Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker

Josep Domingo-Ferrer*, Sara Ricci† and Jordi Soria-Comas‡
Universitat Rovira i Virgili
Dept. of Computer Engineering and Maths
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
* Email: josep.domingo@urv.cat
† Email: icciraras@gmail.com
‡ Email: jordi.soria@urv.cat

***Abstract*—Before releasing an anonymized data set, the data protector must know how safe the data set is, that is, how much disclosure risk is incurred by the release. If no privacy model is used to select specific privacy guarantees prior to anonymization, posterior disclosure risk assessment must be performed based on the anonymized data set and, if the result is not satisfactory, anonymization must be repeated with stricter privacy parameters. Even if a privacy model is used, it may still be advisable to empirically evaluate disclosure on the anonymized data set, especially if the privacy model parameters have been relaxed to improve data utility. Record linkage is a general methodology to posterior disclosure risk assessment, whereby the data protector attempts to recreate the attacker's re-identification scenario. An important limitation of record linkage is that it usually requires the data protector to make restrictive assumptions on the attacker's background knowledge. To overcome this limitation, we present a maximum-knowledge attacker model and then we specify and compare several record linkage tests for such a worst-case attacker. Our tests are based on comparing the distribution of linkage distances between the original and the anonymized data set with the distribution of distances between one of the two previous data sets and one random data set. The more similar the distributions, the more plausibly deniable are record linkages claimed by an attacker. Because attaining zero disclosure risk for all records is too costly in terms of utility, a less demanding alternative is presented whose goal is to reduce the maximum per-record disclosure risk.**

I. Introduction

Governments and companies are collecting large amounts of sensitive and heterogeneous information about individual subjects. These data can be useful for a variety of secondary analyses by third parties other than the data collector. Before these data are released for secondary analysis, the privacy of the individuals in the data set must be protected.

A great variety of statistical disclosure control (SDC) methods are available to protect the privacy of individual subjects [4]. All of them entail some degree of data modification, which decreases the utility of the anonymized output. Depending on the kind of modification performed on the data, methods can be classified into the following categories: non-perturbative masking (which yields an anonymized data set with less information than the original data, but preserves the truthfulness of the anonymized values), perturbative masking (which perturbs the data contained in the original records), and synthetic data generation (which generates from scratch a simulated data set that preserves some statistical properties of the original data set).

Regardless of the SDC method used, there are two main ways to tackle anonymization:

- *Utility-first*. An SDC method with not too utility-damaging parameter values is used to obtain an anonymized data set. Then the disclosure risk of the anonymized data is evaluated. If it is deemed too high, anonymization is repeated using stricter method parameters (or even using a different method); conversely, if disclosure risk is well below the tolerable levels, a relaxation of the SDC method parameters is affordable to improve the utility of the anonymized data set.
- *Privacy-first*. In this case, a privacy model is adopted to choose specific privacy guarantees *before* data anonymization. Example privacy models include k-anonymity [8] and its extensions [9], [6], [5], as well as ε-differential privacy [3]. For most privacy models, there is more than one SDC method that can be used to satisfy the model's requirements.

Under the utility-first paradigm, disclosure risk assessment is carried out *a posteriori* on the anonymized data set. Even in the privacy-first paradigm, it might be advisable to perform an *a posteriori* disclosure risk analysis if the privacy model parameters have been relaxed for the sake of utility (*e.g.* in case ε-differential privacy is used with a large ε); it is a matter of empirically checking what is the actual protection against re-identification that the relaxed model is providing.

In an *a posteriori* disclosure risk analysis, the data protector tries to simulate what an attacker would do by carrying out a record linkage attack. In such a setting, the goal of the attacker is to re-identify records in the anonymized data set by linking them to records in an external *non-de-identified* data set within the attacker's background knowledge. The proportion of correctly re-identified records is a measure of disclosure risk.

Performing a disclosure risk assessment via record linkage is not without issues:

- First of all, to simulate the attacker's record linkage, the data protector needs to make assumptions on the background knowledge available to any possible attacker; specifically, assumptions must be made on which external de-identified data sets are available to attackers and which attributes can be used for linkage because they are common or similar among those external data sets and the anonymized data set. Note that if the attacker has more background knowledge than expected by the data protector, the latter may underestimate the risk of disclosure.
- A second issue with record linkage is that it mainly focuses on the risk of identity disclosure. While identity disclosure risk is important, another disclosure risk type exists: the attacker may increase his knowledge about the confidential attributes of a certain target individual without re-identifying that individual's anonymized record (attribute disclosure).

A. Contribution and Plan of this Paper

This paper proposes a methodology based on record linkage to quantify disclosure risk. The proposed methodology addresses the above-mentioned problematic issues of the usual record linkage scenario. Specifically, it avoids making any restrictive assumptions on the attacker's background knowledge and it does so by assuming a maximum-knowledge attacker. Furthermore, since our attacker has maximum background knowledge on subjects, he cannot increase his knowledge on them, which makes attribute disclosure irrelevant and legitimates focusing just on re-identification disclosure. However, if assessing attribute disclosure is needed, we can still perform it by assuming that our attacker's knowledge excludes a particular original attribute.

Although the proposed methodology is primarily designed for disclosure risk assessment by the data protector/publisher, an attacker can also use it to some extent (depending on his actual background knowledge) to assess the accuracy of any record linkage he wishes to claim. Obviously, the data protector should leverage the methodology to make sure anonymization makes any attacker's record linkages inaccurate enough and/or plausibly random.

Section II provides background on some concepts that play an important role in this paper. Section III states and justifies our maximum-knowledge attacker model. Section IV proposes a common framework together with three different tests to assess disclosure risk based on record linkage. Section V provides experimental results. Conclusions and future research topics are summarized in Section VI.

II. Background

A. Reverse Mapping

Reverse mapping [7], [2] is a post-masking technique that can be applied to any anonymized data set. We refer to the resulting data set as the *reverse-mapped* data set.

The reverse-mapped data set is constructed by taking each attribute Y of the anonymized data set at a time, ranking the values of Y, ranking the values of the corresponding attribute X in the original data set, and replacing each value of Y with the value of X that has the same rank; we call the resulting attribute Z the reverse-mapped attribute. Algorithm 1 formalizes the reverse mapping procedure just described.

Algorithm 1: Reverse-mapping for an attribute X
Require: Original attribute $X = (x_1, x_2, \ldots, x_n)$
Require: Anonymized attribute $Y = (y_1, y_2, \ldots, y_n)$
for $i = 1$ to n **do**
Compute $j = Rank(y_i)$
Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of X of rank j)
end for
return $Z = (z_1, z_2, \ldots, z_n)$

Reverse mapping assumes that the marginal distribution of each of the attributes is not disclosive. As the marginal distribution does not refer to any particular individual but to all individuals, this assumption is quite reasonable. For those cases in which the assumption does not hold (*e.g.* if it is easy to associate the maximum or the minimum value of an attribute to a particular subject, like the top value of a "Salary" attribute being associated to chairman of a company), the marginal distributions must be anonymized prior to performing reverse mapping.

The interesting point about the reverse mapping transformation is that it allows viewing any microdata anonymization method as being *functionally equivalent* to a permutation step (mapping each original attribute X to its reverse-mapped version Z) followed by a residual noise addition step (mapping Z to Y). Note that the noise added is necessarily small, because by construction the ranks of values of Y and Z are the same. Therefore, *disclosure protection essentially comes from permutation*.

B. Permutation Distance

The permutation distance measures the dissimilarity between two records in the context of a data set. Let $\mathbf{x}_1 = (x_1^1, \ldots, x_1^m)$ and $\mathbf{x}_2 = (x_2^1, \ldots, x_2^m)$ be two records of a data set $\mathbf{X}$ with attributes $X^1, \ldots, X^m$. The permutation distance $d(\mathbf{x}_1, \mathbf{x}_2)$ between $\mathbf{x}_1$ and $\mathbf{x}_2$ is the maximum of the rank distances of corresponding attribute values, that is

$$d(\mathbf{x}_1, \mathbf{x}_2) = \max_{1 \le i \le m} |rank_{X^i}(x_1^i) - rank_{X^i}(x_2^i)| \quad (1)$$

In the context of data anonymization we are interested in computing the permutation distance between a record $\mathbf{x}$ in the original data set $\mathbf{X} = (X^1, \ldots, X^m)$ and a record $\mathbf{y}$ in the anonymized data set $\mathbf{Y} = (Y^1, \ldots, Y^m)$. Expression (1) is not directly applicable to this case. To overcome this difficulty, for each x^i we determine a value $z^i \in Y^i$ that is closest to x^i. Then, the distance between $\mathbf{x}$ and $\mathbf{y}$ is defined to be the distance between $\mathbf{z} = (z^1, \ldots, z^m)$ and $\mathbf{y}$:

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \le i \le m} |rank_{Y^i}(z^i) - rank_{Y^i}(y^i)|$$

Furthermore, the permutation distance between a record $\mathbf{x}$ and a data set $\mathbf{Y}$ can be defined as the minimum of the distances between $\mathbf{x}$ and each of the records in $\mathbf{Y}$.

$$d(\mathbf{x}, \mathbf{Y}) = \min_{\mathbf{y} \in \mathbf{Y}} d(\mathbf{x}, \mathbf{y})$$

III. Maximum-Knowledge Attacker Model

To come up with a broadly applicable disclosure risk assessment based on record linkage, we need to address the two above-mentioned shortcomings of the usual record linkage approach. As mentioned in Section I-A, our maximum-knowledge attacker model plays a key role in this endeavor.

We consider a maximum-knowledge attacker in the sense that he knows both the original data set $\mathbf{X}$ and the anonymized data set $\mathbf{Y}$, *his goal being to re-create the mapping between records in* $\mathbf{X}$ *and records in* $\mathbf{Y}$. This is the same attacker model defined in [2].

In general, when an attacker performs record linkage for a target original record corresponding to a specific subject, he uses any available background knowledge about the target subject. Useful background knowledge consists in information on the values for the target subject of as many attributes as possible. These attributes for which the attacker knows some information usable in the linkage are known as *quasi-identifiers*. In case of successful (correct) linkage via the quasi-identifiers, the reward for the attacker is to learn the values for the target subject of the rest of attributes (which include the confidential attributes whose values for the target individual the attacker did not know before the linkage).

Our maximum-knowledge attacker knows *all* original attribute values for all subjects. Hence, he can use *all* attributes as quasi-identifiers (maximum background knowledge), which allows him to compute the best possible linkages. Therefore, if the anonymization performed by the data protector is safe against such an attacker, it will be safe against any other attacker.

Yet, since our maximum-knowledge attacker knows all original attribute values for all subjects, he has no reward to gain from the linkage. One might think of this attacker as being a purely malignant one (*e.g.* whose goal is to tarnish the data protector's reputation). Yet, the attacker's motivations are not important: we just want to model a worst-case attacker who has all the background information he can possibly use.

An additional consequence of using the maximum-knowledge attacker model is that it legitimates focusing on re-identification disclosure and ignoring attribute disclosure. Indeed, our attacker already knows the values of all attributes, so he is not interested in attribute disclosure; he is only interested in the mapping between original and anonymized records, that is, re-identification. However, if attribute disclosure assessment is needed for some particular attribute, we show in the next section how to perform it by excluding that particular attribute from the attacker's assumed knowledge.

IV. Disclosure Risk via Record Linkage: Framework and Tests

In order to decide which record in $\mathbf{Y}$ should be linked to a given record in $\mathbf{X}$, we must choose some criterion. In general, the goodness of a linkage depends on the similarity between the original and the anonymized records. The more similar they are, the better the linkage is. Note that the best linkage according to this similarity criterion is *not* necessarily the one that links a record $\mathbf{x} \in \mathbf{X}$ to its corresponding anonymized version $\mathbf{y} \in \mathbf{Y}$. It may well link $\mathbf{x}$ to another record $\mathbf{y}' \in Y$ that is closer to $\mathbf{x}$ than $\mathbf{y}$.

If we adopt similarity maximization as our record linkage criterion, we are faced with the fact that there are many conceivable ways to define the similarity between two records. We do not need to consider them all to assess disclosure risk. We only need to be concerned with the similarity notion favored by the data protector. Indeed, the data protector's goal is to prevent the attacker from matching records that are close according to the *protector*'s similarity notion.

If there exists a function that, given two records, returns the distance between them (that is, their dissimilarity), then the maximum similarity linkage criterion amounts to finding the pair of records $(\mathbf{x}, \mathbf{y})$ at minimum distance, with $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$. In the experimental evaluation of Section V, we use the permutation distance [2] described in Section II-B above. After computing the best linkages in $\mathbf{Y}$ for all records in $\mathbf{X}$, we consider the distances between the linked pairs $(\mathbf{x}, \mathbf{y})$, and more precisely the distribution of those distances. To assess the disclosure risk incurred when publishing $\mathbf{Y}$, we compare the previous distribution of distances with another distribution of distances that is known to correspond to non-disclosive linkages. In the following subsections we propose different types of non-disclosive linkages that can be used as a benchmark.

Let $dist$ be the distribution of linkage distances obtained from linking records in $\mathbf{X}$ and $\mathbf{Y}$. Let $dist'$ be the distribution of distances from the non-disclosive linkage mentioned above. By comparing the distributions $dist$ and $dist'$ we can assess the extent to which knowing $\mathbf{X}$ helps in the linkage:

- If both distributions are equal, then there is no evidence of any information in $\mathbf{X}$ that the attacker can use to improve the accuracy of the linkage. This can be construed as $\mathbf{X}$ and $\mathbf{Y}$ being independent, possibly due to $\mathbf{Y}$ being a very strongly anonymized version of $\mathbf{X}$ (such a strong anonymization will probably result in a large information loss, though).
- The most likely situation is one in which $dist$ and $dist'$ are not equal. Inequality can take two forms:
  - The more natural case is one in which the probability mass of $dist$ is concentrated in smaller distances than the probability mass of $dist'$. This indicates that $\mathbf{X}$ does indeed contain information that is useful for the linkage with $\mathbf{Y}$; put in another (equivalent) way, it indicates that $\mathbf{Y}$ reveals information on $\mathbf{X}$. The more different $dist$ and $dist'$, the more information $\mathbf{Y}$ reveals on $\mathbf{X}$.
  - A stranger case is one in which the probability mass of $dist$ is concentrated in greater distances than the probability mass of $dist'$. This indicates that knowledge of $\mathbf{X}$ is counterproductive for the linkage; in other words, $\mathbf{Y}$ is misinformative about $\mathbf{X}$. This case is probably a consequence of $\mathbf{Y}$ overprotecting $\mathbf{X}$ and it should be avoided.

Algorithm 2 presents a brief formalization of the steps described above.

$\mathbf{X}$	
A	B
a_1	b_1
a_2	b_2

$\mathbf{D_X}$	
A	B
a_1	b_1
a_1	b_2
a_2	b_1
a_2	b_2

Fig. 1. Dictionary set $\mathbf{D_X}$ corresponding to a data set $\mathbf{X}$ with two attributes A and B, and two records (a_1, b_1) and (a_2, b_2)

$\mathbf{Y}$	
A	B
a_1	b_1
a_2	b_2
$\vdots$	$\vdots$
a_n	b_n

$\mathbf{Y}'$	
A	B
$a_{\sigma(1)}$	$b_{\rho(1)}$
$a_{\sigma(2)}$	$b_{\rho(2)}$
$\vdots$	$\vdots$
$a_{\sigma(n)}$	$b_{\rho(n)}$

Fig. 2. Permuted data set $\mathbf{Y}'$ corresponding to the data set $\mathbf{Y}$. Attributes A and B have been permuted with permutations σ and ρ, respectively, both in $\mathcal{S}_n$.

Algorithm 2: Disclosure risk assessment via record linkage
Require: Original data set $\mathbf{X}$.
Require: Anonymized data set $\mathbf{Y}$.
$dist \leftarrow$ distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$.
$dist' \leftarrow$ distribution of distances of a non-disclosive linkage.
return comparison between $dist$ and $dist'$.

In the experimental section below (Section V) , we will see that distributions $dist$ and $dist'$ are spread over a range of distances. That essentially means that the risk of disclosure is not uniform across the records. Records with smaller linkage distance have a greater risk of disclosure. While this is inevitable, when seeking disclosure risk limitation it seems reasonable to require that all records be protected at least to a certain level. Otherwise said, we should make sure that none of the distances in $dist$ is below a threshold that determines the minimum acceptable protection. In this sense, we must focus on the minimum distance rather than on the distance distribution.

A. Dictionary Linkage Test

In the dictionary linkage test we compare the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ to the distribution of linkage distances between $\mathbf{D_X}$ and $\mathbf{Y}$, where $\mathbf{D_X}$ is an artificially created data set that we call the *dictionary* of $\mathbf{X}$.

$\mathbf{D_X}$ contains all possible combinations of attribute values in $\mathbf{X}$. This is illustrated in Figure 1 for a small data set $\mathbf{X}$ with two attributes and two records. The dictionary $\mathbf{D_X}$ contains the records corresponding to all possible combinations of values of the attributes A_1 and A_2.

Note that because $\mathbf{D_X}$ contains all possible combinations of attribute values, any dependency between attributes in $\mathbf{X}$ is destroyed. In fact, it is easy to check that the empirical distributions of the attributes are independent. Let us illustrate this fact with a simple check on the data set $\mathbf{D_X}$ of Figure 1. We have $\Pr(A, B) = (a_i, b_j) = 0.25$ because each combination of values appears once among the four records of D_X. Also, we have $\Pr(A = a_i) = \Pr(B = b_j) = 0.5$ because each of the values of each attribute appears twice among the four records of D_X. Thus, $\Pr(A = a_i, B = b_j) = \Pr(A = a_i) \times \Pr(B = b_j)$, that is, the empirical distributions of the attributes of $\mathbf{D_X}$ are independent.

Since, in general, what is sensitive is the relation between attributes (*e.g.* the relation between a set of quasi-identifiers and a confidential attribute) and $\mathbf{D_X}$ does not preserve any relation between the attributes of $\mathbf{X}$, no sensitive information about $\mathbf{X}$ can be extracted from $\mathbf{D_X}$. Only the (non-sensitive) marginal distribution of the attributes in $\mathbf{X}$ is preserved in $\mathbf{D_X}$. Thus, as far as $\mathbf{X}$ is concerned, linking $\mathbf{D_X}$ to $\mathbf{Y}$ is essentially the same as linking a randomly generated data set to $\mathbf{Y}$. In other words, the linkage between $\mathbf{D_X}$ and $\mathbf{Y}$ reveals nothing and, thus, it can be used as the non-disclosive linkage whose distribution of distances is $dist'$.

The size of $\mathbf{D_X}$ grows rapidly as the size of $\mathbf{X}$ increases. More precisely, it grows polynomially in the number of records of $\mathbf{X}$ and exponentially in the number of attributes of $\mathbf{X}$. This makes it unfeasible to generate the entire $\mathbf{D_X}$, except for very small $\mathbf{X}$, and thus to compute $dist'$. However, $dist'$ can be approximated by taking a random sample of the records in $\mathbf{D_X}$. The price of sampling is that we lose the good properties of $\mathbf{D_X}$, namely the independence between attributes, but, if the sample is large enough, we should get a good approximation to independence.

B. Linkage to Permuted Data Set

In this test, we modify the masked data set $\mathbf{Y}$ with the aim of reducing any dependencies between attributes that may subsist in it. According to Section IV-A, by generating the dictionary set of $\mathbf{Y}$, we get rid of any dependencies between the attributes of $\mathbf{Y}$. However, using $\mathbf{D_Y}$ as the target set for the linkage may be problematic because it contains every possible combination of attribute values in $\mathbf{Y}$. If $\mathbf{Y}$ contains the same attribute values as $\mathbf{X}$ (*e.g.* data swapping or any other masking method followed by reverse mapping), then $\mathbf{D_Y}$ contains all possible combinations of attributes values of $\mathbf{X}$. In particular, it contains $\mathbf{X}$ and the linkage distances are all equal to zero.

In order to obtain informative linkage distances, we adopt another strategy to remove the dependencies among attributes of $\mathbf{Y}$ that does not increase its size. Specifically, we apply a random permutation to each attribute of $\mathbf{Y}$ and we call the resulting data set $\mathbf{Y}'$. This procedure is illustrated in Figure 2.

Although attributes in $\mathbf{Y}'$ can be expected to be less dependent from each other than those in $\mathbf{Y}$, in a particular instantiation of $\mathbf{Y}'$, anything can happen. For example, if the randomly selected permutations turn out to be the identity permutation for both attributes, then $\mathbf{Y}'$ will be an exact copy of $\mathbf{Y}$. To avoid relying on a single problematic instantiation of $\mathbf{Y}'$, we generate $\mathbf{Y}'$ several times and perform the linkage for each instantiation.

The baseline for the disclosure risk assessment is similar to the one in the previous section. Since $\mathbf{Y}'$ can be expected to be rid of any dependencies that existed between the attributes of $\mathbf{Y}$, we can say that $\mathbf{Y}'$ contains no sensitive information (remember that we assume that marginal distributions are not sensitive, only the relation between attributes is). Therefore,

we can take the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}'$ as $dist'$, the baseline to compare with.

C. Attribute Disclosure Test

In contrast to the previous two tests which were aimed at re-identification disclosure, the one here focuses on attribute disclosure for a single confidential attribute. Rather than assuming a maximum-knowledge attacker that knows all m attributes of the original data set $\mathbf{X}$, *we now exclude one particular original attribute from the attacker's background knowledge*, say A^m without loss of generality. Thus, we assume that the attacker knows all original values for attributes $A^1, \ldots, A^{m-1}$. In this case, *the attacker's goal is to determine the value of A^m as accurately as possible.*

The attacker's strategy is to perform a record linkage between $\mathbf{X}$ and $\mathbf{Y}$ based on $A^1, \ldots, A^{m-1}$. Then the distance between the values of attribute A^m within each pair of linked records is computed.

Let us write the records of $\mathbf{X}$ as $(\mathbf{x}_b, x_m)$ where $\mathbf{x}_b = (x_1, \ldots, x_{m-1})$ and x_i is the value of attribute A^i. Let us use an analogous notation for the records of $\mathbf{Y}$. The attacker links every $(\mathbf{x}_b, x_m)$ in $\mathbf{X}$ with the record $(\mathbf{y}_b, y_m)$ in $\mathbf{Y}$ such that the distance between $\mathbf{x}_b$ and $\mathbf{y}_b$ is shortest. In principle, any distance could be valid but we propose the use of the permutation distance (see Section II-B). Finally, the attacker calculates the rank difference between x_m and y_m.

From the process above, we obtain a distribution $dist$ of distances (the rank differences between x_m and y_m). To assess the level of disclosure risk, we need to compare $dist$ to a distribution of distances $dist'$ that results from a non-disclosive linkage. In the same way as we did in Section IV-B, we take as the non-disclosive linkage the linkage between $\mathbf{X}$ and $\mathbf{Y}'$ based on attributes $A^1, \ldots, A^{m-1}$, where $\mathbf{Y}'$ is a modified version of $\mathbf{Y}$ preserving the size of $\mathbf{Y}$ but *not* the relations between the attributes of $\mathbf{Y}$. As in Section IV-B, we resort to applying a random permutation to each attribute of $\mathbf{Y}$ in order to obtain $\mathbf{Y}'$. Then we take as $dist'$ the distribution of the rank differences between values of A^m in the linked pairs $(\mathbf{x}_b, x_m)$ and $(\mathbf{y}'_b, y'_m)$.

V. Experimental Results

This section details the empirical evaluation carried out for the proposed disclosure risk assessment tests. Experiments were conducted using the "Census" data set [1], a usual test data set in the statistical disclosure control literature that contains 13 numerical attributes and 1080 records.

A. Evaluation Measures and Experiments

Since we aim at assessing the risk of disclosure in an anonymized data set, we have generated several anonymized data sets $\mathbf{Y}$ from the original "Census" data set $\mathbf{X}$. The anonymized data sets have been generated by adding independent normally distributed noise to each of the attributes, followed by reverse mapping to recover the original marginal distributions. The amount of noise has been adjusted to the variability of each of the original attributes; specifically, the standard deviation of the noise is proportional to the standard deviation of the original attribute.

To explore how the distribution of linkage distances changes with the amount of masking noise, we consider four different anonymized data sets: $\mathbf{Y}_{0.5}$, $\mathbf{Y}_1$, $\mathbf{Y}_3$ and $\mathbf{Y}_7$, where a subscript κ means that the standard deviation of the noise added to each original attribute is κ times the standard deviation of the original attribute.

The record linkage criterion chosen is to link an original record to the anonymized record at minimum permutation distance. Disclosure risk assessment is based on comparing distributions of linkage distances. We use the Kolmogorov-Smirnov distance to measure how different two distributions are; if D_1 and D_2 are the cumulative distribution functions of these distributions, the Kolmogorov-Smirnov distance is given by

$$KS(D_1, D_2) = \max_{x \in \mathbb{R}} |D_1(x) - D_2(x)|$$

and it is bounded in the $[0, 1]$ interval.

B. Results of the Dictionary Linkage Test

Since it was unfeasible to generate an exhaustive dictionary for all the "Census" attributes, we took as $\mathbf{D}_\mathbf{X}$ a random sample of 10,000 records from the exhaustive dictionary. Then, for each anonymized data set, we compared the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ and the distribution of linkage distances between $\mathbf{D}_\mathbf{X}$ and $\mathbf{Y}$. The top chart of Figure 3 shows both distributions for different amounts of masking noise:

- The curves labeled "noise=0.5", "noise=1", "noise=3" and "noise=7" correspond to the distribution of the linkage distances between the original data set $\mathbf{X}$ and the anonymized data sets $\mathbf{Y}_{0.5}$, $\mathbf{Y}_1$, $\mathbf{Y}_3$ and $\mathbf{Y}_7$, respectively.
- The curve labeled "dictionary" corresponds to the four distributions of the linkage distances between the dictionary data set and the anonymized data sets. It turns out that the distributions were practically the same no matter the amount of anonymization, so we just plotted one curve for the four of them.

From the top chart of Figure 3 it can be seen that a lot of noise is needed for the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ to be similar to the one between $\mathbf{D}_\mathbf{X}$ and $\mathbf{Y}$, that is, for anonymization to be perfect. For instance, the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}_3$ is still quite different from the distribution between $\mathbf{D}_\mathbf{X}$ and $\mathbf{Y}_3$ (even if $\mathbf{Y}_3$ is already very strongly anonymized and probably useless, with noise whose standard deviation is 3 times the deviation of each original attribute).

The bottom chart of Figure 3 is analogous to the top chart, but replacing the original data set $\mathbf{X}$ by a data set $\mathbf{X}^\sigma$ whose attributes take values that are random permutations of the corresponding attribute in $\mathbf{X}$. Hence, in this case, the attributes in $\mathbf{X}^\sigma$ can be thought as being nearly independent from each other. This bottom chart shows that a moderate amount of noise ensures already perfect anonymization: indeed, the distribution of linkage distances between $\mathbf{X}^\sigma$ and $\mathbf{Y}^\sigma_1$ is already almost identical to the distribution between $\mathbf{D}_{\mathbf{X}^\sigma}$ and $\mathbf{Y}_1$.

In summary, the greater the dependency among attributes in the original data, the more noise is required to attain a

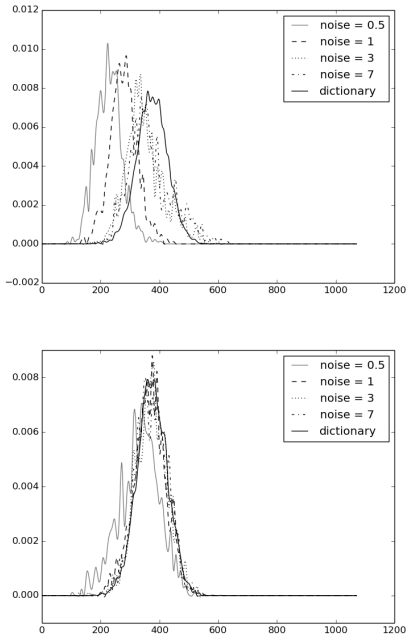

Fig. 3. Top, distributions of linkage distances between the $\mathbf{X}$ and the various anonymized data sets, and distributions of the linkage distances between the dictionary data set $\mathbf{D_X}$ and the various anonymized data sets (a single curve labeled "dictionary" represents the latter distributions, because they are extremely similar regardless the anonymization level used). Bottom, same as the top graph but replacing $\mathbf{X}$ by a random permutation $\mathbf{X}^\sigma$ and $\mathbf{D}^\sigma$ by $\mathbf{D_{X^\sigma}}$.

perfect anonymization. In fact, the distribution between the dictionary and the anonymized data set is unattainable if one wants to preserve some utility in the anonymized data set; this distribution should only be regarded as a benchmark for low disclosure risk.

This is further illustrated in Figure 4. The solid curve shows the variation in the correlation matrix (by means of the Frobenius norm) in terms of the noise being added, more precisely the ratio κ between the standard deviation of the noise applied to each original attribute and the standard deviation of the attribute. For practical purposes, we are more interested in the minimum of the linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ in terms of the noise being added, shown by the dashed curve in Figure 4. This distance represents the maximum level of disclosure risk among the records of $\mathbf{X}$, that is, the maximum level of disclosure protection that can be guaranteed for all the records in $\mathbf{X}$. For several levels of noise, Table I gives numerical values for: i) the minimum of the $(\mathbf{X}, \mathbf{Y})$ linkage distances; ii) the correlation between the $(\mathbf{X}, \mathbf{Y})$ linkage distances and the $(\mathbf{D_X}, \mathbf{Y})$ linkage distances; iii) same as ii) but replacing $\mathbf{X}$ with $\mathbf{X}^\sigma$.

C. Results of the Linkage to Permuted Data Set

For each anonymized data set, we compared the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ and the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}'$, where $\mathbf{Y}'$ is a permuted data set generated from $\mathbf{Y}$ by randomly permuting each attribute. The top chart of Figure 5 shows both distributions for different amounts of noise:

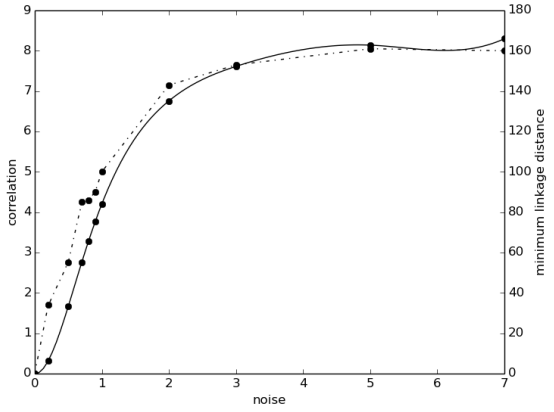

Fig. 4. Solid curve: variation in the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$ (expressed by the Frobenius norm). Dashed curve: variation of the minimum linkage distance between $\mathbf{X}$ and $\mathbf{Y}$. Both curves are functions of the noise being added (expressed by the ratio between the standard deviation of the noise and the standard deviation of each attribute in $\mathbf{X}$).

TABLE I. MINIMUM $(\mathbf{X}, \mathbf{Y})$ LINKAGE DISTANCE, KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE $(\mathbf{X}, \mathbf{Y})$ LINKAGE DISTANCE DISTRIBUTION AND THE $(\mathbf{D_X}, \mathbf{Y})$ LINKAGE DISTANCE DISTRIBUTION, KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE $(\mathbf{X}^\sigma, \mathbf{Y}^\sigma)$ LINKAGE DISTANCE DISTRIBUTION AND THE $(\mathbf{D_{X^\sigma}}, \mathbf{Y}^\sigma)$ LINKAGE DISTANCE DISTRIBUTION, FOR SEVERAL LEVELS OF NOISE.

	$Y_{0.5}$	Y_1	Y_3	Y_7
minimum $(\mathbf{X}, \mathbf{Y})$ linkage distance	55	100	153	160
KS betw. $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{D_X}, \mathbf{Y})$ linkage distances	0.85	0.68	0.25	0.15
KS betw. $(\mathbf{X}^\sigma, \mathbf{Y}^\sigma)$ and $(\mathbf{D_{X^\sigma}}, \mathbf{Y}^\sigma)$ link. dist.	0.32	0.07	0.01	0.02

- The curves labeled "noise=0.5", "noise=1", "noise=3" and "noise=7" correspond to the distribution of the linkage distances between the original data set $\mathbf{X}$ and the anonymized data sets $\mathbf{Y}_{0.5}$, $\mathbf{Y}_1$, $\mathbf{Y}_3$ and $\mathbf{Y}_7$, respectively.

- The curve labeled "permutated" corresponds to the four distributions of the linkage distances between $\mathbf{X}$ and the permuted versions of the anonymized data sets. It turns out that the distributions were practically the same no matter the amount of anonymization, so we just plotted one curve for the four of them.

The bottom chart of Figure 5 is analogous to the top chart, but it replaces the original data set $\mathbf{X}$ with a permuted version $\mathbf{X}^\sigma$ (which has less dependency between attributes), from which new anonymized and permuted anonymized data sets are derived.

The results of Figure 5 are consistent with those of the dictionary linkage test. Attaining perfect anonymization requires a large amount of noise, and the greater the dependency between the attributes in $\mathbf{X}$, the more noise it is required. The distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}'$ should only be used as benchmark for low disclosure risk, because it is unattainable using a reasonably anonymized $\mathbf{Y}$. This is illustrated in Table II, which is analogous to Table I given for the dictionary test.

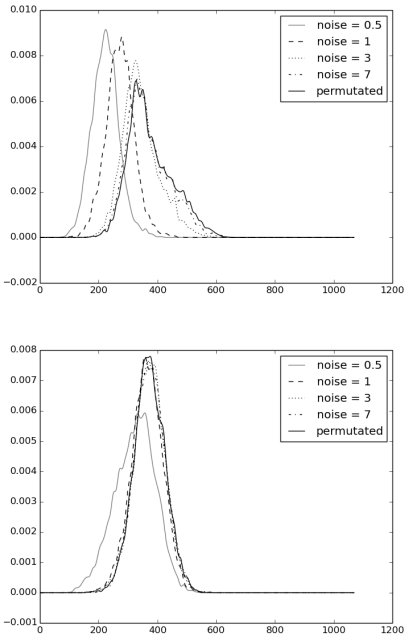

Fig. 5. Top, distributions of linkage distances between the $\mathbf{X}$ and the various anonymized data sets, and distributions of the linkage distances between $\mathbf{X}$ and the various permuted versions of the anonymized data sets (a single curve labeled "permutated" represents the latter distributions, because they are extremely similar regardless the anonymization level used). Bottom, same as the top graph but replacing $\mathbf{X}$ by a random permutation $\mathbf{X}^\sigma$, from which the various anonymized data sets and permuted anonymized data sets are re-derived.

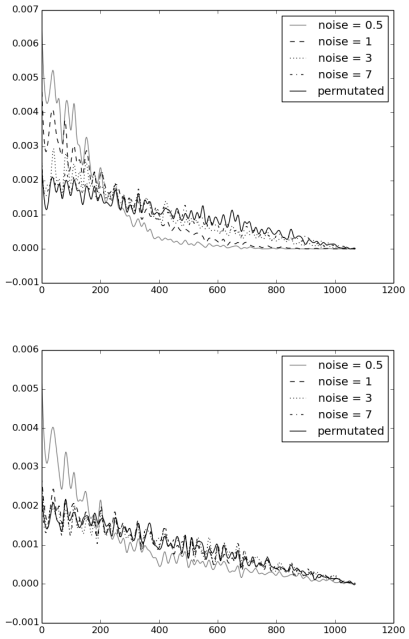

Fig. 6. Top, distributions of linkage distances between the $\mathbf{X}$ and the various anonymized data sets, and distributions of the linkage distances between $\mathbf{X}$ and the various permuted versions of the anonymized data sets (a single curve labeled "permutated" represents the latter distributions, because they are extremely similar regardless the anonymization level used). Bottom, same as the top graph but replacing $\mathbf{X}$ by a random permutation $\mathbf{X}^\sigma$, from which the various anonymized data sets and permuted anonymized data sets are re-derived.

TABLE II. MINIMUM $(\mathbf{X}, \mathbf{Y})$ LINKAGE DISTANCE, KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE $(\mathbf{X}, \mathbf{Y})$ LINKAGE DISTANCE DISTRIBUTION AND THE $(\mathbf{X}, \mathbf{Y}')$ LINKAGE DISTANCE DISTRIBUTION, KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE $(\mathbf{X}^\sigma, \mathbf{Y}^\sigma)$ LINKAGE DISTANCE DISTRIBUTION AND THE $(\mathbf{X}^\sigma, \mathbf{Y}^{\sigma\prime})$ LINKAGE DISTANCE DISTRIBUTION, FOR SEVERAL LEVELS OF NOISE.

	$Y_{0.5}$	Y_1	Y_3	Y_7
minimum $(\mathbf{X}, \mathbf{Y})$ linkage distance	55	100	153	160
KS betw.$(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}, \mathbf{Y}')$ link. dist.	0.83	0.61	0.15	0.033
KS betw. $(\mathbf{X}^\sigma, \mathbf{Y}^\sigma)$ and $(\mathbf{X}^\sigma, \mathbf{Y}^{\sigma\prime}$ link. dist.	0.30	0.055	0.006	0.006

D. Results of the Attribute Disclosure Test

In the experiments for this test, we considered in turn each of the 13 attributes in the "Census" data set as the unknown attribute A^m, and in what follows we report the average results obtained for the 13 batteries of resulting experiments (in fact the results were very similar for all attributes).

For each anonymized data set, we compared the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ and the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}'$, where $\mathbf{Y}'$ is a permuted data set generated from $\mathbf{Y}$ by randomly permuting each attribute. In this case the linkage distances are the rank differences between the values of the unknown attribute A^m in the linked pairs.

The distributions of linkage distances are depicted in Figure 6, which is analogous to Figures 3 and 5 of the previous two tests:

- The top chart of the figure shows the linkage distances when using the original data set $\mathbf{X}$, for several levels of anonymization. When $\mathbf{X}$ is linked to the anonymized $\mathbf{Y}$ data sets, the linkage distances depend on the level of anonymization, and a different curve is given for each anonymization level. However, when $\mathbf{X}$ is linked to permuted anonymized data sets $\mathbf{Y}'$, the level anonymization becomes irrelevant, and a single curve labeled "permutated" is given.
- The bottom chart of the figure is analogous to the top one, but replacing $\mathbf{X}$ with a permuted version $\mathbf{X}^\sigma$ which has less dependency between attributes.

The results in Figure 6 are consistent with those in Figures 3 and 5. Attaining perfect anonymization requires a large amount of noise, and the greater the dependency between the unknown attribute and the known attributes $\mathbf{X}$, the more noise is required.

In this attribute disclosure test, the minimum of the linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ is zero (which means equal rank for the values of the unknown attribute A^m in both linked records). The mean of the linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ is a better indicator of the protection provided to the unknown attribute, because it gives an idea of the average rank change caused by anonymization in that attribute. Table III is analogous to Table I and II, but it reports the average linkage distance for the unknown attribute rather than the minimum linkage distance. Similarly, Figure 7 is analogous to Figure 4 above; the only change is that, instead of the minimum linkage distance, the dashed curve reports the mean linkage distance between $\mathbf{X}$ and $\mathbf{Y}$ for the unknown attribute.

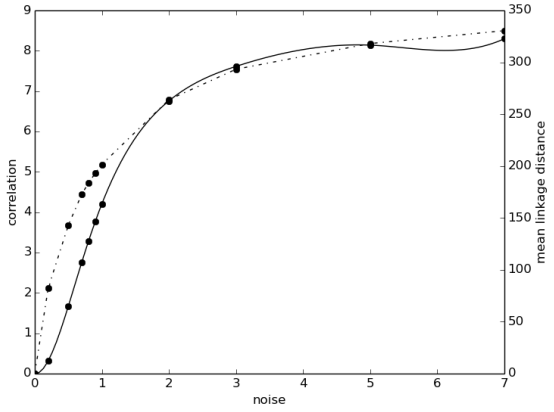

Fig. 7. Solid curve: variation in the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$ (expressed by the Frobenius norm). Dashed curve: variation of the mean linkage distance for the unknown attribute between $\mathbf{X}$ and $\mathbf{Y}$. Both curves are functions of the noise being added (expressed by the ratio between the standard deviation of the noise and the standard deviation of each attribute in $\mathbf{X}$).

TABLE III. MEAN (x^m, y^m) LINKAGE DISTANCE (MEAN RANK DIFERENCE OF UNKNOWN ATTRIBUTE VALUES IN LINKED PAIRS), KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE DISTRIBUTION OF THE (x_m, y_m) RANK DIFFERENCES AND THE DISTRIBUTION OF THE (x_m, y'_m) RANK DIFFERENCES, KOLMOGOROV-SMIRNOV DISTANCE BETWEEN THE DISTRIBUTION OF THE (x^σ_m, y^σ_m) RANK DIFFERENCES AND THE DISTRIBUTION OF THE $(x^\sigma_m, y^{\sigma\prime}_m)$ RANK DIFFERENCE, FOR SEVERAL LEVELS OF NOISE.

	$Y_{0.5}$	Y_1	Y_3	Y_7
mean linkage distance X, Y	144	207	304	342
KS betw. $(\mathbf{X}_p, \mathbf{Y}_p)$ and (X^m, Y^m) link. dist.	0.42	0.27	0.09	0.03
KS betw. $(\mathbf{X}^\sigma_p, \mathbf{Y}^\sigma_p)$ and $(X^{\sigma\,m}, Y^{\sigma\,m})$	0.22	0.03	0.003	0.006

VI. CONCLUSIONS AND FUTURE RESEARCH

We have proposed a general method for disclosure risk assessment based on record linkage by a maximum-knowledge attacker. Unlike the usual record linkage approaches, we do not need to make restrictive assumptions on the attacker's knowledge. Our record linkage analysis is based on comparing the distribution of the linkage distances between the original and the anonymized data set against the distribution of linkage distances of a non-disclosive record linkage between one of the two previous data sets and one random data set. The idea is that the more similar both distributions are, the better protected are the anonymized data.

We have presented three specific record linkage tests, each using a different non-disclosive record linkage as a benchmark. Two of the tests are focused on re-identification disclosure risk and one focused on attribute disclosure risk.

The empirical results that we have presented show that the amount of masking noise needed to attain a safe anonymized data set is proportional to the dependency between the attributes of the original data set. The more independent these attributes are, the less noise is needed to anonymize the original data set.

Empirical results also show that achieving perfect anonymization requires a huge amount of noise, which is likely to damage the utility of data almost entirely. Hence, the benchmark based on a non-disclosive linkage should only be taken as a lower bound on the achievable disclosure risk protection. The actual per-record protection against re-identification risk can be measured as the minimum of the linkage distances over all records in the original data set. With data utility in mind, we should aim to a not too small minimum linkage distance rather than to perfect anonymization.

As future research, we plan to use the presented approach to compare the disclosure protection offered by several anonymization methods under various parameterizations (the empirical study presented here is limited to additive noise).

ACKNOWLEDGMENTS AND DISCLAIMER

Thanks go to Aida Calviño for discussions in the initial stage of this work. The following funding sources are gratefully acknowledged: European Commission (project H2020 "CLARUS"), Government of Catalonia (ICREA Acadèmia Prize to the third author and grant 2014 SGR 537) and Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY"), The third author leads the UNESCO Chair in Data Privacy. The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO.

REFERENCES

[1] R. Brand, J. Domingo-Ferrer, and J.M. Mateo-Sanz. Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC.

[2] J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *CoRR*, abs/1501.04186, 2015.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Proceedings of the Third Conference on Theory of Cryptography*, LNCS 3876, pp. 265–284. Springer, 2006.

[4] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[5] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: privacy beyond k-anonymity and l-diversity. In Rada Chirkova, Asuman Dogac, M. Tamer zsu, and Timos K. Sellis, editors, *Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE 2007)*, pp. 106–115. IEEE, 2007.

[6] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), 2007.

[7] K. Muralidhar, R. Sarathy, and J. Domingo-Ferrer. Reverse mapping to preserve the marginal distributions of attributes in masked microdata. In Josep Domingo-Ferrer, ed., *Privacy in Statistical Databases*, LNCS 8744, pp. 105–116. Springer, 2014.

[8] P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.* Technical Report, SRI International, 1998.

[9] J. Soria-Comas and J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. In *Proc. of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2012)*, pp. 1–8. IEEE, 2012.