important for conducting our daily life activities. The main consequence of this trend is that huge collections of data are easily available and accessible, leading to growing privacy concerns. The research community has devoted many efforts aiming at addressing the complex privacy requirements that characterize the modern Information Society. Although several advancements have been made, still many open issues need to be investigated.

In this paper, we consider a scenario where data are incrementally released and we address the privacy problem arising when sensitive non released properties depend on (and can therefore be inferred from) non-sensitive released data. We propose a model capturing this inference problem, where sensitive information is characterized by peculiar value distributions of non sensitive released data. We then describe how to counteract possible inferences that an observer can draw by applying different statistical metrics on released data. Finally, we perform an experimental evaluation of our solution, showing its efficacy.

keywords: sensitive distribution, inference, continuous data release

*A preliminary version of this paper appeared under the title "Protecting Privacy of Sensitive Value Distributions in Data Release," in *Proc. of the 6th Workshop on Security and Trust Management* (STM 2010), Athens, Greece, September 23-24, 2010 [4].

1 Introduction

Sharing and dissemination of information play a central role in today's Information Society. Governmental, public, and private institutions are increasingly required to make their data electronically available as well as to offer services and data access over the Internet. This implies disclosing to external parties or sharing information once considered classified or accessible only internally, which must now be made partially available for outside interests. A notable side effect of this scenario is that there is a tremendous exposure of private and sensitive information to privacy breaches. Data publication and sharing must then ensure, on one hand, the satisfaction of possible needs for data from external parties and, on the other hand, proper protection of sensitive data, which should be neither directly released nor indirectly leaked. Ensuring privacy to sensitive data is a complex problem, as the possible correlations and dependencies existing among data can introduce inference channels causing leakage of sensitive information even if such information is not explicitly released. This problem has been under the attention of researchers for decades and has been analyzed from different perspectives, resulting in a large body of research that includes: statistical databases and statistical data publications (e.g., [1]); multilevel database systems with the problem of establishing proper classification of data, capturing data relationships and corresponding inference channels (e.g., [12, 24]); ensuring privacy of respondents' identities or of their sensitive information when publishing macro or micro data (e.g., [9, 10]); protection of sensitive data associations due to data mining (e.g., [2]). Several approaches have been proposed addressing all these aspects, and offering solutions to block or limit the exposure of sensitive or private information. However, new scenarios of data publication, coupled with the richness of published data and the large umber of available data sources, raise novel problems that still need to be addressed.

In this paper, we address a specific problem related to inferences arising from the dependency of sensitive (not released) information referred to some entities on other properties (released) regarding such entities. In particular, we are concerned with the possible inferences that can be drawn by observing the distribution of values of non sensitive information associated with these entities. As an illustrating example, the age distribution of the soldiers in a military location may permit to infer the nature of the location itself, such as a headquarter (hosting old officials) or a training campus (hosting young privates), which might be considered sensitive. Such a problem of sensitive information derivation becomes more serious as the amount of released data increases, since external observations will tend to be more representative of the real situations and the confidence in the external observations will increase. Although this problem resembles in some aspects the classical problem of controlling horizontal aggregation of data, it differs from it in several assumptions. In particular, we assume a scenario where an external observer could gather the data released to legitimate users and inference is due to peculiar distributions of data values. Also, we are concerned not only with protecting sensitive information

associated with specific entities, but also with avoiding possible false positives, where sensitive values may be improperly associated (by the observers) with specific entities.

A preliminary version of this work appeared in [4]. Here, we extend our earlier proposal by introducing several metrics to assess the inference exposure due to data release. Our metrics are based on the concepts of *mutual information*, which has been widely used in several security areas ranging from the definition of distinguishers for differential side-channel analysis (e.g., [3, 5, 20, 39]) to data-hiding and watermarking security (e.g., [6]), and of *distance* between the expected and the observed distribution of values of non sensitive information. We then revise the definition of safe release according to the proposed metrics, and describe the controls enforced in a scenario where tuples are released one at a time, upon request. Also, we present an experimental evaluation proving the effectiveness of our solution.

The remainder of this paper is organized as follows. Section 2 introduces our reference scenario of inference in data publication, raised from a real case study that needed consideration. Section 3 formally defines the problem of releasing a dataset without leaking (non released) sensitive information due to the dependency existing between the frequency distribution of some properties of the released dataset and the not released information. Section 4 describes two possible strategies that use the mutual information and distance between distributions for counteracting the considered inference problem. Section 5 illustrates how the two strategies proposed can be concretely implemented by adopting different metrics that determine when a data release is safe with respect to inference channels that may leak sensitive information. Section 6 describes how to control the on-line release of the tuples in a dataset. Section 7 discusses the experimental results proving the effectiveness of our solution. Section 8 presents related work. Finally, Section 9 gives our conclusions.

2 Reference scenario and motivation

We consider a scenario (see Figure 1) where a *data holder* maintains a collection of records stored in a trusted environment. Each record contains different attributes and pertains to a unique data respondent, who is the only authorized party that can require its release. While the records individually taken are not sensitive, their aggregation is considered sensitive since it might enable inferring sensitive information not appearing in the records and not intended for release. We assume all requests for records to be genuine and communication to data respondents of responses to their record release requests to be protected. As a consequence, malicious observers are aware neither of the requests submitted by respondents nor of the data holder answers. We also assume that the number of records stored at the data holder site is kept secret. However, once records are released, the data holder has no control on them and therefore *external observers* can potentially gather all the records released. This may happen even with cooperation of respondents, in the case of external servers where

released data may be stored.

The data holder must ensure that the collection of records released to the external world be safe with respect to potential inference of sensitive (not released) information that could be possible by aggregating the released records. We consider a specific case of horizontal aggregation and inference channel due to the distribution of values of certain attributes with respect to other attributes. In particular, inference is caused by a distribution of values that deviates from expected distributions, which are considered as typical and are known to the observers. In other worlds, a record is released only if, when combined with records already released, does not cause a deviation of the distribution of the records released from the expected distribution.

In the reminder of this paper, we refer our examples to a real case scenario characterized as follows. The data holder is a military organization that maintains records on its personnel. Each record refers to a soldier and reports attributes `Name`, `Age`, and `Location` where the soldier is on duty. Some of the military locations are headquarters of the army. The information that a location is a headquarter is considered sensitive and neither appears in the soldiers' records nor it is released in other forms. Soldiers' records can be released upon request of the soldiers. In addition, the age distribution of soldiers is a distribution that can be considered common and widely known to the external world and, in general, typically expected at each location. However, locations where headquarters are based show a different age distribution, characterized by an unusual peak of soldiers middle age or older. Such a distribution clearly differs from the expected age distribution, where the majority of soldiers are in their twenties or thirties. The problem is therefore that, while single records are considered non sensitive, an observer aggregating all the released records could retrieve the age distribution of the soldiers in the different locations and determine possible deviations from the expected age distribution for certain locations, thus inferring that a given location hosts a headquarter. As an example, consider an insurance company offering special rates to military personnel. If all the soldiers subscribe a policy with this company to take advantage of the discount, the insurance company (as well as any user accessing its data) has knowledge of the complete collection of released records and can therefore possibly discover headquarter locations. Our problem consists in ensuring that the release of records to the external world be safe with respect to such inferences. The solution we describe in the following provides a response to this problem by adopting different metrics to assess the inference exposure of a set of records and, based on that, to decide whether a record (a set thereof) can be released.

3 Data model and problem definition

We provide the notation and formalization of our problem. Our approach is applicable to a generic data model with which the data stored at the data holder site could be organized. For concreteness, we assume data to

be maintained as a relational database. Consistently with other proposals (e.g., [34]), we consider the data collection to be a single table T characterized by a given set A of attributes; each record in the data collection is a tuple t in the table. Among the attributes contained in the table, we distinguish a set $Y \subset A$ of attributes whose values represent entities, called *targets*.

Example 3.1 *In our running example, table T is defined on the set A=*{`Name`, `Age`, `Location`}* of attributes, with Y=*{`Location`}*. We assume that the domain of attribute* `Location` *includes values L_1, L_2, L_3, L_4, L_5, representing five different military locations.*

While targets, that is, the entities identified by Y (locations in our example), are non sensitive, they are characterized by *sensitive properties*, denoted $s(Y)$, which are not released. In other words, for each $y \in Y$ the associated sensitive information $s(y)$ does not appear in any released record. However, inference on it can be caused by the distribution of the values of a subset of some other attributes $X \subseteq A$ for the specific y. We denote by $P(X)$ the set of *relative frequencies* $p(x)$ of the different values x in the domain of X which appear in table T. Also, we denote by $P(X|y)$ the relative frequency of each value in the domain of X appearing in table T and restricted to the tuples for which Y is equal to y. We call this latter the *y-conditioned distribution* of X in T.

Example 3.2 *In our running example, $s(Y)$ is the type of the location (e.g., headquarter). The sensitive information $s(y)$ of whether a location y is a headquarter (L_2, in our example) can be inferred from the distribution of the age of soldiers given the location. Figure 2(a) shows how tuples stored in table T are distributed with respect to the values of attributes* `Age` *and* `Location`*. For instance, of the* 10000 *tuples,* 2029 *refer to location L_1,* 72 *refer to soldiers with age lower than* 18*. Figure 2(b) reports the corresponding relative frequencies of age distributions. In particular, each column L_i, $i = 1, \ldots, 5$, reports the L_i-conditioned distribution $P($`Age`$|L_i)$ (for convenience expressed in percentage). For instance,* 3.55% *of the tuples of location L_1 refer to soldiers with age lower than* 18*. The last column of the table reports the distribution of the age range regardless of the specific location and then corresponds to $P($`Age`$)$ (expressed in percentage). For instance, it states that* 2.56% *of the tuples in the table refer to soldiers with age lower that* 18*. Figure 2(c) reports the distribution of soldiers in the different locations regardless of their age (again expressed in percentage). For instance,* 20.29% *of the* 10000 *soldiers are based at L_1.*

The existence of a correlation between the distribution of values of attributes X for a given target y and the sensitive information $s(y)$ is captured by the definition of *dependency* as follows.

Definition 3.3 (Dependency) *Let T be a table over attributes A, let X and Y be two disjoint subsets of A, and let $s(Y)$ be a sensitive property of Y. A* dependency*, denoted $X \rightsquigarrow Y$, represents a relationship existing between the conditional distribution $P(X|y)$ and the value of the sensitive property $s(y)$, for any $y \in Y$.*

The existence of a dependency between the y-conditioned distribution of X and the sensitive property $s(y)$ introduces an inference channel, since the visibility on $P(X|y)$ potentially enables an observer to infer the sensitive information $s(y)$ even if not released. For instance, with respect to our running example, `Age`$\rightsquigarrow$`Location`.

Definition 3.3 simply states the existence of a dependency but does not address the issue of possible leakages of sensitive information. In this paper, we consider the specific case of leakage caused by *peculiar* value distributions that differ from what is considered typical and expected. We then start by characterizing the expected distribution, formally defined as *baseline distribution* as follows.

Definition 3.4 (Baseline distribution) *Let A be a set of attributes, and X be a subset of A. The* baseline distribution *of X, denoted $B(X)$, is the expected distribution of the different values (or range thereof) of X.*

The baseline distribution is the distribution publicly released by the data holder and can correspond to the real distribution of the values of attributes X in table T (i.e., $B(X)$=$P(X)$) at a given time or can be a "reference" distribution considered typical. We assume the data holder to release truthful information and, therefore, that the baseline distribution resembles the distribution of the values of X in T at a given point in time (note that T may be subject to changes over time, for example, due to the enrollment of new soldiers and the retirement of old soldiers). This being said, in the following, for simplicity, we assume the baseline distribution $B(X)$ to coincide with $P(X)$. When clear from the context, with a slight abuse of notation, we will use $P(X)$ to denote the baseline distribution.

Example 3.5 *The baseline distribution P(`Age`) corresponds to the values (expressed in percentage) in the last column of Figure 2(b), which is also graphically reported as a histogram in Figure 3(a). Figures 3(b)-(f) report the histogram representation of the L_i-conditioned distributions for the different locations in T. As clearly visible from the histograms, while locations L_1, L_3, L_4, and L_5 enjoy a value distribution that resembles the expected baseline, location L_2 (the headquarter) shows a considerably different distribution.*

Our goal is to avoid the inference of the sensitive information caused by *unusual* distributions of values of X, with respect to specific targets y, in Y that the observer can learn from viewing released tuples (i.e., the y-conditioned distributions computed over released tuples present some peculiarities that distinguish it from the baseline distribution). To this purpose, in the following sections we illustrate a solution that the data holder can adopt for verifying whether the release of a tuple referred to a target y, together with the previously released tuples, may cause the inference of the sensitive property $s(y)$ and then whether the release of such a tuple can be permitted or should be denied.

4 Characterization of the inference problem

In our characterization of the problem, X and Y can be intended as two dependent random variables, meaning that there is a correlation between the values of X and Y. Due to this dependency, a potential observer can exploit the distribution of values of X for a given target y (i.e., the y-conditioned distribution) for inferring sensitive property $s(y)$. To counteract this type of inference, we obfuscate the dependency between X and Y in the released dataset, by adopting one of the following two strategies: *i)* make X and Y appear as two statistically independent random variables; or *ii)* minimize the distance between the y-conditioned distribution $P(X|y)$ and the baseline distribution $P(X)$.

Statistical independence. The first strategy ensures that the joint probability $P(X,Y)$ be "similar" to $P(X)P(Y)$. Since when X and Y are two independent variables the joint probability $P(X,Y)$ is equal to $P(X)P(Y)$, this strategy aims at releasing tuples such that the correlation between X and Y is not visible. As a consequence, the knowledge of the distribution of X does not give any information about the sensitive property $s(y)$ for each target y in Y. A classical measure of the dependency between two random variables is the *mutual information*, denoted $I(X,Y)$. It expresses the amount of information that an observer can obtain on Y by observing X, and viceversa. The mutual information $I(X,Y)$ of two random variables X and Y is defined as follows.

$$I(X,Y) = \sum_{x\in X, y\in Y} p(y)p(x|y)\ \log_2 \frac{p(x|y)}{p(x)}$$

The lower the mutual information in the released dataset, the more random variables X and Y resemble statistical independent variables.

Example 4.1 *Consider the distributions of the* `Age` *values for the different locations and* $P($`Age`$)$ *in Figure 2*(b)*, and the values* $p(L_i)$*,* $i = 1,\ldots,5$*, reported in Figure 2*(c)*. We have:*
$I($`Age`$,$`Location`$) = p(L_1)[p(< 18|L_1)\log_2 \frac{p(<18|L_1)}{p(<18)} + \ldots + p(\geq 55|L_1)\log_2 \frac{p(\geq 55|L_1)}{p(\geq 55)}] + \ldots + p(L_5)[p(< 18|L_5)\log_2 \frac{p(<18|L_5)}{p(<18)} + \ldots + p(\geq 55|L_5)\log_2 \frac{p(\geq 55|L_5)}{p(\geq 55)}] = 0.063285$

Distance between distributions. The second strategy ensures that when tuples are released, the y-conditioned distribution of all targets y in Y be "similar" to the baseline distribution. Intuitively, this strategy aims at hiding the peculiarities of the distribution of variable X with respect to a specific y so that an observer cannot infer anything about sensitive property $s(y)$. This strategy is then based on the evaluation of the distance between the baseline distribution $P(X)$ and the y-conditioned distribution $P(X|y)$. The distance between two distributions can be computed in different ways. The metrics that will be considered in the following section

adopt either the classical notion of *Kullback-Leibler distance* between distributions, denoted Δ, or the *Pearson's cumulative* statistic, denoted F.

The Kullback-Leibler distance nicely fits our scenario since it has a straightforward interpretation in terms of Information Theory. In fact, it represents a possible decomposition of the mutual information [17]. Given two distributions $P(X)$ and $P(X|y)$ their Kullback-Leibler distance is defined as follows.

$$\Delta(X, y) = \sum_{x \in X} p(x|y) \log_2 \frac{p(x|y)}{p(x)}$$

It is easy to see that the mutual information represents the weighted average of the Kullback-Leibler distance for the different targets, where the weight corresponds to the frequency of value y.

Example 4.2 *Consider the distributions of* `Age` *values for the different locations and the baseline distribution* $P(\texttt{Age})$ *in Figure 2(b). We have:*

$$\Delta(\texttt{Age}, L_1) = p(<18|L_1) \log_2 \frac{p(<18|L_1)}{p(<18)} + \ldots + p(\geq 55|L_1) \log_2 \frac{p(\geq 55|L_1)}{p(\geq 55)} = 0.047349$$

Similarly, we obtain: $\Delta(\texttt{Age}, L_2) = 0.358836$*,* $\Delta(\texttt{Age}, L_3) = 0.013967$*,* $\Delta(\texttt{Age}, L_4) = 0.007375$*, and* $\Delta(\texttt{Age}, L_5) = 0.010879$*.*

The Pearson's cumulative statistic is a well known measure, traditionally used in statistics for evaluating how much two probability distributions are similar. Given two distributions $P(X)$ and $P(X|y)$, their Pearson's cumulative statistic is defined as follows.

$$F(X, y) = \sum_{x \in X} \frac{(O_x^y - E_x)^2}{E_x}$$

where O_x^y is the frequency of value x for X with respect to y (i.e., the number of tuples in T such that $x = t[X]$ and $y = t[Y]$), and E_x is the expected frequency distribution of the same value x for X according to the baseline distribution $P(X)$.

Example 4.3 *Consider the distributions of the* `Age` *values for the different locations and the baseline distribution* $P(\texttt{Age})$ *in Figure 2(b). We have:*

$$F(\texttt{Age}, L_1) = \frac{\left(O_{<18}^{L_1} - E_{<18}\right)^2}{E_{<18}} + \ldots + \frac{\left(O_{\geq 55}^{L_1} - E_{\geq 55}\right)^2}{E_{\geq 55}} = 104.532750$$

Similarly, we obtain: $F(\texttt{Age}, L_2) = 878.201780$*,* $F(\texttt{Age}, L_3) = 30.837391$*,* $F(\texttt{Age}, L_4) = 17.340740$*, and* $F(\texttt{Age}, L_5) = 39.875054$

The lower the distance between $P(X|y)$ and $P(X)$ in the released dataset, the more the correlation between variables X and Y has been obfuscated. To determine when the distance between the y-conditioned distribution $P(X|y)$ and the baseline distribution $P(X)$ can be considered significant (and then exploited to infer a possible dependency between X and Y), we can adopt either an *absolute* or a *relative* approach. The absolute approach compares the distance between $P(X|y)$ and $P(X)$ for each value y of Y with a fixed threshold. The relative approach compares instead the distance between $P(X|y)$ and $P(X)$ for a given value y, with the distances obtained for the other values of Y.

Both the strategy based on statistical independence and the strategy based on minimizing the distance between distributions described above for obfuscating the correlation between X and Y can be concretely applied through specific metrics. Before describing such metrics in the following section, it is important to note that an external observer can only see and learn the distribution of values computed on tuples that have been released. In the remainder of this paper, we will then use T_r to denote the set of tuples released to the external world at a given point in time, and P_r to denote the value distributions observable on T_r (in contrast to the P observable on T). The knowledge of an external observer includes the different observations $P_r(X|y)$ she can learn by collecting all the released tuples (i.e., T_r), and the baseline distribution $P(X)$ publicly available.

5 Statistical tests for assessing inference exposure

In this section, we describe four statistical tests that can be adopted for verifying whether the release of a set of tuples is safe, that is, a potential observer can neither identify the entities associated with a sensitive value (e.g., an observer cannot identify that L_2 is a headquarter), nor improperly associate sensitive values with released entities in the dataset (i.e., false positives). Figure 4 summarizes such tests, classifying them depending on the strategy they follow to obfuscate the dependency between statistical variables X and Y, as illustrated in Section 4.

The statistical tests described in this section are based on the definition of a metric to measure how much the release of a subset T_r of tuples of T is exposed to inferences (*inference exposure*), and on the computation of a threshold that this measure should not exceed to guarantee that the data release is safe. In the following, we define different properties that the released dataset should satisfy to guarantee that a potential observer cannot infer the existence of a dependency between the random variables X and Y.

5.1 Significance of the mutual information

This statistical test aims at ensuring that mutual information $I_r(X,Y)$ characterizing the released dataset T_r is *statistically not significant*. The rationale is that the mutual information between X and Y, as illustrated in Section 4, measures the average amount of knowledge about Y that an observer acquires looking at X (and vice-versa). In other words, the mutual information $I_r(X,Y)$ between X and Y quantifies the (linear or non linear) dependency between the considered statistical variables. When $I_r(X,Y)$ is close to zero an observer does not have enough confidence on the existence of a dependency between X and Y in the released dataset T_r. Hence, the observer cannot infer anything about the sensitive property $s(y)$ associated with a target y that belongs to the released dataset.

From a practical point of view, to verify when the release of a given subset T_r of T can be considered safe, it is sufficient to check whether the mutual information $I_r(X,Y)$ of T_r is below a predefined threshold I_{rc} close enough to zero. For instance, the release of a set T_r of tuples related to a subset of the soldiers in our running example does not disclose information on the dependency between `Age` and `Location` if $I_r(\texttt{Age},\texttt{Location}) < I_{rc}$. A safe release is formally defined as follows.

Definition 5.1 (Safe release w.r.t. Mutual Information – MIS) *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, T_r be a subset of tuples in T, and I_{rc} be the critical value for the mutual information. The release of T_r is* safe *iff $I_r(X,Y) < I_{rc}$.*

The problem becomes now how to compute I_{rc}. The solution we propose is based on the following property [7].

Property 5.2 *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, and T_r be a subset of tuples in T. Under the independence hypothesis between X and Y:*

$$2N_r \log(2) I_r(X,Y) \sim \chi^2((N_{X_r} - 1)N_{Y_r})$$

where $N_r = |T_r|$ is the number of released tuples, N_{X_r} is the number of values of X in T_r, and N_{Y_r} is the number of values of Y in T_r.

Property 5.2 states that under the hypothesis of independence between X and Y, $2N_r \log(2) I_r(X,Y)$ is asymptotically chi-square distributed with $(N_{X_r} - 1)N_{Y_r}$ degrees of freedom.[1]

[1] In [7] the mutual information was computed by comparing each y-conditioned distribution $P(X|y)$ with a sample distribution $P(X)$ estimated on the same dataset. Hence, the number of degrees of freedom was $(N_{X_r} - 1)(N_{Y_r} - 1)$. In this paper, the baseline distribution $P(X)$ is assumed to be known to the observer. Coherently, Property 5.2 is derived under the assumption that the observer tests the mutual information at hand by comparing it to the case where samples (x,y) are drawn from the distribution $P(X,Y) = P(X)P(Y)$. Then, the number of degrees of freedom increases to $(N_{X_r} - 1)N_{Y_r}$.

Example 5.3 *Figure 5 compares the distribution of the rescaled (by factor* $2N_r \log(2)$*, with* $N_r = 5000$*) mutual information* $I_r(\texttt{Age}, \texttt{Location})$ *of our dataset, with the chi-square distribution with* $(10-1)5 = 45$ *degrees of freedom, where* 10 *is the number of different values for attribute* `Age` *and* 5 *is the number of different values for attribute* `Location`*. The histogram in the figure has been obtained with* 10000 *Monte Carlo iterations, considering the baseline distribution* $P(\texttt{Age})$ *and the distribution* $P(\texttt{Location})$ *of the sensitive information of our running example. From the figure, it is easy to see that the approximation of our rescaled mutual information to the chi-square distribution nicely holds.*

Since, by Property 5.2, $I_r(X,Y)$ is distributed as a chi-square distribution with $(N_{X_r}-1)N_{Y_r}$ degrees of freedom, we propose to compute the critical value I_{rc} for the mutual information by selecting a *significance level* α (i.e., a residual probability) and imposing $P(I_r(X,Y) > I_{rc}) = \alpha$ (i.e., the probability that $I_r(X,Y)$ is greater than threshold I_{rc} should be equal to α). As a consequence, I_{rc} can be obtained by constraining $\int_0^{2N_r \log(2) I_{rc}} \chi^2[(N_{X_r}-1)N_{Y_r}](x)\mathrm{d}x = 1-\alpha$. The significance level α represents the confidence in the result of a statistical analysis. Indeed, the higher the value of α, the more restrictive the condition that a release must satisfy to be considered safe. In fact, a lower value for α represents a low probability of error in drawing conclusions starting from the mutual information measured on the data. The value of the significance level α must be chosen in such a way to limit the confidence that an observer can have in the test results, thus preventing the observer from exploiting this test for drawing inferences. For instance, if an observer can evaluate the statistical test with significance level $\alpha = 5\%$, the inference she can draw from the result obtained has a high probability of being right (i.e., a high mutual information is due to chance only in 5% of the cases). The value chosen for α by the data holder should then be higher than the risk that an observer is willing to take when trying to guess the sensitive property $s(y)$ of a target y in Y. If the cost of the observer for her attack is low (e.g., the observer is interested in detecting which location is a headquarter for curiosity), she will be probably willing to take a high risk of making a wrong guess and she will therefore choose a high significance level for her analysis. In this case, α should be high to guarantee a better protection of the sensitive property (e.g., 15%-20%). On the other hand, if the cost of an observer for her attack is high (e.g., the observer wants to destroy headquarters), she will be probably willing to take a low risk of error, and α could be lower, thus permitting the release of a larger subset of tuples (e.g., 5% represents the typical value adopted in statistical hypothesis testing). Since it is unlikely for the data holder to know the significance level considered by a possible observer in the analysis, the data holder should estimate it and choose a value for α trying to balance the need for data protection on one side and the need for data release on the other side. In fact, the released dataset is protected against those analyses that assume a risk of error lower than α.

Once the data holder has fixed the significance level and computed the critical value I_{rc} for the mutual

information, she can decide whether to release a tuple when its respondent requires it. Let T_r be a safe set of released tuples and t be a tuple in T that needs to be released. To decide whether to release t, it is necessary to check if the mutual information $I_r(X, Y)$ associated with $T_r \cup \{t\}$ is lower than critical value I_{rc}. If this is the case, tuple t can be safely released; otherwise tuple t cannot be released since it may cause leakage of sensitive information.

Example 5.4 *Consider the military dataset in Figure 2(a), the release of the subset T_r of tuples in Figure 6(a), and assume that the data holder chooses a significance level $\alpha = 20\%$. The mutual information $I_r(\texttt{Age}, \texttt{Location})$ of T_r is* 0.025522*, while the critical value I_{rc} is* 0.025527*. Since $I_r(\texttt{Age}, \texttt{Location}) < I_{rc}$, the release of T_r is safe.*

Consider the release of the whole dataset T in Figure 2(a), and assume that the data holder adopts a less restrictive significance level $\alpha = 5\%$. The mutual information $I(\texttt{Age}, \texttt{Location})$ of the whole dataset is 0.063285 *(see Example 4.1) and its critical value I_{rc} is* 0.004448*. Therefore, as expected, the release of the whole dataset is not safe.*

5.2 Significance of the distance between distributions

The evaluation of the significance of the distance between distributions aims at verifying whether there are specific targets in the released dataset that can be considered as *outliers*, that is, whose y-conditioned distribution is far from the expected distribution represented by the baseline $P(X)$. The rationale is that peculiarities of the y-conditioned distribution can be exploited for inferring the sensitive property $s(y)$. This statistical test, operating on the single values y of Y, works at a finer granularity level than the previous one, based on the mutual information.

As already noted in Section 4, a possible way for the data holder to verify whether the y-conditioned distribution presents some peculiarities consists in computing the Kullback-Leibler distance $\Delta_r(X, y)$ between the y-conditioned distribution $P_r(X|y)$ of the released dataset and the baseline distribution $P(X)$. Following an approach similar to that illustrated in Section 5.1, the disclosure of the sensitive property $s(y)$ can be prevented by ensuring that $\Delta_r(X, y)$ is *statistically not significant*, for all targets y in the released dataset.

From a practical point of view, we can verify if the release of a given subset T_r of T can be considered safe by checking whether the distance $\Delta_r(X, y)$ is smaller than a predefined threshold $\Delta_{rc}(y)$ for all targets y. A safe release is formally defined as follows.

Definition 5.5 (Safe release w.r.t. KL Distance – KLD) *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, T_r be a subset of tuples in T, and $\Delta_{rc}(y)$ be the critical value for $\Delta_r(X, y)$, for all values y of Y in T_r. The release of T_r is* safe *iff for all values y of Y in T_r, $\Delta_r(X, y) < \Delta_{rc}(y)$.*

According to Definition 5.5, if $\Delta_r(X, y) < \Delta_{rc}(y)$ for all released targets y, the release of T_r is safe. If there exists at least a target y' such that $\Delta_r(X, y') \geq \Delta_{rc}(y')$, the release of T_r is not safe and y' is considered exposed.

The approach we propose to compute threshold $\Delta_{rc}(y)$ is based on the observation that the mutual information $I_r(X, Y)$ is by definition equal to $\sum_{y \in Y} p(y)\Delta_r(X, y)$, and that Property 5.2 can be adapted for the Kullback-Leibler distance $\Delta_r(X, y)$ as follows.

Property 5.6 *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, y be a value of Y, and T_r be a subset of tuples in T. Under the independence hypothesis between X and Y:*

$$2N_r(y)\log(2)\Delta_r(X, y) \sim \chi^2(N_{X_r} - 1)$$

where $N_r(y)$ is the number of released tuples with $Y = y$, and N_{X_r} is the number of values of X in T_r.

Property 5.6 states that under the hypothesis of independence between X and Y, $2N_r(y)\log(2)\Delta_r(X, y)$ is asymptotically chi-square distributed with $(N_{X_r} - 1)$ degrees of freedom.

Example 5.7 *Figures 7(a)-(e) compare the distribution of the rescaled (by factor $2N_r(y)\log(2)$ with $N_r(L_1) = 1014$, $N_r(L_2) = 649$, $N_r(L_3) = 826$, $N_r(L_4) = 1003$, and $N_r(L_5) = 1506$) Kullback-Leibler distance $\Delta_r(\texttt{Age}, L_i)$, $i = 1, \ldots, 5$, with the chi-square distribution with $10 - 1 = 9$ degrees of freedom. The histograms in the figures have been obtained with 10000 Monte Carlo iterations, considering the baseline distribution $P(\texttt{Age})$ and the distribution $P(\texttt{Location})$ of the sensitive information of our running example. From the figures, it is easy to see that our rescaled $\Delta_r(\texttt{Age}, L_i)$ fit the considered chi-square distribution.*

For each target y, Property 5.6 can be used to compute the critical value $\Delta_{rc}(y)$ for $\Delta_r(X, y)$ by selecting a *significance level* α and requiring $P(\Delta_r(X, y) > \Delta_{rc}(y)) = \alpha$. As a consequence, $\Delta_{rc}(y)$ can be obtained by constraining $\int_0^{2N_r(y)\log(2)\Delta_r(X,y)} \chi^2(N_{X_r} - 1)(x)\mathrm{d}x = 1 - \alpha$. As already observed for the mutual information, higher values of α guarantee better protection against inference exposure of the sensitive property.

Once the data holder has fixed the significance level and computed the critical values $\Delta_{rc}(y)$ for each target y, she can decide whether to release a tuple when its respondent requires it. Let T_r be a safe set of released tuples and t be a tuple in T whose release has been requested. To decide whether to release t, it is necessary to check if the distance $\Delta_r(X, y)$ for target $y = t[Y]$, computed on $T_r \cup \{t\}$, is lower than the critical value $\Delta_{rc}(y)$. If such a control succeeds, the release of t, that is, the disclosure of $T_r \cup \{t\}$, is considered safe. Otherwise, target y is considered exposed (i.e., y is an outlier) and the release of t is blocked. Note that condition $\Delta_r(X, y) < \Delta_{rc}(y)$ is certainly satisfied for all the targets different from $t[Y]$ because T_r is assumed to be safe.

Example 5.8 *Consider the military dataset in Figure 2(a) and the release of the subset T_r of tuples in Figure 8(a), and assume that the data holder adopts a significance level α=20%. The distances between each*

L_i-conditioned distribution $P_r(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, and the baseline distribution $P(\texttt{Age})$ are: $\Delta_r(\texttt{Age}, L_1) = 0.026582$, $\Delta_r(\texttt{Age}, L_2) = 0.056478$, $\Delta_r(\texttt{Age}, L_3) = 0.028935$, $\Delta_r(\texttt{Age}, L_4) = 0.029818$, and $\Delta_r(\texttt{Age}, L_5) = 0.014996$. The critical values are: $\Delta_{rc}(L_1) = 0.026599$, $\Delta_{rc}(L_2) = 0.057343$, $\Delta_{rc}(L_3) = 0.028954$, $\Delta_{rc}(L_4) = 0.029834$, and $\Delta_{rc}(L_5) = 0.015018$. Since the distance $\Delta_r(\texttt{Age}, L_i)$ computed for each location L_i, $i = 1, \ldots, 5$, is lower than the corresponding critical value, the release of T_r is safe.

Consider the release of the whole dataset T in Figure 2(a) and assume that the data holder adopts a less restrictive significance level α=5%. The distances between each L_i-conditioned distribution and the baseline distribution are: $\Delta(\texttt{Age}, L_1) = 0.047349$, $\Delta(\texttt{Age}, L_2) = 0.358836$, $\Delta(\texttt{Age}, L_3) = 0.013967$, $\Delta(\texttt{Age}, L_4) = 0.007375$, and $\Delta(\texttt{Age}, L_5) = 0.010879$ (see Example 4.2). Their critical values are: $\Delta_{rc}(L_1) = 0.006015$, $\Delta_{rc}(L_2) = 0.009395$, $\Delta_{rc}(L_3) = 0.007388$, $\Delta_{rc}(L_4) = 0.006081$, and $\Delta_{rc}(L_5) = 0.004051$. Since the distance $\Delta(\texttt{Age}, L_i)$ of each location L_i, $i = 1, \ldots, 5$, exceeds the corresponding critical value, the release of T is, as expected, not safe.

By comparing the two metrics discussed so far, it is easy to see that the metric based on the mutual information does not distinguish the exposures of the different targets. Hence, if for a given y, $p_r(y)$ represents a small portion of the released dataset, a high value for $\Delta_r(X, y)$ has a limited influence on the decision of whether the release of T_r is safe or not, since the contribution of $\Delta_r(X, y)$ in the computation of $I_r(X, Y)$ is limited. On the contrary, the test based on the Kullback-Leibler distance results more restrictive than the evaluation of the significance of the mutual information since the safety control is performed at the level of each single target y of Y.

5.3 Chi-square goodness-of-fit test

The *chi-square goodness-of-fit* test aims at verifying, like the statistical test described in Section 5.2, whether the released dataset includes a target y that can be considered an *outlier*. The chi-square goodness-of-fit test [32] is a well known statistical test, traditionally used to determine whether a probability distribution ($P_r(X|y)$) fits into another (theoretical) probability distribution ($P(X)$), that is, if the two probability distributions are similar. The test is based on the computation of Pearson's cumulative statistic $F_r(X, y)$ that measures how "close" the observed y-conditioned distribution $P_r(X|y)$ is to the expected (baseline) distribution $P(X)$. When $F_r(X, y)$ is close to zero, $P_r(X|y)$ appears as a distribution that fits $P(X)$ (i.e., the values of $P_r(X|y)$ appear as randomly extracted from the baseline distribution $P(X)$) and therefore nothing can be inferred about the sensitive property $s(y)$ associated with target y.

From a practical point of view, we verify if the release of a given subset T_r of T can be considered safe by checking whether the Pearson's cumulative statistic $F_r(X, y)$ is smaller than a predefined threshold F_{rc}. Formally, a safe release is defined as follows.

Definition 5.9 (Safe release w.r.t. Chi-Square Goodness-of-Fit – CST) *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, T_r be a subset of tuples in T, and F_{rc} be the critical value for $F_r(X, y)$. The release of T_r is* safe *iff for all values y of Y in T_r, $F_r(X, y) < F_{rc}$.*

According to Definition 5.9, if all the released targets y satisfy condition $F_r(X, y) < F_{rc}$, the release of T_r is safe; if there exists at least a target y' that violates the condition, the release of T_r is not safe and y' is considered exposed.

The threshold F_{rc} is computed by exploiting the following statistical property enjoyed by the chi-square goodness-of-fit test [32].

Property 5.10 *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, y be a value of Y, and T_r be a subset of tuples in T. Under the independence hypothesis between X and Y:*

$$F_r(X, y) = \sum_{x \in X} \frac{(O_x^y - E_x)^2}{E_x} \sim \chi^2(N_{X_r}(y) - 1)$$

where $N_{X_r}(y)$ is the number of values of X for the tuples in T_r with $Y = y$.

Property 5.10 states that, under the hypothesis of independence between X and Y, the Pearson's cumulative statistic $F_r(X, y)$ is asymptotically chi-square distributed with $(N_{X_r}(y) - 1)$ degrees of freedom. Like for the metrics already discussed, we compute the critical value $F_{rc}(y)$ for the Pearson's cumulative statistic by selecting a *significance level* α and requiring $P(F_r(X, y) > F_{rc}(y)) = \alpha$. As a consequence, $F_{rc}(y)$ can be obtained by constraining $\int_0^{\sum_{x \in X} \frac{(O_x^y - E_x)^2}{E_x}} \chi^2(N_{X_r}(y) - 1)(x)\mathrm{d}x = 1 - \alpha$. It is important to note that the number of degrees of freedom of the chi-square distribution depends on the number N_{X_r} of values of variable X that have been released for target y, which may be different from the number of values in the domain of attribute X (for more details see Section 6).

Once the data holder has fixed the significance level and computed the critical value F_{rc}, she can decide whether to release a tuple when its respondent requires it. Let T_r be a safe set of tuples and t be a requested tuple in T. To evaluate whether the release of tuple t is safe, it is necessary to check whether the Pearson's cumulate statistic $F_r(X, y)$ for target y=$t[Y]$, computed on $T_r \cup \{t\}$ is lower than the fixed threshold F_{rc}. If this is the case, tuple t can be safely released; otherwise the release of t is blocked since it reveals that y is an outlier. We note that it is not necessary to check the Pearson's cumulate statistics of the other targets in T_r, since they are not affected by the release of t, and their associated $F_r(X, y)$ are lower than F_{rc}, as T_r is supposed to be safe.

Example 5.11 *Consider the military dataset in Figure 2(a) and the release of the subset T_r of tuples in Figure 9(a) and assume that the data holder adopts a significance level α=20%. The Pearson's cumulative statistics for the five locations are: $F_r(\texttt{Age}, L_1) = 8.550683$, $F_r(\texttt{Age}, L_2) = 0.961415$, $F_r(\texttt{Age}, L_3) = 9.717669$,*

$F_r(\texttt{Age}, L_4) = 8.293681$, *and* $F_r(\texttt{Age}, L_5) = 8.554984$. *The critical values are:* $F_{rc}(L_1) = 8.558059$, $F_{rc}(L_2) = 1.642374$, $F_{rc}(L_3) = 9.803249$, $F_{rc}(L_4) = 11.030091$, *and* $F_{rc}(L_5) = 8.558059$. *It is immediate to see that* $F_r(\texttt{Age}, L_i) < F_{rc}(L_i)$, *for all* $i = 1, \ldots, 5$. *As a consequence, the release of* T_r *is safe.*

Consider the release of the whole dataset T *in Figure 2*(*a*) *and assume that the data holder adopts a less restrictive significance level* α=5%. *The Pearson's cumulative statistics for the five locations are:* $F(\texttt{Age}, L_1) = 104.532750$, $F(\texttt{Age}, L_2) = 878.201780$, $F(\texttt{Age}, L_3) = 30.837391$, $F(\texttt{Age}, L_4) = 17.340740$, *and* $F(\texttt{Age}, L_5) = 39.875054$ (*see Example 4.3*). *The critical values are:* $F_{rc}(L_1) = 15.507313$, $F_{rc}(L_2) = 16.918978$, $F_{rc}(L_3) = F_{rc}(L_4) = F_{rc}(L_5) = 15.507313$. *Therefore,* $P(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, *is not close enough to* $P(\texttt{Age})$ *and the release of the whole dataset is not safe. This result is not surprising since none of the* L_i*-conditioned distribution* $P(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, *in our running example exactly fits the baseline distribution* $P(\texttt{Age})$.

5.4 Dixon's Q-test

The Dixon's Q-test, similarly to the statistical tests described in Section 5.2 and Section 5.3, aims at verifying whether there is one target in the released dataset that can be considered an *outlier*. The Dixon's Q-test is a well-known solution for outlier detection in a given dataset that can be adopted whenever there is at most one outlier and at least three targets in the considered dataset [15]. This statistical test differs from the ones illustrated in Section 5.2 and Section 5.3 since, instead of comparing each distance between $P_r(X|y)$ and $P(X)$ against a fixed threshold, it evaluates if one of the distances between $P_r(X|y)$ and $P(X)$ is significantly higher than the others. The Dixon's Q-test can be applied considering any definition of distance between distributions (e.g., Kullback-Leibler distance, or Pearson's cumulative statistic). In line with the rest of the paper, we apply the Dixon's Q-test to the Kullback-Leibler distance $\Delta_r(X, y)$ between $P_r(X|y)$ and $P(X)$. We note that different versions of this test have been proposed in the literature, and we adopt r_{10} [15]. This test assumes the presence of at most one outlier at the upper hand of the dataset (i.e., one outlier characterized by a high value for the distance between distributions) and no outlier at the lower hand of the dataset (i.e., no outlier is characterized by a low distance between distributions).

The Dixon's Q-test requires to first organize the values on which it needs to be evaluated (i.e., $\Delta_r(X, y)$ in our scenario) in ascending order. Starting from the last two values in the ordered sequence (i.e., the two highest values), it computes coefficient $Q_r(X)$ as their relative distance. More formally, Dixon's coefficient is computed as:

$$Q_r(X) = \frac{\Delta_r(X, y_n) - \Delta_r(X, y_{n-1})}{\Delta_r(X, y_n) - \Delta_r(X, y_1)},$$

where $\Delta_r(X, y_1)$,$\ldots$,$\Delta_r(X, y_n)$ is the sequence, in ascending order, of distance values.

The Dixon's Q-test is not able to identify any outlier in the dataset if $Q_r(X)$ is close enough to zero, since

the distance between each pair of subsequent values in the sequence is almost the same. In this case, there is no target y such that the distance between its y-conditioned distribution $P_r(X|y)$ and the baseline $P(X)$ stands out from the other distances.

From a practical point of view, we verify if the release of a given subset T_r of T can be considered safe by checking whether the Dixon's coefficient $Q_r(X)$ is smaller than a predefined threshold Q_{rc}. The critical value Q_{rc} is computed by fixing a *significance level* α and imposing $P(Q_r(X) > Q_{rc}) = \alpha$. Figure 10 summarizes the critical values Q_{rc} when the number of distinct values in the domain of Y ranges between 3 and 10 and the significance level is fixed to 20%, 10%, 5%, and 1%, respectively. If $Q_r(X) < Q_{rc}$, the release of T_r does not reveal the presence of any outlier and the release of T_r is *safe*. A safe release is formally defined as follows.

Definition 5.12 (Safe release w.r.t. Dixon's Q-test – DQT) *Let T be a table over attributes A, X and Y be two subsets of A such that $X \rightsquigarrow Y$, T_r be a subset of tuples in T, and Q_{rc} be a critical value for $Q_r(X)$. The release of T_r is* safe *iff $Q_r(X) < Q_{rc}$.*

If condition $Q_r(X) < Q_{rc}$ does not hold, an observer can infer that the target y characterized by the maximum distance $\Delta_r(X, y)$ between $P_r(X|y)$ and $P(X)$ is an outlier.

Once the data holder has fixed the significance level and computed the critical value Q_{rc} for the Dixon's Q-test, she can decide whether to release a tuple when its respondent requires it. Let T_r be a safe set of released tuples and t be a requested tuple in T. To decide whether to release t, it is necessary to check if Dixon's coefficient $Q_r(X)$ associated with $T_r \cup \{t\}$ is lower than critical value Q_{rc}. If this is the case, tuple t can be safely released; otherwise tuple t is not released since it may cause leakage of sensitive information.

Example 5.13 *Consider the military dataset in Figure 2(a) and the release of the subset T_r of tuples in Figure 11(a), and assume that the data holder adopts a significance level $\alpha = 20\%$. The distance values between $P_r(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, and the baseline $P(\texttt{Age})$ are equal to: $\Delta_r(\texttt{Age}, L_1) = 0.209188$, $\Delta_r(\texttt{Age}, L_2) = 0.361504$, $\Delta_r(\texttt{Age}, L_3) = 0.037932$, $\Delta_r(\texttt{Age}, L_4) = 0.018421$, and $\Delta_r(\texttt{Age}, L_5) = 0.021103$. To apply the Dixon's Q-test, these distance values are considered in ascending order and then the Dixon's coefficient is computed as $Q_r(X) = \frac{0.361504 - 0.209188}{0.361504 - 0.018421} = 0.443963$. Since attribute* `Location` *has 5 distinct values in its domain, we consider the third column in the table in Figure 10 for the definition of critical value Q_{rc}. In particular, the critical value is fixed to 0.451 for the considered significance level. Since Dixon's coefficient is lower than the critical value, the release of T_r is safe.*

Consider the release of the whole dataset T in Figure 2(a) and assume that the data holder adopts a less restrictive significance level $\alpha = 5\%$. The distance values in Example 4.2 are considered in ascending order and Dixon's coefficient is computed as $Q_r(X) = \frac{0.358836 - 0.047349}{0.358836 - 0.07375} = 0.886263$, which is greater than 0.642. Therefore, the release of the whole dataset of our running example is not safe, since it discloses that L_2 is an outlier.

6 Controlling exposure and regulating releases

We now illustrate how the incremental release of tuples is controlled and regulated according to the metrics discussed in the previous section.

The data holder first chooses the metric and the significance level α she wants to adopt. Every time a tuple t is requested, it is necessary to check if the release of t, combined with all the tuples already released and potentially known to an observer T_r, may cause the unintended disclosure of sensitive information. In particular, if $T_r \cup \{t\}$ satisfies the definition of safe release for the considered metric (see Section 5), t is released. If tuple t cannot be released when it is requested, its release might simply be denied. However, this choice represents a restrictive solution, since it does not take into consideration the fact that if a tuple cannot be released when it is requested, it may be safely released at a later time (i.e., after the release of other tuples in the dataset). Indeed, the grant or denial of the release of a tuple depends on the set of tuples that has already been released. Exploiting this observation, we propose to insert the tuples that cannot be released when requested into a queue. Every time a tuple t is released, the tuples in the queue are analyzed to check whether a subset of them can be safely released.

Particular attention has to be payed on the release of the first few tuples because they will produce random value distributions that usually do not resemble the actual distributions existing in the dataset. Such random distributions may characterize the data release as not safe, thus blocking any further release and raising many false alarms (since also targets that are not outliers will have a random initial distribution that will differ from the baseline). However, no observer could put confidence on statistics computed over a few releases as they cannot be considered accurate and their distribution can be completely random. With reference to the release of the first few tuples, it is also important to note that the metrics illustrated in Section 5 are based on approximation properties that hold only when a sufficient number of tuples has been released. There is therefore a starting time at which the data holder should define an alternative condition for determining if a release should be considered safe. In the following we discuss, for each of the metrics in Section 5, how to check whether the release of a tuple t is safe when only few tuples have been released.

Significance of the mutual information and significance of the Kullback-Leibler distance between distributions. The definition of the critical value for the mutual information described in Section 5.1 is based on Property 5.2, which is an asymptotic approximation of $I_r(X, Y)$ to a chi-square distribution that holds only if a sufficient number of tuples has been released. Using the traditional Monte Carlo approach, we propose to compute the critical value of the mutual information for the release of a small number n of tuples as the α-th percentile of the mutual information obtained by extracting a sufficient number of samples (10000 in

our experimental evaluation) of n tuples each from a simulated dataset composed of $|T|$ tuples, where X is distributed following $P(X)$, and X and Y are statistically independent. Indeed, if the mutual information of the released dataset is close to the mutual information of a sample of the same size extracted from a dataset where X and Y are statistically independent, the observer cannot exploit the released tuples for drawing inferences. The remaining aspect to consider is when to start adopting the critical value computed exploiting Property 5.2. A nice approximation is represented by $2N_XN_Y$ tuples (100 in our example), which is confirmed by our experimental evaluation illustrated in Figure 12. In this figure, the curve representing the critical value for the mutual information, corresponding to the value computed through the Monte Carlo method in the interval [0-100] and exploiting Property 5.2 in interval [100-10000], presents a smooth trend. This result also confirms that Property 5.2 holds in our framing of the problem.

The same approach can be adopted for the metric based on the Kullback-Leibler distance since Property 5.6 derives from Property 5.2, and the mutual information is a weighted average of the Kullback-Leibler distances for the different targets y in the dataset.

Chi-square goodness-of-fit test. The approximation on which this statistical test is based holds on a data collection only if, for each target y and for each $x \in X$, a sufficient number of tuples (typically 5 [32]) has been released. In other words, considering a target y, for each $x \in X$, there must be at least 5 tuples in T_r with $t[Y] = y$ and $t[X] = x$. If, for a given target y, there are less than 5 tuples with value x for attribute X, we can combine x with either its preceding or subsequent value in the domain of X and sum their relative frequencies. With reference to our example, if only 2 soldiers located at L_2 in the age range [20-24] have been released, range [20-24] for L_2 can be combined either with [18-19] or with [25-29] for the same location. Suppose now that the relative frequency for age range [25-29] is 4. By merging [20-24] with [25-29] for location L_2, we obtain a new value [20-29] of the domain of attribute `Age` for location L_2, with relative frequency equal to 6. This process is iteratively applied, possibly combining a set of contiguous values for attribute X, until all the relative frequencies of the values in the domain of X are greater than or equal to 5. If all the values in X are combined in a unique value, the test cannot be applied and the release is considered safe. If at least 2 values in the domain of X are maintained, the test can be evaluated. We note however that when multiple original values of X are combined, the approximation in Property 5.10 should be revised to consider the correct number of degrees of freedom, which is equal to the number of values in the domain of X in T_r after the possible merge operation. For instance, with reference to our example, suppose that the values for attribute `Age` for location L_2 have been combined obtaining the following domain values: ≤ 24, $[25-39]$, $[40-44]$, $[45,49]$, ≥ 50. The critical value of Pearson's cumulative statistic for L_2 should be computed considering a chi-square distribution with 4 (instead of 9) degrees of freedom.

Dixon's Q-test. As already noted, this statistical test can be applied only on data collections that include at least 3 elements [15]. In our scenario, it can then be used only if 3 different distances between the y-conditioned distributions and the baseline can be computed. Consequently, datasets with less than 3 different distance values are considered safe since an observer could not gain any information.

7 Experimental results

To evaluate the behavior of the metrics presented in Section 5, we implemented the data release strategy described in Section 6 with a Matlab prototype and executed a series of experiments. For the experiments, we considered the dataset T introduced in Example 3.2, which has been obtained by randomly extracting 10000 tuples from the baseline distribution $P(\texttt{Age})$ of the age of soldiers of the UK Regular Forces as at 1 April 2006 [38] (Figure 3(a)). The experiments evaluated the inference exposure (computed as the mutual information, Kullback-Leibler distance between distributions, Pearson's cumulative statistic, or Dixon's coefficient), and the information loss (i.e., the number of tuples not released upon request) caused by our privacy protection technique. We also compared the results obtained adopting the different metrics.

7.1 Inference exposure

We evaluated how the metrics discussed in Section 5 vary with the release of tuples and compared them with the corresponding critical values. The experiments have been conducted on 20 randomly extracted sequences of 10000 requests each. For the sake of readability, in this section we illustrate the graphs showing the evolution of the inference exposure and of its critical value for one of the 20 sequences; the results obtained with the other sequences present a similar trend.

Mutual information. Figure 12 shows the evolution of both the mutual information, and the corresponding critical value, varying the number of released tuples (the scale of the axis in Figure 12 is logarithmic). The two curves are close to each other and their distance decreases as the number of released tuples increases. It is easy to see that the mutual information of released data is always lower than the critical value. The figure also shows a smooth trend for the curve representing the critical value, confirming that the approximation in Property 5.2 nicely holds in our scenario. In fact, the discontinuity in the critical value of the mutual information when the 100th tuple is released, due to the fact that the critical value is computed using the Monte Carlo based approach in the interval [1-100] and the approach using Property 5.2 in the interval [100-10000], is small and cannot be noticed in the figure.

Kullback-Leibler distance. Figures 13(a)-(e) show the evolution of both the Kullback-Leibler distance between $P_r(\texttt{Age}|L_i)$ and $P(\texttt{Age})$, $i = 1, \ldots, 5$, and the corresponding critical values, varying the number of released tuples (the scale of the axis in Figures 13(a)-(e) is logarithmic). It is not surprising that the trends shown in these figures are similar to that illustrated in Figure 12. Indeed, the mutual information is the weighted average of the Kullback-Leibler distance values of all the locations in the dataset. It is interesting to note that all the locations present a similar trend for the evolution of both the Kullback-Leibler distance and its critical value. Also, like for the mutual information, Figures 13(a)-(e) present a smooth trend in the curves representing the critical values for the five locations, confirming that the approximation in Property 5.6 holds. In fact, the discontinuity in the critical value of the Kullback-Leibler distance when the 100th tuple is released cannot be noticed from the figure.

Chi-square goodness-of-fit. Figures 14(a)-(e) show the evolution of both the Pearson's cumulative statistic of each location, and the corresponding critical values, varying the number of released tuples. As discussed in Section 5.3, when a sufficient number of tuples have been released the critical value F_{rc} is the same for all the locations. On the contrary, when a limited number of tuples have been released, the critical value may be different for each location, depending on the number of distinct values in the domain of attribute X for each location. As it is visible from Figure 14, the curve representing the critical value has different steps. Each step corresponds to a change in the number of values in the domain of X and therefore a different (higher) number of degrees of freedom of the chi-square distribution in Property 5.10. When the number of released tuples does not permit to correctly evaluate if the Chi-square goodness-of-fit test is passed or not, the release is considered safe since an observer cannot gain knowledge by looking at the released data. This is the reason why the Pearson's cumulative statistic and its critical value are not computed for the first few (about 10) released tuples in Figures 14(a)-(e). For all the locations, the value of the Pearson's cumulative statistic increases while tuples are released. In particular, this growing trend is more visible when less than 100 tuples have been released. Also in this case, as expected, the distance between the Pearson's cumulative statistic and its critical value decreases while data are released.

Dixon's Q-test. Figure 15 shows the evolution of both the Dixon's coefficient and the corresponding critical value, varying the number of released tuples. The distance between Dixon's coefficient and the critical value decreases while tuples are released. As it is visible from Figure 15, the Dixon's coefficient and its critical value are not reported for the first 5 tuples released. This is due to the fact that, for the first 5 tuples, it is not possible to compute 3 different distance values between y-conditioned distributions and the baseline. The curve representing the critical value presents three steps. Each step corresponds to the release of a tuple that permits

to compute an additional difference. In other words, it corresponds to the release of a tuple t such that $t[Y]$ is a target that either was not represented in T_r or that was characterized by a distance from the baseline equal to the distance of another target.

We note that, for all the considered metrics, the distance between the exposure and its critical value decreases as more data are released, since the fluctuations in the value distribution characterize the release of the first few tuples. In fact, as the number of tuples in the released dataset increases, the impact of the release of a single tuple on the distribution of released values decreases.

7.2 Information loss

To evaluate the quality of the results obtained adopting our metrics, we consider the number of released and discarded tuples. Figures 16(a)-(b) summarize the average number of tuples released by each of our metrics with significance level α equal to 20% and 5%, respectively, for the 20 sequences of 10000 requests that we generated for our experiments, distinguishing also how many requests for each location have been fulfilled.

Comparing the results in Figures 16(a)-(b) we note that, as expected, a lower significance level permits to release a higher number of tuples for all the considered metrics. Indeed, most of the cells in the table in Figure 16(b) have higher values than the corresponding cells in Figure 16(a). It is also easy to see that there is not a metric that is always better than the others in terms of the number of tuples released. For instance, Dixon's Q-test is less restrictive that the other metrics, since it releases the highest number of tuples as a whole and for each locations when $\alpha = 20\%$, and as a whole and for each locations but L_3 when $\alpha = 5\%$. From our analysis of the results reported in the two tables, we can conclude that the considered metrics adopt a different approach to protect the released data: CST and KLD block the release of the tuples of the outlier, while MIS and DQT block the release of the tuples from all the locations.

The location with the fewest released tuples is L_2 for both MIS and CST metrics, and for DQT in the case $\alpha = 20\%$. This is a non-surprising result, since L_2 is the headquarter (i.e., the outlier that needs to be protected). On the contrary, metric KLD blocks more tuples from L_1 than from L_2, and DQT, for $\alpha = 5\%$, blocks more tuples from location L_3 than from L_2. The location that enjoys the largest number of tuples released with $\alpha = 20\%$ is L_3 for all the metrics but DQT, which privileges location L_5. With $\alpha = 5\%$, the location with the highest percentage of released tuples is L_4 for all the metrics but MIS, which privileges location L_3.

It is interesting to note that all the metrics proposed in this paper to evaluate if a release is safe permit to release a considerable number of tuples, especially if compared with the (more intuitive) approach of *fitting the baseline distribution* within each L_i-conditioned distribution. Fitting the baseline within an L_i-conditioned distribution forces a maximum number of tuples that could be released for each age range in L_i, since the relative

frequency of the tuples in each age range must be exactly that of the baseline for each location in the released dataset. For instance, in the baseline distribution almost 19.67% soldiers are in the range [25-29], while in L_2 only 8.78% of tuples (140 tuples) fall in such range. Respecting the baseline distribution requires, even in the case where all tuples in the range [25-29] of L_2 are released to not release tuples in other ranges (so that the 140 tuples above actually correspond to 19.67%). Figure 17 graphically depicts this reasoning of fitting the baseline distribution (in black) within the L_2-conditioned distribution (gray going over the black). For each value range, no more than the number reached by the baseline distribution should be released. Figure 18 summarizes the number of tuples for each location that would be released adopting the approach of fitting the baseline within each L_i-conditioned distribution, $i = 1, \ldots, 5$. It is easy to see that this approach is far more restrictive than our solution and blocks the release of a larger number of tuples. Each of the proposed metrics permits to release a higher number of tuples for most of the locations (but for CST in the case of location L_4 with $\alpha = 20\%$ and L_3 with $\alpha = 5\%$). In particular, our approach permits to release in most cases more than twice the number of tuples that would be released by fitting the baseline distribution within each L_i-conditioned distribution. This is mainly due to the fact that, when fitting the baseline within each $P(\texttt{Age}|L_i)$, the presence of a low number of tuples in an age-range for a location (e.g., 2 soldiers with age greater than 55 in L_3, L_4, and L_5) hardly constraints the release of the tuples in all the other age ranges. In our example, the two tuples representing soldiers older than 55 must represent the 0.21% of all the tuples released for locations L_3, L_4, and L_5. As a consequence, the data holder can release at most 952 tuples of L_3, L_4, and L_5. Our metrics try to loosen this constraint, by evaluating the distance (or its average) between the distributions, instead of the value that the distribution has at each age value.

7.3 Comparison

To further compare the behavior of the metrics proposed, we have randomly generated 100 request sequences of 5000 tuples each, out of the 10000 in our dataset of the UK Regular Forces. For each of the metrics proposed in the paper, and for each of the 100 random request sequences, we run our algorithm. For this series of experiments, we fixed the significance level α to 20%, which represents the most restrictive release scenario. We then checked, for each of the metrics, how many of the 100 safe releases obtained running our algorithm with the considered metric represents a safe release also with respect to each of the other three metrics. Figure 19 summarizes the number of datasets obtained adopting each metric (on the row) that are safe also with respect to the other metrics (on the column). It is immediate to see that DQT is the less restrictive metric, confirming the results illustrated in the previous subsection. In fact, none of the 100 datasets obtained adopting DQT metric is safe with respect to the other three metrics (fourth row in Figure 19). On the contrary, 54 (61

and 45, respectively) datasets obtained using MIS metric (KLD and CST metrics, respectively) also satisfy the definition of safe release of Dixon's Q-test. The most restrictive metric is instead KLD, since no dataset obtained adopting a different metric resulted safe with respect to KLD metric (second column in Figure 19) while at least one dataset obtained adopting KLD metric is safe with respect to each of the other three metrics (second row in Figure 19). It is interesting to note that this result is different from the conclusions drawn in the previous subsection, where we noted that MIS and CST are the metrics that minimize the release of tuples. It is however not surprising since the analysis illustrated in Figure 19 is different from the one summarized in Figures 16(a)-(b). In fact, the results illustrated in Figure 19 are obtained analyzing a dataset that is considered safe by one metric with respect to the other metrics introduced in Section 5. On the contrary, the results in Figures 16(a)-(b) are obtained analyzing the safe datasets produced by each of the metrics of interest, starting from the same original data collection and considering the same order in the request of tuple. The results in Figure 19 confirm the fact that the considered metrics measure the exposure of the released dataset in different ways and that the considered metrics obtain a different result if applied to the same sequence of tuple requests. Each metric is therefore suited for protecting a different statistical characteristic of the data that could be exploited for inference purposes. For instance, MIS metric is the ideal solution to protect the released data against attacks that exploit the mutual information between X and Y (i.e., their statistical dependency) to gain information about the sensitive property. To decide the metric and the value for α to be adopted for protecting the release of her dataset, the data holder needs to estimate the attacks that a possible observer could exploit to gain sensitive information. If the data holder wants to achieve a higher protection for her data, she can combine (a subset of) the metrics introduced in Section 5. This approach, while better preserving privacy of sensitive data, has the drawback of limiting the number of tuples released, since the released dataset must satisfy all the conditions in Figure 4 (or a subset thereof). Analogously, to take a safe approach, the data holder can choose a high value for the significance level.

8 Related work

Several research efforts have been recently dedicated to the problem of protecting privacy in data publication (e.g., [9, 18, 27, 34]). In particular, considerable attention has been devoted to the problem of protecting respondents' identities and the sensitive information associated with them. Most of these proposals use the notion of k-anonymity [34] as a starting point or adopt some extensions of k-anonymity (e.g., [18, 25, 27, 29]), while others are based on the idea of fragmenting data and publishing associations at the group level (e.g., [13, 41]). Among them, t-closeness [27] and (α_i, β_i)-closeness [18] present some similarities with our work. t-closeness protects attribute disclosure by imposing that the distribution of sensitive values in the equivalence classes of

the released table (i.e., in the groups of tuples with the same value for the quasi-identifying attributes) must be similar to the distribution in the private table. To this purpose, t-closeness approach adopts the Earth Mover's Distance (EMD) for measuring the distance between the global distribution computed on the whole private table and the distributions computed within each equivalence class. The distance between these distributions should be no more than t. In [18], the authors present an extension of t-closeness that overcomes some of its limitations (e.g., the difficulty in choosing a correct value for t and the impossibility to specify that some attribute values are more sensitive than others). With this approach, the data publisher defines a different range $[\alpha_i, \beta_i]$ for each value v_i of a sensitive attribute. A table can then be released if, for each equivalence class, the fraction of tuples in the class with a given sensitive value v_i falls in the corresponding range $[\alpha_i, \beta_i]$. Although our proposal and these two approaches have in common the fact that they consider inference issues caused by anomalous value distributions, our work addresses a different and more complex scenario characterized by incremental releases of detailed data. Also, in our scenario the sensitive information is not released but can be inferred due to a value distribution dependency between a set of attributes appearing in the released dataset and the sensitive property itself.

The problem and scenario we consider resemble the scenarios where data are continuously generated and may need to be immediately released (e.g., [26, 40, 42]). In this case, data have to be timely released without violating the privacy of the individuals to whom they refer. The solutions proposed are typically based on the generalization of the data to be released coupled with the introduction of a limited delay in data publication (e.g., [40, 42]), or on the addition of noise (e.g., [26]). Such solutions share with our work the need of incrementally releasing data in a way that sensitive information is properly protected. Our work however aims at avoiding inferences from the released data in contrast to the protection of respondents' identities, and does not allow the use of generalization but requires the release of detailed data.

Inference problems have been extensively studied in the context of multilevel database systems (e.g., [11, 24, 28, 30]). Most inference research addresses detection of inference channels within a stored database or at query processing time. In the first case, inference channels are removed by upgrading selected schema components or redesigning the schema (e.g., [33]). In the second case, database transactions are evaluated to determine whether they lead to illegal inferences and, if so, deny the query (e.g., [21, 23, 31, 36]). None of these approaches is however applicable to the problem under consideration. As a matter of fact, the inference problem we address is due to a dependency existing between the value distributions observable aggregating all the released tuples and the sensitive information that we want to protect. Previous work on inference focuses instead on locating inference channels based on semantic relationships among attributes or on queries submitted to the system.

Our problem has also common aspects with the problem that arises when the aggregation of two or more data items is considered more sensitive than the data items singularly taken. A well-known example is the Secret

Government Agency (SGA) Phonebook [35]: the entire phonebook is classified as confidential and it is accessible only by users with the appropriate clearance but single entries are unclassified and available to any requester. Although our problem is conceptually similar, the classical solutions developed for addressing the aggregation problem (e.g., [14, 22, 24]) are not directly applicable in our context. These approaches define a threshold on the amount of data that can be released to each user and focus on maintaining history and establishing how to control collusion among users.

Other related proposals are those used to assess the *interestingness* of association rules in knowledge discovery problems. In [37], the authors introduced the J-measure to assess the relevance of an association rule. In some sense, these proposals are complementary to our work, as they can be used for assessing dependencies among the attributes characterizing a data collection. The information they produce can then be used as input to our approach for the definition of appropriate dependencies.

9 Conclusions

We considered the problem of protecting sensitive information in an incremental data release scenario, where the data holder releases non sensitive data on demand. As more and more data are released, an external observer can aggregate such data and infer the sensitive information by exploiting a dependency between the distribution of the non sensitive released data and the sensitive information itself. In this paper, we presented an approach for characterizing when data can be released without incurring to such inference. To this purpose, we defined different metrics that can be considered to determine when the released data can be exploited for inference, and introduced the concept of safe release according to such metrics. We also discussed how to enforce the information release control at run-time, and provided an experimental evaluation of the proposed solution, proving its efficacy. Our work leaves space for further investigations that can extend our solution in several directions. Interesting open issues that can be addressed include the consideration of inferences arising from information other than value distributions differing from a given pre-defined one, and the consideration of different types of knowledge that observers can exploit for inference, such as the order in which tuples are released or the time between two subsequent releases.

Acknowledgements

This work was supported in part by the EU within the 7FP project "PrimeLife" under grant agreement 216483, by the Italian Ministry of Research within the PRIN 2008 project "PEPPER" (2008SY2PH4), and by the Università degli Studi di Milano within the "UNIMI per il Futuro - 5 per Mille" project "PREVIOUS".

References

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

[2] C. Aggarwal and P.S. Yu, editors. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[3] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F. Standaert, and N. Veyrat-Charvillon. Mutual information analysis: A comprehensive study. *Journal of Cryptology*, 24(2):269–291, April 2011.

[4] M. Bezzi, S. De Capitani di Vimercati, G. Livraga, and P. Samarati. Protecting privacy of sensitive value distributions in data release. In *Proc. of the 6th Workshop on Security and Trust Management (STM 2010)*, Athens, Greece, September 2010.

[5] E. Brier, C. Clavier, and F. Olivier. Correlation power analysis with a leakage model. In *Proc. of the 6th International Workshop on Cryptographic Hardware and Embedded Systems (CHES 2004)*, Cambridge, MA, USA, August 2004.

[6] F. Cayre, C. Fontaine, and T. Furon. Watermarking security: Theory and practice. *IEEE Transactions on Signal Processing*, 53(10):3976–3987, October 2005.

[7] P.E. Cheng, J.W. Liou, M. Liou, and J.A.D. Aston. Data information in contingency tables: A fallacy of hierarchical loglinear models. *Journal of Data Science*, 4(4):387–398, October 2006.

[8] S. Cimato, M. Gamassi, V. Piuri, R. Sassi, and F. Scotti. Privacy-aware biometrics: Design and implementation of a multimodal verification system. In *Proc. of the 24th Annual Computer Security Applications Conference (ACSAC 2008)*, Anaheim, CA, USA, December 2008.

[9] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. k-Anonymity. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.

[10] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Microdata protection. In T. Yu and S. Jajodia, editors, *Secure Data Management in Decentralized Systems*. Springer-Verlag, 2007.

[11] S. Dawson, S. De Capitani di Vimercati, P. Lincoln, and P. Samarati. Minimal data upgrading to prevent inference and association attacks. In *Proc. of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 1999)*, Philadelphia, PA, USA, May/June 1999.

[12] S. Dawson, S. De Capitani di Vimercati, P. Lincoln, and P. Samarati. Maximizing sharing of protected information. *Journal of Computer and System Sciences*, 64(3):496–541, May 2002.

[13] S. De Capitani di Vimercati, S. Foresti, S. Jajodia, S. Paraboschi, and P. Samarati. Fragments and loose associations: Respecting privacy in data publishing. *Proc. of the VLDB Endowment*, 3(1):1370–1381, September 2010.

[14] D.D. Denning, T.F. Lunt, R.R Schell, M. Heckman, and W.R. Shockley. The SeaView security model. *IEEE Transactions of Software Engineering*, 16(6):593–607, June 1990.

[15] W. J. Dixon. Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4):488–506, December 1950.

[16] W. J. Dixon. Ratios involving extreme values. *The Annals of Mathematical Statistics*, 22(1):58–78, March 1951.

[17] R. M. Fano. *Transmission of Information; A Statistical Theory of Communications*. MIT University Press, New York, NY, USA, 1961.

[18] K.B Frikken and Y. Zhang. Yet another privacy metric for publishing micro-data. In *Proc. of the 7th ACM Workshop on Privacy in the Electronic Society (WPES 2008)*, Alexandria, VA, USA, October 2008.

[19] M. Gamassi, V. Piuri, S. Sana, and F. Scotti. Robust fingerprint detection for access control. In *Proc. of the 2nd RoboCare Workshop (RoboCare 2005)*, Rome, Italy, May 2005.

[20] B. Gierlichs, L. Batina, P. Tuyls, and B. Preneel. Mutual information analysis - A generic side-channel distinguisher. In *Proc. of the 10th International Workshop on Cryptographic Hardware and Embedded Systems (CHES 2008)*, Washington, DC, USA, August 2008.

[21] J.A. Goguen and J. Meseguer. Unwinding and inference control. In *Proc. of the 1984 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, May 1984.

[22] J.T. Haigh, R.C. O'Brien, and D.J. Thomsen. The LDV secure relational DBMS model. In *Proc. of the 4th IFIP WG 11.3 Workshop on Database Security (DBSec 1990)*, Halifax, UK, September 1990.

[23] T.H. Hinke, H.S. Delugach, and A. Chandrasekhar. A fast algorithm for detecting second paths in database inference analysis. *Journal of Computer Security*, 3(2/3):147–168, June 1995.

[24] S. Jajodia and C. Meadows. Inference problems in multilevel secure database management systems. In M. Abrams, S. Jajodia, and H. Podell, editors, *Information Security: An integrated collection of essays*. IEEE Computer Sociery Press, 1995.

[25] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, April 2006.

[26] F. Li, J. Sun, S. Papadimitriou, G.A. Mihaila, and I. Stanoi. Hiding in the crowd: Privacy preservation on evolving streams through correlation tracking. In *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, April 2007.

[27] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, Istanbul, Turkey, April 2007.

[28] T.F. Lunt. Aggregation and inference: Facts and fallacies. In *Proc. of the 1989 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, May 1989.

[29] A. Machanavajjhala, J. Gehrke, and D. Kifer. ℓ-density: Privacy beyond k-anonymity. In *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, Atlanta, GA, USA, April 2006.

[30] D.G. Marks, A. Motro, and S. Jajodia. Enhancing the controlled disclosure of sensitive information. In *Proc. of the 4th European Symposium on Research in Computer Security (ESORICS 1996)*, Rome, Italy, September 1996.

[31] M. Morgenstern. Controlling logical inference in multilevel database systems. In *Proc. of the 1988 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, May 1988.

[32] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition, 2007.

[33] X. Qian, M.E. Stickel, P.D. Karp, T.F. Lunt, and T.D. Garvey. Detection and elimination of inference channels in multilevel relational database. In *Proc. of the 1993 IEEE Symposium on Research in Security and Privacy*, Oakland, CA, May 1993.

[34] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, November/December 2001.

[35] M. Schaefer, editor. *Multilevel Data Management Security*. Air Force Studies Board Committee on Multilevel Data Management Security, National Academy Press, 1983.

[36] G.W. Smith. Modeling security-relevant data semantics. *IEEE Transactions on Software Engineering*, 17(11):1195–1203, November 1991.

[37] P. Smyth and R.M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316, August 1992.

[38] TSP 8 - Age distribution of UK regular forces, Edition - 01 Apr 2006. `http://www.dasa.mod.uk/applications/newWeb/www/index.php?page=67&pubType=1&thiscontent=80`.

[39] N. Veyrat-Charvillon and F. Standaert. Mutual information analysis: How, when and why? In *Proc. of the 11th International Workshop on Cryptographic Hardware and Embedded Systems (CHES 2009)*, Lausanne, Switzerland, September 2009.

[40] K. Wang, Y. Xu, R. Wong, and A. Fu. Anonymizing temporal data. In *Proc. of the 10th IEEE International Conference on Data Mining (ICDM 2010)*, Sydney, Australia, December 2010.

[41] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proc. of the 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Korea, September 2006.

[42] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia. Continuous privacy preserving publishing of data streams. In *Proc. of the 12th International Conference on Extending Database Technology (EDBT 2009)*, Saint Petersburg, Russia, March 2009.

Figure 1: Reference scenario

Figure 2: Number of tuples in table T by `Age` and `Location` (a), L_i-conditioned distributions P(`Age`$|L_i$), $i = 1, \ldots, 5$, over table T (b), and location frequencies (c)

Figure 3: Histogram representation of the baseline distribution (a) and of the L_i-conditioned distributions $P(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, in Figure 2(b)

Figure 4: Statistical tests and safe release control

Figure 5: Comparison between the chi-square distribution with 45 degrees of freedom and the distribution of $2N_r \log(2) I_r(\texttt{Age}, \texttt{Location})$

Figure 6: Number of tuples by `Age` and `Location` in a safe dataset T_r w.r.t. mutual information significance with $\alpha = 20\%$ (a), L_i-conditioned distributions $P_r(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, over T_r (b), and location frequencies (c)

Figure 7: Comparison between the chi-square distribution with 9 degrees of freedom and the distribution of $2N_r(L_1) \log(2) \Delta_r(\texttt{Age}, L_1)$ (a), $2N_r(L_2) \log(2) \Delta_r(\texttt{Age}, L_2)$ (b), $2N_r(L_3) \log(2) \Delta_r(\texttt{Age}, L_3)$ (c), $2N_r(L_4) \log(2) \Delta_r(\texttt{Age}, L_4)$ (d), and $2N_r(L_5) \log(2) \Delta_r(\texttt{Age}, L_5)$ (e)

Figure 8: Number of tuples by `Age` and `Location` in a safe dataset T_r w.r.t. Kullback-Leibler distance with $\alpha = 20\%$ (a), L_i-conditioned distributions $P_r(\texttt{Age}|L_i)$, with $i = 1, \ldots, 5$, over T_r (b), and location frequencies (c)

Figure 9: Number of tuples by `Age` and `Location` in a safe dataset T_r w.r.t. Chi-Square Goodness-of-Fit with $\alpha = 20\%$ (a), L_i-conditioned distributions $P_r(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, over T_r (b), and location frequencies (c)

Figure 10: Critical values Q_c for the Dixon's Q-test with significance levels 20%, 10%, 5%, 1% and [3-10] distinct values in Y domain [16]

Figure 11: Number of tuples by `Age` and `Location` in a safe dataset T_r w.r.t. Dixon's Q-test with $\alpha = 20\%$ (a), L_i-conditioned distributions $P_r(\texttt{Age}|L_i)$, $i = 1, \ldots, 5$, over T_r (b), and location frequencies (c)

Figure 12: Evolution of the mutual information and its critical value

Figure 13: Evolution of the Kullback-Leibler distance between $P_r(\texttt{Age}|L_i)$ and $P(\texttt{Age})$ and its critical value for each location

Figure 14: Evolution of the Pearson's cumulative statistic and its critical value for each location

Figure 15: Evolution of the Dixon's coefficient and its critical value

Figure 16: Average number of requested tuples that have been released by each metric for each location

Figure 17Fitting the baseline distribution within the L_2-conditioned distribution

Figure 18: Number of requested tuples released fitting the baseline

Figure 19: Number of datasets obtained adopting a metric that are safe also with respect to the other metrics

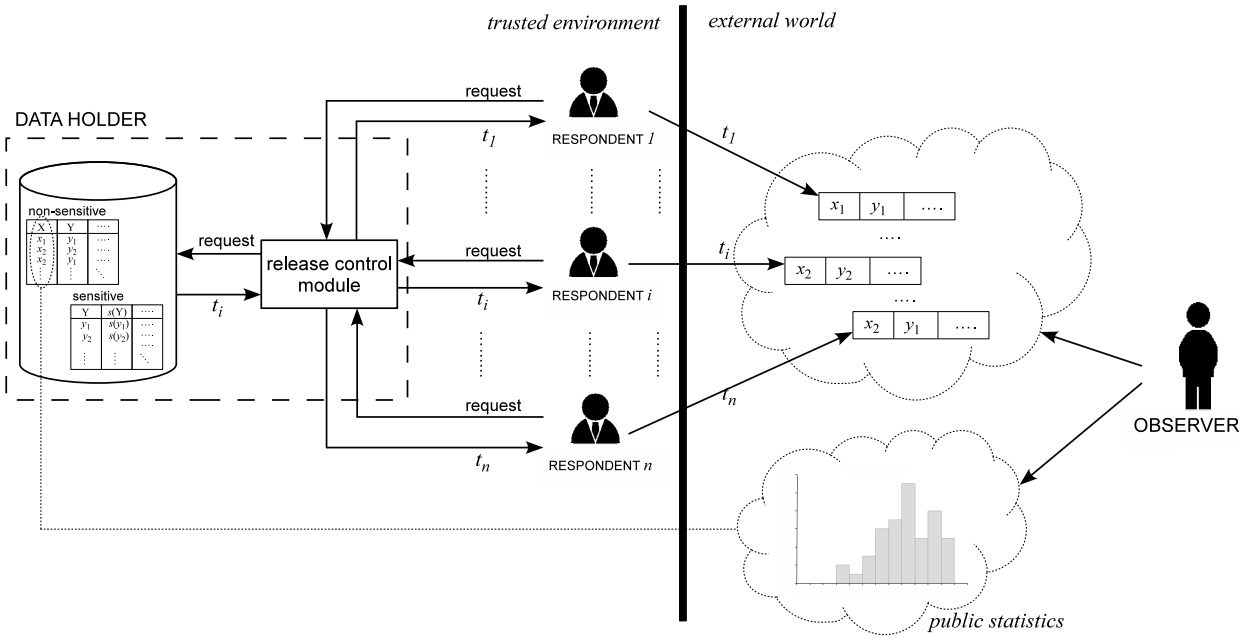

trusted environment
external world
DATA HOLDER
non-sensitive
sensitive
request
release control module
RESPONDENT 1
RESPONDENT i
RESPONDENT n
OBSERVER
public statistics

Figure 1

Age	Number of tuples					
	L1	L2	L3	L4	L5	**Total**
<18	72	26	38	47	73	256
18-19	151	53	82	140	223	649
20-24	539	147	449	505	736	2376
25-29	452	114	370	418	613	1967
30-34	335	213	234	318	501	1601
35-39	321	238	277	332	538	1706
40-44	128	219	122	162	220	851
45-49	20	205	50	49	76	400
50-54	9	71	28	34	31	173
≥55	2	13	2	2	2	21
Total	2029	1299	1652	2007	3013	10000

(a)

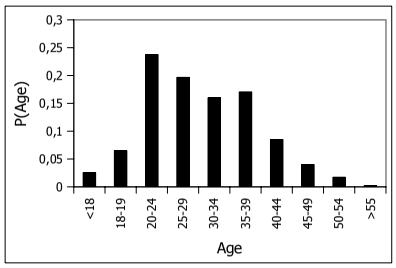
Age	$\mathbf{P(Age\|}L_i\mathbf{)}$					
	L1	L2	L3	L4	L5	**P(Age)**
<18	3.55	2.00	2.31	2.34	2.42	2.56
18-19	7.44	4.08	4.96	6.98	7.40	6.49
20-24	26.56	11.32	27.18	25.16	24.44	23.76
25-29	22.28	8.78	22.40	20.83	20.35	19.67
30-34	16.51	16.40	14.16	15.84	16.63	16.01
35-39	15.82	18.32	16.77	16.54	17.86	17.06
40-44	6.31	16.86	7.38	8.07	7.30	8.51
45-49	0.99	15.78	3.03	2.44	2.52	4.00
50-54	0.44	5.46	1.69	1.69	1.03	1.73
≥55	0.10	1.00	0.12	0.11	0.05	0.21

(b)

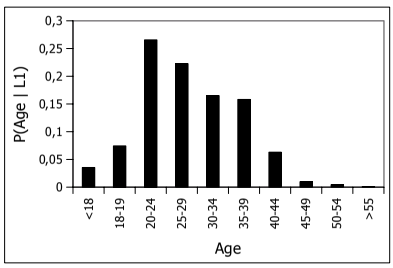
L_i	$\mathbf{P}(L_i)$
L_1	20.29
L_2	12.99
L_3	16.52
L_4	20.07
L_5	30.13

(c)

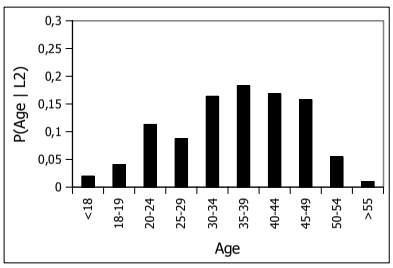
Figure 2



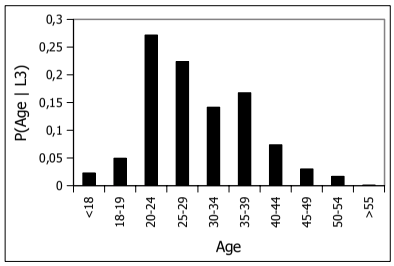
(a) $P(\texttt{Age})$



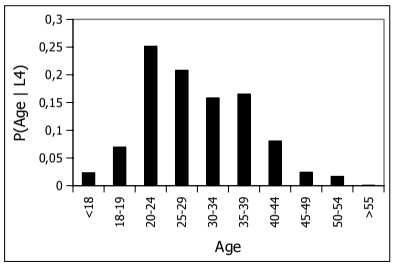
(b) $P(\texttt{Age}|L_1)$



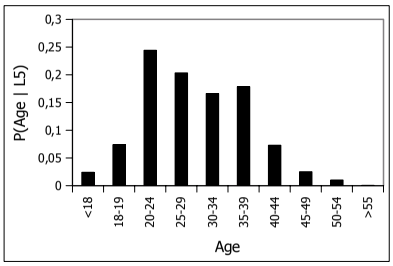
(c) $P(\texttt{Age}|L_2)$



(d) $P(\texttt{Age}|L_3)$



(e) $P(\texttt{Age}|L_4)$



(f) $P(\texttt{Age}|L_5)$

Figure 3

		Test	Safe release control
Statistical Independence		MIS (Section 5.1)	$I_r(X,Y) < I_{rc}$
Distance	**Absolute**	KLD (Section 5.2)	$\forall y \in Y,\ \Delta_r(X,y) < \Delta_{rc}(y)$
		CST (Section 5.3)	$\forall y \in Y,\ F_r(X,y) < F_{rc}$
	Relative	DQT (Section 5.4)	$Q_r(X) < Q_{rc}$

Figure 4

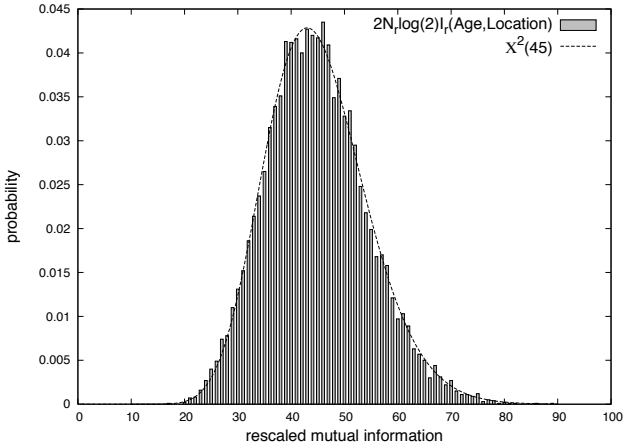

2N_r log(2) I_r(Age,Location)
X^2(45)
probability
rescaled mutual information
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
10
20
30
40
50
60
70
80
90
100

Figure 5

Age	Number of tuples					
	L1	L2	L3	L4	L5	**Total**
<18	9	5	7	8	11	40
18-19	23	11	12	19	29	94
20-24	80	30	68	70	109	357
25-29	71	18	55	58	88	290
30-34	51	30	43	47	74	245
35-39	55	28	46	50	76	255
40-44	25	24	23	25	38	135
45-49	2	10	11	11	13	47
50-54	2	8	4	5	6	25
≥55	1	1	0	0	0	2
Total	319	165	269	293	444	1490

(a)

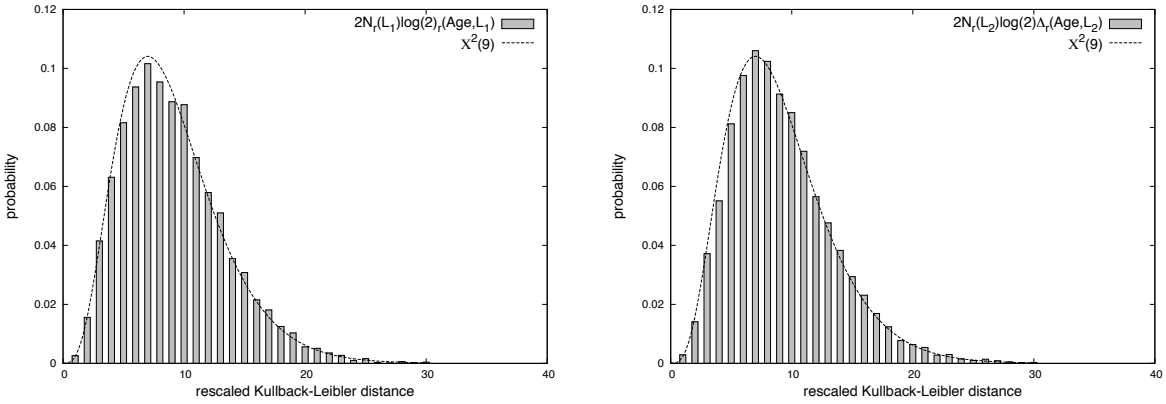
Age	$Pr(\mathbf{Age}\|L_i)$					
	L1	L2	L3	L4	L5	$Pr(\mathbf{Age})$
<18	2.82	3.03	2.60	2.73	2.48	2.68
18-19	7.21	6.67	4.46	6.49	6.53	6.31
20-24	25.08	18.18	25.28	23.89	24.55	23.96
25-29	22.26	10.91	20.45	19.80	19.81	19.46
30-34	15.99	18.18	15.98	16.04	16.67	16.44
35-39	17.24	16.97	17.10	17.06	17.12	17.11
40-44	7.84	14.55	8.55	8.53	8.56	9.07
45-49	0.63	6.06	4.09	3.75	2.93	3.15
50-54	0.63	4.85	1.49	1.71	1.35	1.69
≥55	0.30	0.60	0.00	0.00	0.00	0.13

(b)

L_i	$Pr(L_i)$
L_1	21.41
L_2	11.08
L_3	18.05
L_4	19.66
L_5	29.80

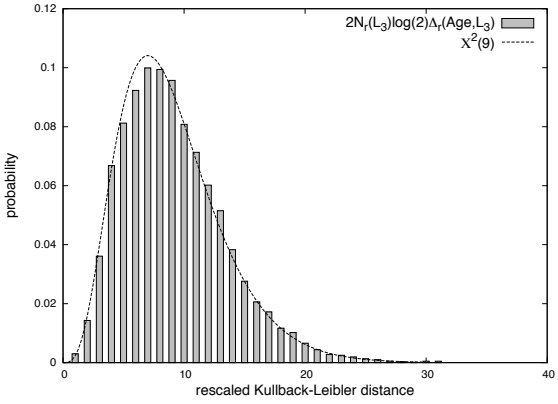
(c)

Figure 6

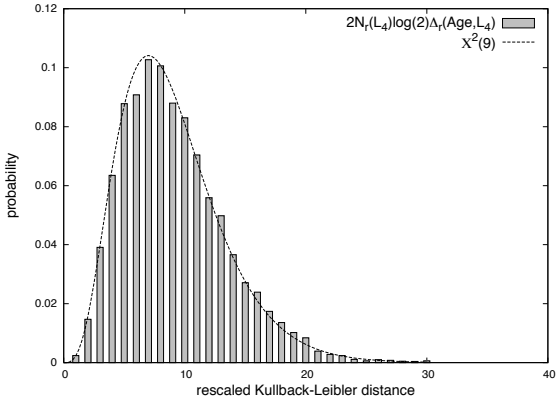


(a) L_1

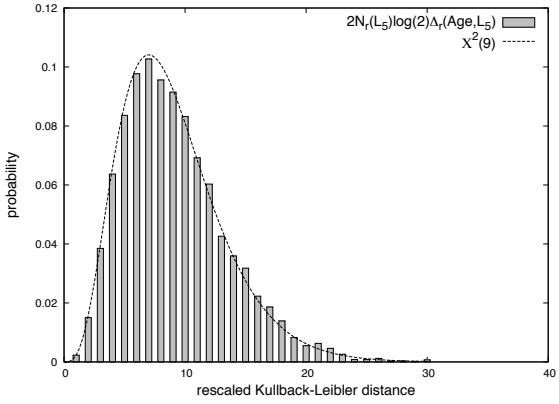
(b) L_2



(c) L_3



(d) L_4



(e) L_5

Figure 7

Age	Number of tuples					
	L1	L2	L3	L4	L5	**Total**
<18	12	4	6	5	16	43
18-19	25	11	18	18	43	115
20-24	86	29	90	72	141	418
25-29	66	19	65	67	112	329
30-34	56	31	37	49	94	267
35-39	57	29	55	51	115	307
40-44	19	18	19	27	47	130
45-49	9	8	8	4	13	42
50-54	2	4	6	2	7	21
≥55	0	1	1	1	0	3
Total	332	154	305	296	588	1675

(a)

Age	$Pr(\mathbf{Age}\|L_i)$					
	L1	L2	L3	L4	L5	$Pr(\mathbf{Age})$
<18	3.61	2.60	1.97	1.69	2.72	2.57
18-19	7.53	7.14	5.90	6.08	7.31	6.87
20-24	25.90	18.83	29.51	24.32	23.98	24.96
25-29	19.89	12.34	21.31	22.64	19.05	19.64
30-34	16.87	20.13	12.13	16.55	15.99	15.94
35-39	17.17	18.83	18.03	17.23	19.56	18.33
40-44	5.72	11.69	6.23	9.12	7.99	7.75
45-49	2.71	5.19	2.62	1.35	2.21	2.51
50-54	0.60	2.60	1.97	0.68	1.19	1.25
≥55	0.00	0.65	0.33	0.34	0.00	0.18

(b)

L_i	$Pr(L_i)$
L_1	19.82
L_2	9.20
L_3	18.21
L_4	17.67
L_5	35.10

(c)

Figure 8

Age	Number of tuples					
	L1	L2	L3	L4	L5	**Total**
<18	13	0	8	6	4	31
18-19	25	1	13	35	35	109
20-24	92	0	80	100	135	407
25-29	74	0	76	94	117	361
30-34	65	3	55	63	98	284
35-39	64	38	48	71	94	315
40-44	32	7	21	29	41	130
45-49	3	3	11	13	18	48
50-54	0	0	3	8	4	15
≥55	0	0	0	0	0	0
Total	368	52	315	419	546	1700

(a)

Age	$Pr(\mathbf{Age}\|L_i)$					
	L1	L2	L3	L4	L5	$Pr(\mathbf{Age})$
<18	3.53	0.00	2.53	1.43	0.73	1.82
18-19	6.79	1.92	4.13	8.35	6.41	6.41
20-24	25.00	0.00	25.4	23.87	24.73	23.94
25-29	20.11	0.00	24.13	22.43	21.43	21.24
30-34	17.66	5.77	17.46	15.04	17.95	16.71
35-39	17.39	73.08	15.24	16.95	17.21	18.53
40-44	8.70	13.46	6.67	6.92	7.51	7.65
45-49	0.82	5.77	3.49	3.10	3.3	2.82
50-54	0.00	0.00	0.95	1.91	0.73	0.88
≥55	0.00	0.00	0.00	0.00	0.00	0.00

(b)

L_i	$Pr(L_i)$
L_1	21.65
L_2	3.06
L_3	18.52
L_4	24.65
L_5	32.12

(c)

Figure 9

	Number of elements							
Significance	3	4	5	6	7	8	9	10
20%	0.781	0.560	0.451	0.386	0.344	0.314	0.290	0.273
10%	0.886	0.679	0.557	0.482	0.434	0.399	0.370	0.349
5%	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
1%	0.988	0.889	0.780	0.698	0.637	0.590	0.555	0.527

Figure 10

Age	Number of tuples					
	L1	L2	L3	L4	L5	**Total**
<18	14	3	5	8	15	45
18-19	36	10	10	34	43	133
20-24	104	30	77	84	176	471
25-29	96	18	73	76	134	397
30-34	69	50	48	77	109	353
35-39	64	32	49	64	120	329
40-44	0	36	18	30	42	126
45-49	0	34	17	10	18	79
50-54	3	14	5	6	4	32
≥55	1	3	0	1	0	5
Total	387	230	302	390	661	1970

(a)

Age	$Pr(\mathbf{Age}\|L_i)$					
	L1	L2	L3	L4	L5	$Pr(\mathbf{Age})$
<18	3.62	1.30	1.66	2.05	2.27	2.28
18-19	9.30	4.35	3.30	8.72	6.51	6.75
20-24	26.87	13.04	25.50	21.54	26.63	23.91
25-29	24.81	7.83	24.17	19.49	20.27	20.15
30-34	17.83	21.75	15.89	19.74	16.49	17.92
35-39	16.54	13.91	16.23	16.41	18.15	16.70
40-44	0.00	15.65	5.96	7.69	6.35	6.40
45-49	0.00	14.78	5.63	2.56	2.72	4.01
50-54	0.78	6.09	1.66	1.54	0.61	1.63
≥55	0.25	1.30	0.00	0.26	0	0.25

(b)

L_i	$Pr(L_i)$
L_1	19.64
L_2	11.68
L_3	15.33
L_4	19.80
L_5	33.55

(c)

Figure 11

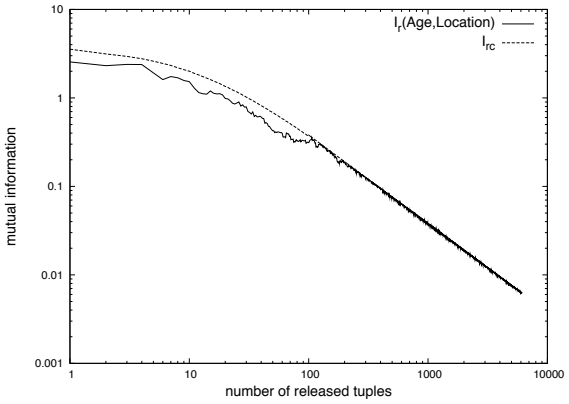
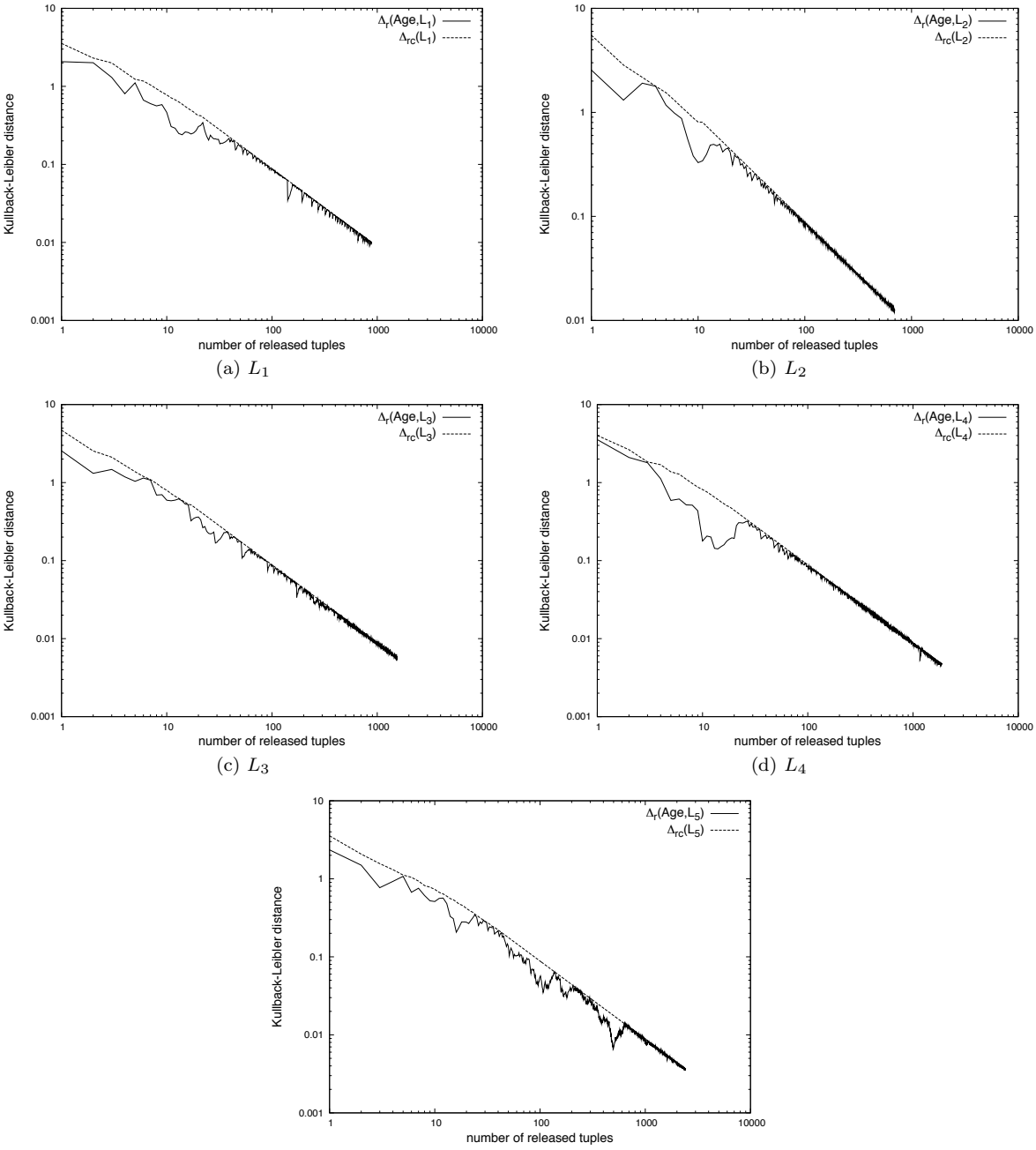

mutual information
number of released tuples
I_r(Age,Location)
I_rc
10
1
0.1
0.01
0.001
1
10
100
1000
10000

Figure 12

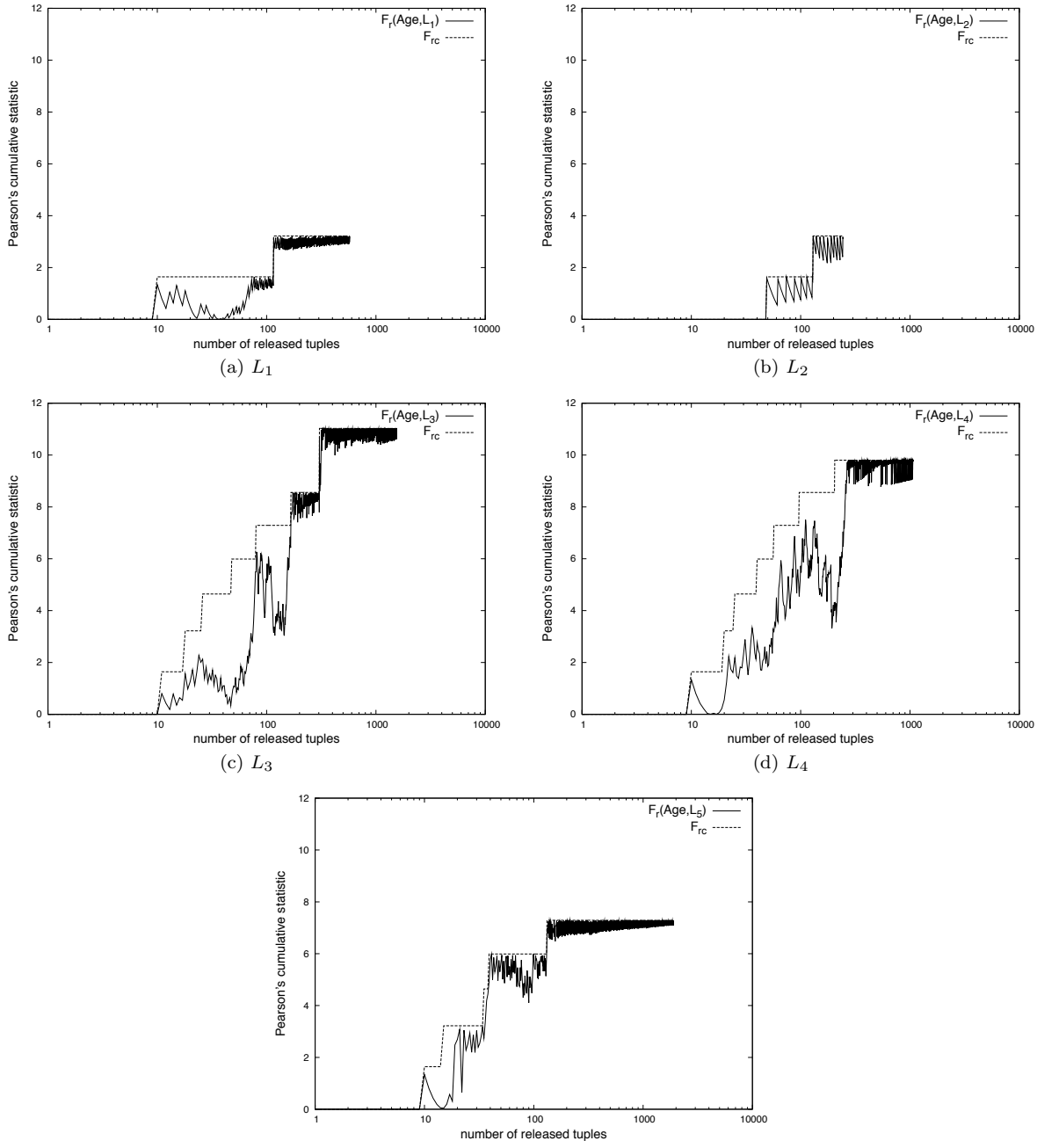


(a) L_1 (b) L_2

(c) L_3 (d) L_4

(e) L_5

Figure 13



(a) L_1

(b) L_2

(c) L_3

(d) L_4

(e) L_5

Figure 14

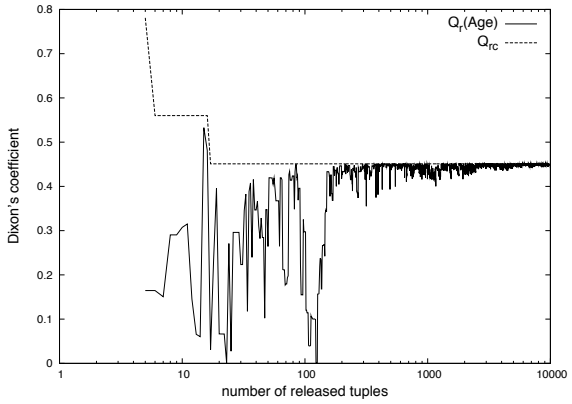

Dixon's coefficient
number of released tuples
Q_r(Age)
Q_{rc}
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
10
100
1000
10000

Figure 15

	Original	MIS		KLD		CST		DQT	
L_1	2029	1156.00	(56.97%)	871.85	(42.97%)	994.55	(49.02%)	1935.85	(95.41%)
L_2	1299	705.20	(54.29%)	697.65	(53.71%)	255.35	(19.66%)	1262.65	(97.20%)
L_3	1652	1119.00	(67.74%)	1549.75	(93.81%)	1300.00	(78.69%)	1565.45	(94.76%)
L_4	2007	1256.95	(62.63%)	1874.75	(93.41%)	1361.85	(67.86%)	1990.20	(99.16%)
L_5	3013	1876.65	(62.29%)	2415.65	(80.17%)	1899.25	(63.04%)	3013.00	(100.00%)
Total	10000	6095.78	(60.96%)	7408.67	(74.09%)	5119.88	(51.20%)	9631.55	(96.32%)

(a) $\alpha = 20\%$

	Original	MIS		KLD		CST		DQT	
L_1	2029	1187.55	(58.53%)	918.35	(45.26%)	1021.85	(50.36%)	1996.90	(98.42%)
L_2	1299	720.05	(55.43%)	713.30	(54.91%)	322.30	(24.81%)	1275.80	(98.21%)
L_3	1652	1145.90	(69.36%)	1576.20	(95.41%)	1151.90	(69.73%)	1571.80	(95.15%)
L_4	2007	1283.50	(63.95%)	1951.85	(97.25%)	1698.15	(84.61%)	1996.25	(99.46%)
L_5	3013	1907.85	(63.32%)	2530.20	(83.98%)	2344.55	(77.81%)	2996.75	(99.46%)
Total	10000	6290.58	(62.91%)	7757.14	(77.57%)	6478.14	(64.78%)	9846.14	(98.46%)

(b) $\alpha = 5\%$

Figure 16

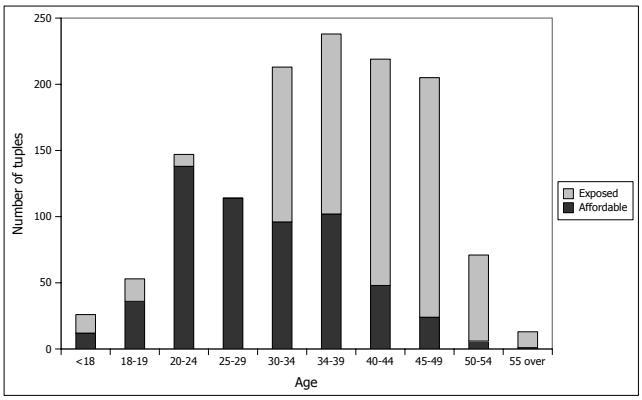


Figure 17: Fitting the baseline distribution within the L_2-conditioned distribution

	Original	**Released**	
L_1	2029	500	(24.6%)
L_2	1299	580	(44.6%)
L_3	1652	952	(57.7%)
L_4	2007	952	(47.5%)
L_5	3013	952	(31.6%)
Total	10000	3937	(39.37%)

Figure 18

	MIS	KLD	CST	DQT
MIS	100	0	0	54
KLD	1	100	1	61
CST	0	0	100	45
DQT	0	0	0	100

Figure 19