Assessing Disclosure Risk via Record Linkage by a Maximum-Knowledge Intruder

Jordi Soria-Comas, Sara Ricci, Josep Domingo-Ferrer

UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Maths,
Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia,
{jordi.soria, sara.ricci, josep.domingo}@urv.cat

Abstract. Assessing the risk of disclosure of an anonymized data set prior to public release is of great importance. When a privacy model (e.g. k-anonymity and its extensions, or differential privacy) is used, prior privacy guarantees are enforced. Without the use of privacy models, a posterior evaluation of the risk of disclosure is needed. In fact, a posterior evaluation is advisable even if a privacy model is used. We focus in two types of disclosure risk: identity disclosure and attribute disclosure. Record linkage, which mimics the approach an intruder would take, stands out as a standard approach to evaluate the disclosure risk associated to an anonymized data set. However, record linkage needs to make assumptions about the information available to the intruder. For an intruder with more information than the assumed one, the risk evaluation may not be reliable. To overcome this difficulty, we postulate a maximum-knowledge intruder who knows the entire original data. This maximum-knowledge intruder can perform the best linkage and, thus, is interesting for risk assessment. To evaluate the disclosure risk based on record linkage, instead of counting the number of record successfully linked, we consider the distance between each pair of linked records. The specific distance to be used is determined by the data holder and should be modeled after his notion of similarity between records. Distributions of distances for different anonymization methods can be compared to determine which methods are safer. Furthermore, when comparing against a non-disclosive linkage we get an absolute assessment of the disclosure risk.

1 Introduction

Statistical disclosure control [6] seeks to allow releasing statistical information about subpopulations while preserving the privacy of the individual data subjects to which the released records correspond. Rephrasing Dalenius [2], disclosure happens when the release of some data enables the intruder to learn something about a target individual that he did not know before. Preventing disclosure in the above sense when releasing a statistical database is, however, unfeasible in presence of specific

background knowledge [5]. As a result, on focuses on more modest and specific disclosure notions: identity disclosure and attribute disclosure.

Record linkage is a technique that is especially relevant in both types of disclosure. Essentially, with the support of some external non-de-identified data set, the intruder tries to assign an identity to the records in the anonymized data set. Identity disclosure happens when the intruder is able to successfully assign an identity to a record in the anonymized data set. Even without identity disclosure, it could be possible to learn confidential data about a specific individual if the intruder is able to trace back the identity to a groups of records with small variability for the confidential attribute under consideration.

In general it is not possible for an intruder to be sure about the correctness of the linkage for a specific anonymized record. However, we show that the intruder can assess whether the linkages he has obtained look plausibly random or not; if not, it means the linkages are likely to be good ones. In this paper we propose several approaches for the intruder to assess whether his linkages are plausibly random ones, based on the comparison of distributions of record linkage distances. These approaches can also be used by the data protector to make sure he has sufficiently protected the data to make all linkages look plausibly random.

1.1 Contribution and Plan of this Paper

Section 2 introduces some concepts that are used in the rest of this paper. Section 3 discusses the effect of background knowledge in record linkage attacks and proposes an intruder model of maximum knowledge. Section 4 discusses how to compare the relative strength of different anonymization methods based on record linkage. Section 5 shows how to turn the relative measure of disclosure risk into a measure for assessing identity disclosure and attribute disclosure. Section 6 reports some experimental results. Conclusions are summarized in Section 7.

2 Background

2.1 Permutation Distance

The permutation distance [7, 3] 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

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

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 $\mathbf{Y} = (Y^1, \ldots, Y^m)$. Assuming that the original

and anonymized data sets have the same number of records, we can consider the permutation distance $d(\mathbf{x}, \mathbf{y})$ as the maximum of the rank distances of corresponding attribute values, that is

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq m} |rank_{X^i}(x^i) - rank_{Y^i}(y^i)| \tag{2}$$

3 Maximum-Knowledge Intruder Model

In record linkage, an intruder links the records in the anonymized data set to an external non-de-identified data set (that contains the intruder's background knowledge), in order to assign an identity to the anonymized records. The linkage is usually based on a set of attributes that are common to both the anonymized and the external data sets. In an attempt to maximize the accuracy of the linkage, the intruder uses all the background knowledge available to her.

When the record linkage is performed by the data protector to assess the risk of disclosure by mimicking the intruder, the protector needs to make assumptions regarding the intruder's background knowledge. If the intruder actually has more background knowledge than assumed by the data protector, the risk of disclosure may be underestimated. To forestall this, we assume a maximum-knowledge intruder.

To assess the risk of identity disclosure, we assume that the intruder 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 intruder, it will be safe against any other intruder.

In attribute disclosure, the intruder's goal is to learn a the value of a confidential attribute for a specific individual. To assess the risk of attribute disclosure, we assume that, except for the attribute he is trying to learn in the attack, the intruder knows the original values of all other attributes.

4 A Relative Measure of Disclosure Risk

The result of record linkage depends on two main factors: the background knowledge and the linkage strategy used. In Section 3 a maximum knowledge intruder was stated not to underestimate the risk of disclosure. Regarding the linkage strategy, the experiments in Section 6 are based on minimizing the permutation distance. While both are important choices when performing record linkage, this section is independent of the actual choices being made.

For each record $\mathbf{x} \in \mathbf{X}$ we define its linked record $\mathbf{y_x} \in \mathbf{Y}$ as one of the anonymized records in $\mathbf{Y}$ at shortest distance from $\mathbf{x}$. We assume that a function $M_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y_x})$ measuring the amount of masking between $\mathbf{x}$ and $\mathbf{y_x}$ is available. This function can be defined in terms of the linkage distance or as a completely

different function.

To measure the relative disclosure risk between two different anonymizations $\mathbf{Y}_1$, $\mathbf{Y}_2$ of the same original data set $\mathbf{X}$, we compare the distributions of masking amounts in the linked record pairs between $\mathbf{X}$ and $\mathbf{Y}_1$ and in the linked pairs between $\mathbf{X}$ and $\mathbf{Y}_2$. Let these two distributions be, respectively, $dist_M(\mathbf{X}, \mathbf{Y}_1)$ and $dist_M(\mathbf{X}, \mathbf{Y}_2)$. The more information the anonymized data set contains about the original data, the smaller is the masking amount, and vice versa. Thus we can compare the risk of disclosure of both anonymizations by comparing the distributions of masking amounts.

If $dist_M(\mathbf{X}, \mathbf{Y}_1)$ and $dist_M(\mathbf{X}, \mathbf{Y}_2)$ are not significantly different, then there is no evidence that any of the anonymized data sets contains more information about $\mathbf{X}$. Thus, we conclude that $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are equally good as anonymizations of $\mathbf{X}$.

If there is evidence that one of the distributions of masking amounts (say $dist_M(\mathbf{X}, \mathbf{Y}_1)$) is more biased towards smaller values, then we can conclude that the corresponding anonymized data set ($\mathbf{Y}_1$ in this case) contains more information about the original data and, thus, incurs a greater risk of disclosure.

5 Disclosure Risk via Record Linkage: Framework and Tests

In this section we are interested in providing an absolute assessment of the disclosure risk, rather than comparing the relative risk of two different anonymizations. This is done by comparing the distribution of masking amounts $dist_M(\mathbf{X}, \mathbf{Y})$, between the original data set and the anonymized data set, to a distribution of masking amounts that is known to incur in no risk of disclosure.

5.1 Dictionary Linkage Test

In this test, we seek to assess the risk of record re-identification against maximum-knowledge intruders. To this end, we compare the distribution of masking amounts between $\mathbf{X}$ and $\mathbf{Y}$ to the distribution of masking amounts between $\mathbf{D_X}$ and $\mathbf{Y}$, where $\mathbf{D_X}$ is an artificially created data set that we call the *dictionary* of $\mathbf{X}$. Specifically, $\mathbf{D_X}$ 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.

Note that because $\mathbf{D_X}$ contains all possible combinations of attribute values, any dependency between attributes in $\mathbf{X}$ is destroyed. In other words, the only information about $\mathbf{X}$ that $\mathbf{D_X}$ preserves are the marginals of the attributes, which we consider to be non-disclosive. Thus, the linkage between $\mathbf{D_X}$ and $\mathbf{Y}$ reveals nothing about $\mathbf{X}$.

By comparing the distributions of linkage masking amounts $dist_M(\mathbf{X}, \mathbf{Y})$ and $dist_M(\mathbf{D_X}, \mathbf{Y})$, we get an absolute assessment of the re-identification risk.

5.2 Linkage to Permuted Data Set

Similarly to the dictionary linkage test, we propose here another test to assess the risk of record re-identification against maximum-knowledge intruders.

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. 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.

We can say that $\mathbf{Y}'$ contains no sensitive information (recall that we assume that marginal distributions are not sensitive, only the relation between attributes is). Therefore, we can take the distribution $dist_M(\mathbf{X}, \mathbf{Y}')$ of masking amounts between $\mathbf{X}$ and $\mathbf{Y}'$ as the baseline to compare with $dist_M(\mathbf{X}, \mathbf{Y})$.

5.3 Attribute Disclosure Test

In this section we seek to assess the risk of record attribute disclosure against maximum-knowledge intruders. Recall that, for attribute disclosure, we assume that, except for the attribute he is trying to learn, the intruder knows the original values of all other attributes.

The attribute disclosure test assesses the potential of an intruder to correctly determine the value of the unknown attribute. Let $\mathbf{X} = (X^1, \ldots, X^m)$ be the original data set and $\mathbf{Y} = (Y^1, \ldots, Y^m)$ be the anonymized one. Without loss of generality, let the unknown attribute be X^m. The intruder performs a record linkage (*e.g.* based on the permutation distance) using only the first $m-1$ attributes of $\mathbf{X}$ and $\mathbf{Y}$. Unlike in the previous test, where the distance used in the linkage was meaningful to evaluate the re-identification risk, here the target attribute X^m is not involved in the record linkage; thus, the specific linkage distance is meaningless for the risk of attribute disclosure. For this reason, we consider a different function to measure the amount of anonymization: we take $M_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y_x}) = |rank_{X^m}(\mathbf{x}) - rank_{Y^m}(\mathbf{y_x})|$.

Also in this case, we need a "non-disclosive linkag" that is obtained by adapting the linkage to permuted data set test to this attack: the reference distribution is drawn from the attribute disclosure test applied to $\mathbf{X}$ and $\mathbf{Y}'$, where $\mathbf{Y}'$ is the data

$\mathbf{X}$

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

Figure 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).

set obtained by applying a random permutation to each attribute of $\mathbf{Y}$.

6 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. Since it was unfeasible to generate an exhaustive dictionary for all the "Census" attributes, we took as $\mathbf{D_X}$ a random sample of 10,000 records from the exhaustive dictionary.

6.1 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 $\mathbf{X}$. In the first comparison, the anonymized data sets have been generated by adding independent normally distributed noise to each of the attributes. 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.

In the second comparison, we want to compare a method using a privacy model and one without privacy model. Hence, the anonymized data sets have been generated by applying k-anonymity based on microaggregation [4] and ϵ-differential privacy [8]. We denote by $\mathbf{W}_p$ the anonymized data set obtained using microaggregation, where p is the number of clusters created (we take $p = 15$ and $p = 50$). The differential privacy approach we applied works as follows: each attribute is microaggregated and the original values replaced by the corresponding centroid, then a Laplace noise is added. To explore how the distribution of linkage distances changes

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

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

Figure 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 $\mathcal{S}_n$.

with different values of ϵ, we consider three different anonymized data sets: $\mathbf{Z}_{0.005}$, $\mathbf{Z}_{0.05}$ and $\mathbf{Z}_{0.5}$, where ϵ takes values $\{0.005, 0.05, 0.05\}$.

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.

6.2 Figures

The Figures 3, 4 and 5 show both distributions for different anonymizations:

- 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 curves labeled "gen num_part" correspond to the distribution of the linkage distances between the original data set $\mathbf{X}$ and the anonymized data sets $\mathbf{W}_{15}$ and $\mathbf{W}_{50}$, respectively.

- The curves labeled with "diff_eps" correspond to the distribution of the linkage distances between the original data set $\mathbf{X}$ and the anonymized data sets $\mathbf{Z}_{0.005}$, $\mathbf{Z}_{0.05}$ and $\mathbf{Z}_{0.5}$, respectively.

- The curve labeled "dictionary" corresponds to the distribution of the linkage distances between the dictionary data set and the anonymized data set.

- The curves labeled "permuted" or "permutated" correspond to the distributions of the linkage distances between $\mathbf{X}$ and the permuted versions of the anonymized data set.

The curves labeled with "dictionary","permutated" or "permuted" are not disclosive, and hence, these distributions are practically the same no matter the amount of anonymization. Thus, we just plotted one curve in every chart. In the left-hand side charts of Figures 3, 4 and 5, we used the "Census" data set as original data set; instead, in the right-hand side charts of the same figures, we used a random permutation of the "Census" data set.

In Figure 6, 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

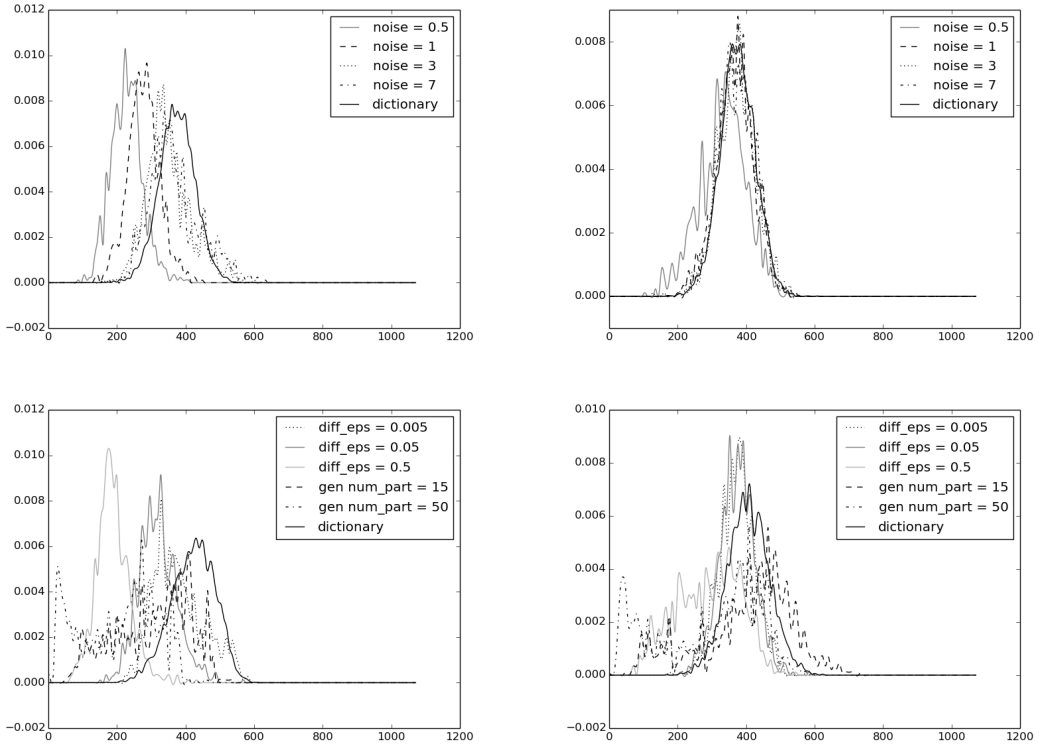

Figure 3: Dictionary linkage test

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 curves in Figure 6. 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 $\mathbf{X}$.

6.3 Results of the tests

From the left charts of Figures 3, 4 and 5 it can be seen that a lot of noise (ϵ is inversely proportional to the noise) is needed for the distribution of linkage distances between $\mathbf{X}$ and $\mathbf{Y}$ (or $\mathbf{Z}$) to be similar to the one between $\mathbf{D_X}$ and $\mathbf{Y}$; (or $\mathbf{Z}$), that is, for the 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 $\mathbf{D_X}$ 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 right charts of Figure 3, 4 and 5 are analogous to the left charts, 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

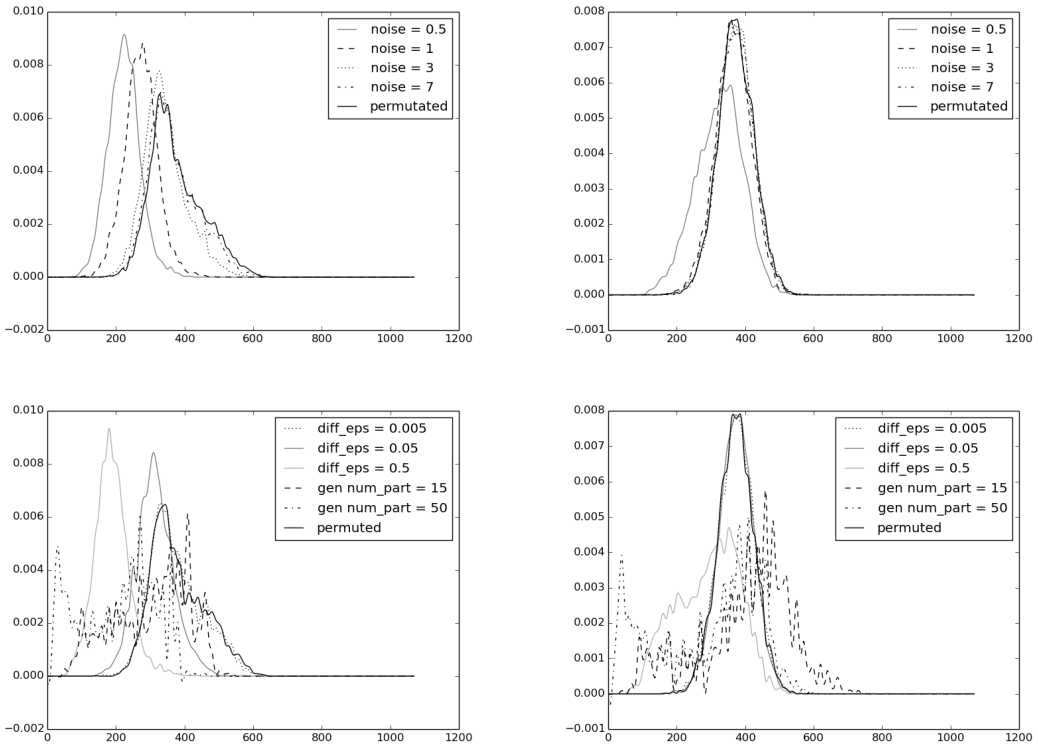

Figure 4: Linkage to permuted data set test

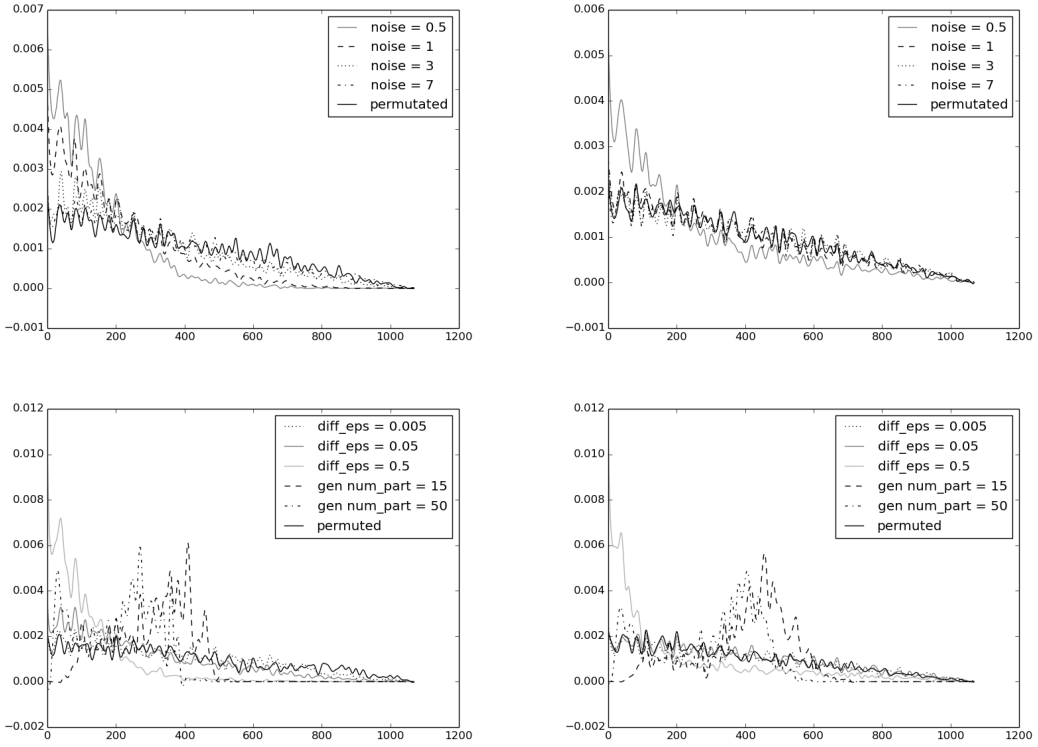

Figure 5: Attribute disclosure test

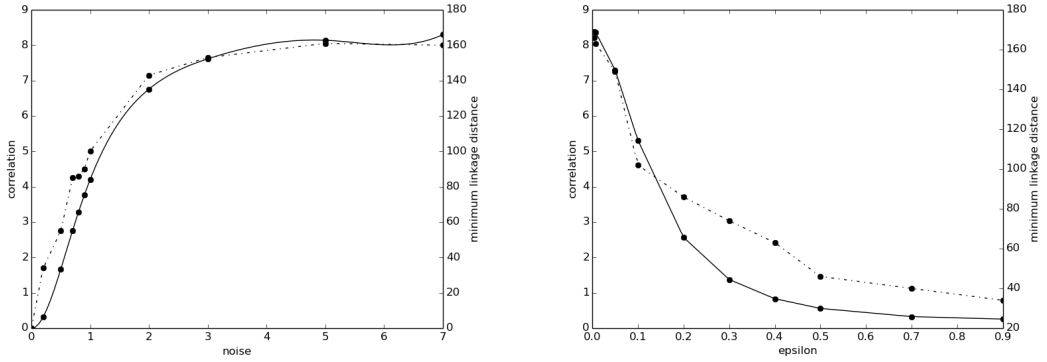

Figure 6: Correlation Test. In the left figure anonymization is attained through addition of noise while in the right figure differential privacy has been used.

in $\mathbf{X}^\sigma$ can be thought as being nearly independent from each other. For instance, the right charts of Figure 3 shows that a moderate amount of noise ensures already perfect anonymization: indeed, the distribution of linkage distances between $\mathbf{X}^\sigma$ and $\mathbf{Y}_1^\sigma$ 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 perfect anonymization. In fact, the non-disclosive distribution is unattainable if one wants to preserve the correlations of the original data set.

Figures 3, 4 and 5 show that microaggregation gives better linkage than differential privacy, and therefore is more disclosive. This comes as no surprise, since the differentially private data set was computed by adding noise to the result of the microaggregation. For differential privacy, the more noise is added (smaller ϵ), the worst is linkage.

Figure 6 shows the effect of anonymization on both correlation between attributes and minimum linkage distance. In case of small or no anonymization (small noise or large ϵ) the utility is preserved (solid curves are close to zero), but a record linkage attack is effective; hence, there are no privacy guarantees. The more noise is added, the less utility we have, and the more privacy is gained.

7 Conclusions

We have proposed a general method for disclosure risk assessment based on record linkage by a maximum-knowledge intruder. Unlike the usual record linkage approaches, we do not need to make restrictive assumptions on the intruder's knowledge. Our record linkage analysis is based on comparing the distribution of the linkage distances. Thus, we can determine the relative strength of anonymization methods. Also, by comparing against the distribution of linkage distances of a non-

disclosive record linkage we get an absolute measure of disclosure risk.

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.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: Government of Catalonia (ICREA Acadèmia Prize to the last author and grant 2014 SGR 537), Spanish Government (projects TIN2011-27076-C03-01 "CO-PRIVACY" and TIN2014-57364-C2-R "SmartGlacis"), and European Commission (project H2020 644024 "CLARUS"), The authors are with 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] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15:429-444, 1977.

[3] 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.

[4] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.

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

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

[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] D. Sánchez, J. Domingo-Ferrer and S. Martínez, Improving the utility of differential privacy via univariate microaggregation. In J. Domingo-Ferrer, ed., *Privacy in Statistical Databases-PSD 2014*, LNCS 8744, pp. 130-142. Springer, 2014.