Rank-Based Record Linkage for Re-Identification Risk Assessment

Krishnamurty Muralidhar[1] and Josep Domingo-Ferrer[2]

[1]Dept. of Marketing & Supply Chain Management, Price College of Business, University of Oklahoma, Norman, OK, USA (krishm@ou.edu)
[2]UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Catalonia (josep.domingo@urv.cat)

Abstract. There is a tradition of data administrators using record linkage to assess the re-identification risk before releasing anonymized microdata sets. In this paper we describe a record linkage procedure based on ranks, and we compare the performance of this rank-based record linkage against the more usual distance-based record linkage to re-identify records masked using several different masking methods. We try to elicit the reasons why RBRL performs better than DBRL for certain methods and worse than DBRL for other methods.

Keywords: Data masking, Rank-based record linkage, Re-identification risk

1 Introduction

In statistical disclosure limitation, record linkage procedures are used to assess the effectiveness of the masking mechanism in preventing re-identification by the adversary [15], that is, preventing the adversary from linking a particular masked record in the anonymized microdata set with the corresponding original record in the original microdata set. Such an assessment makes sense because an adversary may have access to the values of some of the original attributes together *with identification information* from an auxiliary source; the adversary could then attempt to link this identified original record to the masked records using the original attributes whose values he knows (these attributes are called quasi-identifiers). If successful, the adversary would have linked a masked record as belonging to an identified subject: in that case, the adversary could link the confidential attributes in the masked record (e.g. health condition, salary, religion, etc.) to the subject's identity. This is clearly a privacy violation whose risk needs to be evaluated.

Record linkage procedures are often implemented assuming knowledge of the entire original and masked data sets. The procedure then attempts to link records from the masked data set to the original data set. The level of information available to the adversary is one common disagreement when performing record linkage. Data administrators often argue that, to realistically assess disclosure risk, record linkage should assume a level of knowledge by the adversary that is plausible in practice. In

particular the claim that, without knowledge of the actual correspondence between original and masked records, the adversary cannot verify the accuracy of his record linkage, that is, he cannot decide whether a linked pair (original record, masked record) is a correct match. For these reasons, they dismiss record linkage results as being worst-case and hence overly pessimistic. While there is considerable merit in these arguments, we view record linkage as being useful for two reasons. First, an assessment of the worst-case disclosure scenario is, in our opinion, important information for the data administrator. Second, even if record linkage is not used for assessing the risk of disclosure, it serves an important purpose by providing the administrator with a means to compare the disclosure protection offered by different masking mechanisms. Hence, for the purposes of this study, we will stick to the tradition and perform record linkage assuming that the entire original and masked data are available.

Our objective in this paper is to evaluate the performance of an alternative approach for record linkage, namely, perform record linkage based on the ranks of the original and masked data. Admittedly, we need to assume that the values of attributes can be ranked, but such an assumption is less problematic than it appears: for numerical or categorical ordinal attributes, ranking is straightforward; for categorical nominal attributes, semantic distance metrics are available that can be used to rank them (for instance, the marginality distance [3, 14]). While rank based record linkage (RBRL) has been used in the literature, there are no comprehensive evaluations of its performance. This is surprising considering that there are rank-based procedures (such as rank swapping [12,9]) that are frequently used to mask numerical data. When assessing identity disclosure, the commonly used record linkage procedure is distance-based record linkage (DBRL). Our goal is to investigate the application of RBRL for re-identification of masked data and compare its performance to DBRL.

The rest of this paper is structured as follows. Section 2 gives some background on record linkage, including probabilistic and distance-based record linkage. Section 3 describes rank-based record linkage. Section 4 presents empirical results of apply RBRL to re-identify records masked using several different masking methods. Conclusions and future research directions are summarized in Section 5.

2 Background on Record Linkage

Record linkage procedures are used to measure the ability of an adversary to link the original and the masked records. The most common methods of record linkage are probabilistic record linkage and distance based record linkage [15].

Distance-based record linkage (DBRL) methods link a masked record with the original record at the shortest distance, for some previously chosen distance. The main difficulty with such methods is to establish appropriate distances for the attributes under consideration: for numerical attributes, the Euclidean distance seems a natural option, but for non-numerical attributes deciding on a distance is less obvious (although, as mentioned above, distances for categorical values are available). An advantage of DBRL is that it allows including subjective information in the re-identification process (e.g. the attributes can be given different weight when computing distances, that is, those attributes that are considered more relevant for re-identification can be given more weight).

Probabilistic record linkage [7, 10, 16], unlike DBRL, does not require standardizing all attributes to a common range or specifying weights for each attribute or choosing a distance. The user (the data administrator in our case) only needs to provide two probabilities as input: the maximum acceptable probability that a pair of records that is not a real match is classified as a linked pair (probability of false positive) and the maximum acceptable probability that a pair of records that is a real match is classified as a non-linked pair (probability of false negative). Probabilistic record linkage works for any type of attribute (numerical, categorical, etc.) without any special adaptation. However, the linkage algorithm itself is less simple and less intuitive than DBRL, which remains the most usual choice if all attributes are numerical.

As mentioned above, for numerical data the most common choice is the simplest one: DBRL with the Euclidean distance. This is what we will take as a benchmark for comparison with our rank-based record linkage. In practice, DBRL is implemented as follows. Let $X = \left(x_{ij}; i = 1, 2, \dots, n; j = 1, 2, \dots, m\right)$ represent the original data set and let $Y = \left(y_{lj}; l = 1, 2, \dots, n; j = 1, 2, \dots, m\right)$ represent the masked data set, where n is the number of records and m is the number of attributes. In DBRL, for a given original record, we compute the Euclidean distance to this record from every masked record. However, since the variance of the individual attributes can be different, the distance is measured using the standardized (to mean 0 and unit variance) attributes, rather than the original ones. Let $x_{i*} = (x_{i1}, x_{i2}, \dots, x_{im})$ represent the target original record. Then for each masked record y_{l*} we compute the distance measure as

$$d_{il}^2 = \sum_{j=1}^{m} \left\{ \left(\frac{x_{ij} - \mu_{X_j}}{\sigma_{X_j}} \right) - \left(\frac{y_{lj} - \mu_{Y_j}}{\sigma_{Y_j}} \right) \right\}^2 .$$

Then x_{i*} is linked to the masked record l with $Min(d_{il}^2)$. The process is then repeated for every record i in the original data. Many variations of the basic DBRL have been proposed. A comprehensive discussion of these variations is beyond the scope of this paper. We refer the interested reader to Winkler [16].

3 Rank-Based Record Linkage

The rationale for developing a rank-based procedure is predicated on the simple premise that many masking techniques such as rank swapping are based on the ranks. For these techniques, it makes more sense to consider a rank-based record linkage procedure rather than traditional (distance-based) record linkage procedures based on magnitude. Specialized procedures based on ranks have been developed to assess the disclosure risk characteristics of rank swapping [13] and micro-aggregation [13, 17]. The interval-based procedure described in [11] can also be considered as rank based as well: given the value of a masked attribute, the procedure checks whether the corresponding original value falls within an interval centered on the masked value, where the interval width can be in terms of rank (a rank difference).

While the above rank-based procedures have been used in practice, they have been designed for application to specific procedures and/or specific data sets. There have been no general comparisons of the performance of RBRL and that of DBRL. In this study we offer such a comparison.

Rank Based Record Linkage Description

Let $\boldsymbol{R} = (r_{ij})$ and $\boldsymbol{S} = (s_{lj})$ represent the rank matrices corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively; that is, r_{ij} represents the rank of record i in $\boldsymbol{X}$ with respect to attribute j; similarly, s_{lj} represents the rank of record l in $\boldsymbol{Y}$ with respect to attribute j. The procedure is implemented as follows:

$For\ i\ =\ 1\ to\ n$
    $For\ l\ =\ 1\ to\ n$
        $Compute\ d_{il} = Criterion\big(Abs(r_{i1} - s_{l1}), \dots, Abs(r_{im} - s_{lm})\big)$
    $Next\ l$
    $Linked\ index\ of\ i\ = \arg min_l\ (d_{il})$
$Next\ i$

The procedure described above can be considered as the *standardized* L_1 metric version of DBRL performed on the rank matrices $\boldsymbol{R}$ and $\boldsymbol{S}$ rather than $\boldsymbol{X}$ and $\boldsymbol{Y}$. The use of ranks has several advantages. First, no further scaling or standardization is necessary to account for the difference in the magnitudes. Second, while DBRL is almost always implemented using the minimum distance criterion, rank-based procedures lend themselves to different criteria for selecting the match based on the characteristics of the masking procedure.

In this study, we consider two criteria, namely Sum and $Maximum$. The rationale for the Sum criterion is evident – it is simply the application of the "minimum sum of squared distance" criterion for DBRL. The $Maximum$ criterion is intended for procedures such as rank swapping where the records are swapped but are subject to an upper bound in the swap. In these cases, using the Sum criterion may be misleading. Consider rank swapping with an upper bound of p (that is, only records within a rank difference of p are eligible to be swapped). Even if p is not released to the public, if the adversary knows that rank swapping has been performed, then it is clear that the absolute difference in rank between the target record and a masked record has an upper bound (albeit unknown). Hence, the adversary would use the maximum (across all attributes) as the measure of match and then choose the masked record that has the minimum (across all masked records) of the maximum (across all attributes) as the appropriate match.

In the following section, we describe an experimental investigation to compare the performance of DBRL and RBRL across several masking methods.

4 Experimental Investigation

The objective of this experimental investigation is to evaluate the effectiveness of RBRL and DBRL at linking masked records to original records for different masking mechanisms. The masking mechanisms that we choose are among the commonly used ones, namely: independent additive noise, correlated additive noise, multiplicative noise, rank swapping, and microaggregation. This choice of mechanisms represents well some of the diversity of principles used in microdata masking: noise-based, rank-based and cluster-based. We have used as experimental data sets [1]: (1) the Census data set , consisting of 1080 records and 13 numerical attributes, and (2) the EIA data set consisting of 4092 records and 10 numerical attributes. These data sets have been used many times in the literature to evaluate masking mechanisms in terms of utility preservation and disclosure risk protection.

In each of our experiments, we first generated the masked version of the data using a specific masking mechanism; then DBRL and RBRL (both versions) were used to perform record linkage; finally, the number of records correctly linked by each of the approaches was recorded. For masking methods involving randomness (all except microaggregation), the experiment was replicated by generating 100 masked versions of the original data; the results report the percentage of correctly linked records on average over the 100 replications. Since microaggregation involves no randomness, the corresponding experiment was performed only once, and the result is the percentage of correctly linked records in that single run.

Additive Noise

First we consider independent additive noise which is implemented as follows. The original observation is modified by adding noise independently for each attribute to result in the masked observation:

$$y_{ij} = x_{ij} + e_{ij};\ e_{ij} \sim N\left(0, a\sigma_j\right),$$

where a represents the perturbation level and $\mathrm{N}\left(0, a\sigma_{\mathrm{j}}\right)$ represents a normal distribution with mean 0 and the specified standard deviation, with σ_{j} being the standard deviation of X_j. Since the noise terms are generated independently for each attribute, the correlation between the noise terms is zero. This can also be viewed as univariate noise addition where the noise added is based only on the standard deviation σ_{j} of the given attribute (and independent of the correlation between the attributes).

Given that DBRL is almost exactly the reversal of the noise addition procedure, we expected DBRL to perform better than RBRL for masking using independent noise. The results in Table 1 confirm this. For every perturbation level, DBRL outperforms both RBRL-Sum and RBRL-Max (with the Sum and $Maximum$ criteria, respectively). In some cases (Census data, $a = 0.1$), the difference is large with DBRL correctly identifying 98.3% versus RBRL-Sum correctly identifying 90.0% of the records. For higher levels of perturbation, the difference is smaller, but consistent.

The re-identification by all methods is much lower for the EIA data. However, in terms of relative performance, the results for the EIA data are very similar to the ones for the Census data: DBRL consistently outperforms both RBRL methods. Thus, these results show that for independent noise addition, DBRL is the best record linkage approach.

Between the two RBRL approaches, Table 1 shows that RBRL-Sum consistently performs better than RBRL-Max: there is a significant drop in the percentage of correctly linked records when using RBRL-Max compared to RBRL-Sum. This again is not surprising since the $Maximum$ criterion was intended to perform well only in specific scenarios (such as data swapping).

Table 1. Percentage of correctly linked records with distance-based and rank-based record linkage for independent additive noise and several values of the perturbation level a

a	Census Data			EIA Data		
	DBRL	RBRL-Sum	RBRL-Max	DBRL	RBRL-Sum	RBRL-Max
0.10	98.4%	90.0%	57.4%	20.2%	12.1%	7.9%
0.25	69.4%	53.7%	33.7%	7.6%	3.7%	2.3%
0.50	26.4%	18.9%	12.6%	2.5%	1.2%	0.7%
0.75	11.6%	7.8%	5.4%	1.1%	0.6%	0.4%
1.00	6.2%	4.2%	2.9%	0.6%	0.3%	0.2%

An alternative to independent noise addition is correlated noise, where the noise added has the same correlation structure as the original attributes as follows:

$$\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{e}_i;\ \boldsymbol{e_i} \sim N(\mathbf{0}, a\boldsymbol{\Sigma_{XX}})$$

where $\boldsymbol{x}_i$ is the i-th original record, $\boldsymbol{y}_i$ is the corresponding masked record, $\mathbf{e}_i$ is a noise vector and $N(\mathbf{0}, a\boldsymbol{\Sigma_{XX}})$ represents a multivariate normal distribution with mean vector equal to $\mathbf{0}$ for all attributes and covariance matrix a times $\boldsymbol{\Sigma_{XX}}$, with the latter being the covariance matrix of the original data set. With correlated noise, we observe a smaller percentage of records correctly linked at all perturbation levels. Comparing the different methods of linkage, we find that the results for correlated noise are very similar to those observed for independent noise. Due to space limitations, we have not included these results.

Rank Swapping

For numerical data, data swapping is usually implemented by swapping values based on the ranks [12]. The swapping parameter p (the maximum allowable distance between the ranks of the swapped values expressed as a percentage of n) is

specified. For each attribute j, the rank of a given record is computed and swapped with another record within rank distance $\pm np/100$ (with respect to attribute j). The process is repeated for all records, and for all attributes (independently for each attribute). As we mentioned earlier, the primary reason to develop the RBRL approach is to investigate its effectiveness in the context of rank-based masking procedures. In addition, the $Maximum$ criterion was developed specifically for rank swapping. Hence, we expect the performance of RBRL-Max to be the best. The results of record linkage for rank swapping are provided in Table 2.

Table 2. Percentage of correctly linked records with distance-based and rank-based record linkage for rank swapping and several values of the swapping parameter p

p	Census Data			EIA Data		
	DBRL	RBRL-Sum	RBRL-Max	DBRL	RBRL-Sum	RBRL-Max
1%	98.8%	100.0%	100.0%	75.6%	85.8%	93.6%
5%	88.8%	99.4%	100.0%	13.8%	18.2%	45.1%
10%	60.1%	84.0%	98.5%	2.5%	4.0%	10.0%
25%	7.2%	10.5%	37.5%	0.2%	0.3%	0.5%
50%	0.6%	0.7%	1.0%	0.1%	0.1%	0.1%
100%	0.1%	0.1%	0.1%			

Table 2 indicates that, as expected, RBRL-Max outperforms the other two methods of record linkage. The interesting aspect is that the percentage of records re-identified with RBRL-Max is higher than with the other two methods by a considerable margin. For the Census data, even when the swapping distance is 25% (records with rank difference as high as 270 are swapped), RBRL-Max correctly links over 37.5% of the records. This is much higher than the results observed for rank swapping by [4] using DBRL. Our results are consistent with those observed by [13]. However, the *ad hoc* re-identification procedure in [13] relied on knowledge of the swapping parameter p. In RBRL, we do not assume that this information is available to the adversary. The performance of RBRL-Max is superior across the board. For the Census data,, *rank swapping provides little protection against rank-based record linkage* (re-identification) unless the swapping parameter is $(n/2)$ or higher. But with this high level of swapping, data utility is likely to be very poor.

For the EIA data, fewer records are correctly linked across all methods. We again find that RBRL-Max performs significantly better than the other two procedures.

RBRL-Sum also outperforms DBRL for both data sets across all swapping levels. This indicates that for rank swapping, a rank-based record linkage approach is

superior to a distance-based one. While this may not seem a surprising result, what is surprising is that we do not see rank-based record linkage normally used in the literature to evaluate rank-based masking.

p-Distribution Rank Swapping

One of the key issues with rank swapping is that p imposes a strict upper bound on the swapping distance. When p is known, Nin et al. [13] showed that this could allow the adversary to link swapped records to the original ones. Even when p is unknown, the results above indicate that this is a serious threat. To overcome this problem, [13] proposed a modified rank swapping mechanism, which they call p-Distribution Rank Swapping. In this approach, rather than using p as a constant maximum swapping distance for all records, they use it as a parameter for a normal distribution with $\mu = \sigma = p/2$, from which, for each record, they sample a maximum swapping distance for that record, with the usual restriction that one cannot swap the current record to an index below 0 or greater than n; then the actual swapping distance for the record is randomly chosen between 0 and the record's maximum swapping distance. The advantage of this approach is that there is no strict upper bound on the swapping distance across the records. Theoretically, it is possible for the swap value for a given record to be as high as n. When this modified rank swapping mechanism is implemented, the procedure that [13] used for re-identification performs poorly. In this section, we investigate the performance of the three record linkage procedures for p-distribution rank swapping.

Table 3. Percentage of correctly linked records with distance-based and rank-based record linkage for p-distribution rank swapping and several values of parameter p

p	Census Data			EIA Data		
	DBRL	RBRL-Sum	RBRL-Max	DBRL	RBRL-Sum	RBRL-Max
1%	98.4%	100.0%	100.0%			
5%	81.3%	97.4%	99.1%			
10%	40.8%	63.5%	81.9%			
25%	2.9%	3.7%	10.1%			
50%	0.2%	0.2%	0.5%			
100%	0.0%	0.0%	0.0%			

Table 3 provides the percentage of correct linkage resulting from the application of the three linkage procedures for selected parameter values of p-distribution rank swapping. The results indicate that, as before, RBRL-Max performs better than the other three approaches. Not surprisingly, the percentages of correct linkages are

somewhat lower than those obtained for simple rank swapping (Table 2), since some records could possibly have been swapped by a distance potentially as high as n. While the average swapping distance for a given value of p is likely to be the same as before, the variability introduced by the use of a random variable does lower the record linkage success. In terms of relative performance among the record linkage procedures, the results are similar to those of simple rank swapping: RBRL-Max correctly links the most records, followed by RBRL-Sum, and DBRL correctly links the fewest records.

Multiplicative noise

One of the features of noise addition is that the magnitude of the noise is independent of the magnitude of the original value. When the original values are from a skewed distribution, this leads to a situation where too much noise is added to small values and too little noise is added to large values. To overcome this problem, multiplicative noise has been used to mask data as follows:

$$y_{ij} = x_{ij} \times e_{ij}$$

The selection of the distribution of e_{ij} must be done carefully so as to ensure that the resulting masked values are in the appropriate range. In this study, we use a simple distribution for $e_{ij} \sim Uniform(1-b, 1+b)$ where b represents the perturbation level. Multiplicative perturbation is performed independently for each attribute. The results of applying the three record linkage procedures to multiplicative perturbation are provided in Table 4.

Table 4. Percentage of correctly linked records with distance-based and rank-based record linkage for multiplicative noise and several values of the perturbation level b.

b	Census Data			EIA Data		
	DBRL	RBRL-Sum	RBRL-Max	DBRL	RBRL-Sum	RBRL-Max
0.10	99.0%	99.7%	98.9%	64.2%	76.5%	79.4%
0.25	64.7%	81.5%	73.5%	19.2%	32.6%	38.5%
0.50	18.2%	31.5%	22.5%	4.1%	11.4%	13.7%
0.75	6.1%	10.7%	7.9%	1.5%	4.6%	5.6%
1.00	3.0%	4.1%	3.5%	0.8%	1.7%	2.0%

We had hypothesized that DBRL would perform better for magnitude-based masking mechanisms and RBRL (Sum or Max) would perform better for rank-based masking mechanisms. Hence, for multiplicative noise (which is magnitude based), we

expected DBRL to perform better than both RBRL procedures. However, Table 4 indicates that RBRL performs better than DBRL for multiplicative perturbation. For the Census data, across all perturbation levels, RBRL-Sum results in a higher percentage of correctly linked records, in some cases, by a large margin. For instance, for the Census data, when $b = 0.25$, RBRL-Sum correctly links approximately 81.5% of the records, while DBRL correctly links only 64.7% of the records. The performance of RBRL-Max is in the middle, correctly linking 73.5% of the records. This is observed consistently across all perturbation levels for the Census data (with the exception of the perturbation level of 0.05 where RBRL-Max performs slightly worse than DBRL).

As before, for the EIA data set, fewer records are correctly linked for every perturbation level. RBRL performs better than DBRL for this data set as well. Between the two RBRL methods, RBRL-Max performs consistently better than RBRL-Sum. This is different from the results observed for the Census data and somewhat surprising. But perhaps more surprising is the fact that RBRL methods outperform DBRL in all cases across both data sets. This result was not expected and deserves further investigation.

Microaggregation

The last masking mechanism that we considered is individual-ranking microaggregation. In this mechanism, an aggregation parameter k is specified. For each attribute, the individual original values are replaced by the average of the k closest neighboring values. The procedure is then repeated for each attribute. Table 5 shows the record linkage results using the three approaches for different values of k. Unlike the previous methods where a common masking parameter was used across both data sets, for microaggregation we used different values of k for the two data sets (larger values for the larger data set). As indicated earlier, there is no random component in the implementation of microaggregation. Hence, the results presented correspond to a single run of the mechanism on the Census data.

For the Census data, Table 5 indicates that both RBRL methods outperform DBRL by a significant margin (except when k is small, in which case all methods are equally successful). For instance, with $k = 216$, DBRL correctly links 60.2% of the records while both RBRL methods correctly link approximately 96.5% of the records. Even with what would be considered an extremely high value of k $(= 360)$, the two RBRL approaches correctly link 65% of the records in the Census data. This again confirms the conclusions of [2,4] that individual-ranking microaggregation is highly susceptible to re-identification (in contrast, multivariate microaggregation satisfies the k-anonymity privacy model and hence offers probability of re-identification $1/k$ [6]).

Table 5. Percentage of correctly linked records with distance-based and rank-based record linkage for microaggregation and several values of the aggregation parameter k.

Census Data				EIA Data			
k	DBRL	RBRL-Sum	RBRL-Max	k	DBRL	RBRL-Sum	RBRL-Max
10	99.8%	100.0%	100.0%	341	6.4%	61.4%	61.4%
54	96.5%	100.0%	100.0%	372	5.4%	56.3%	56.3%
108	89.7%	100.0%	100.0%	682	0.9%	30.7%	30.7%
270	38.8%	87.9%	87.9%	1023	0.1%	15.4%	15.4%
360	21.3%	65.0%	65.0%	1364	0.0%	8.7%	8.7%
540	9.6%	24.5%	24.5%	2046	0.0%	3.3%	3.3%

The difference between RBRL and DBRL is even more dramatic for the EIA data. With $k = 341$, DBRL correctly links only a small percentage (6.4%) of the records, while both RBRL methods correctly link almost ten times as many records (61.4%). Across all values of k, we note that both RBRL methods provide significantly higher linkage percentages compared to DBRL methods.

Another result is of particular interest, is that the performance of both RBRL methods is exactly the same; *every correctly linked observation using RBRL-Sum is also correctly linked by RBRL-Max*. The possible explanation for this result is the following. With rank swapping, the masking is performed using only the ranked values. With multiplicative noise, the masking is performed by relative magnitude. Micro-aggregation is a combination of the two – the groups are identified by ranks and then aggregated by magnitude. This combination yields the interesting result that both RBRL methods perform exactly the same.

5 Conclusions and Future Research

We have presented a general rank-based record linkage procedure, as an alternative to distance-based record linkage and probabilistic record linkage. We then have given an empirical performance comparison with distance-based record linkage at re-identifying records masked with several different masking methods. Our results indicate that, whenever a masking method explicitly or implicitly ranks the original data, RBRL outperforms distance-based record linkage. Furthermore, and rather unexpectedly, RBRL also outperforms distance-based record linkage for masking by multiplicative noise. We have considered rank-based record linkage with two different criteria (the sum of rank differences and the maximum rank difference). None of the two criteria outperforms the other for all the methods tried: which one is best de-

pends on the masking method being attacked. Also interesting is the fact that for microaggregation, both RBRL procedures (Sum and Maximum) perform in an identical manner, correctly linking exactly the same records.

Originally, we had hypothesized that distance-based record linkage would perform better for masking procedures based on magnitude (additive and multiplicative noise) and rank based record linkage for masking procedures based on ranks (data swapping). Microaggregation was a special case where the identification of the records to be aggregated is based on ranks but the actual aggregation is performed on the values. Here we expected both record linkage methods to be equally effective. The results however indicate that rank-based record linkage outperforms distance-based record linkage for both multiplicative noise and microaggregation. This suggests an alternative explanation, namely that *rank-based record linkage performs better for masking methods where the level of masking is a function of the magnitude or rank of the actual original value*. The only masking methods in our experiment where the masking is independent of the actual value are independent and correlated noise procedures. Distance-based record linkage performs better only in this case. In all other methods, the masking is a function of either the magnitude (multiplicative noise), or the rank (rank swapping, p-distribution swapping), or both (microaggregation), of the actual original value. Rank-based record linkage performs better in all these cases. We believe that this is an interesting phenomenon that deserves further investigation.

As additional future work, we plan to extend our experimental work by comparing RBRL also against probabilistic record linkage, and increasing the number of masking methods tested. Particularly intriguing is the re-identification effectiveness of RBRL for other masking methods implicitly based on ranks, such as PRAM [8].

Acknowledgments and disclaimer. The second author is partly supported by the European Commission (projects H2020-644024 “CLARUS” and H2020-700540 “CANVAS”), by the Government of Catalonia (ICREA-Acadèmia prize and grant 2014 SGR 537) and by the Spanish Government (projects TIN2014-57364-C2-1-R “SmartGlacis” and TIN2015-70054-REDC). The second author leads the UNESCO Chair in Data Privacy, but the views expressed in this paper are the authors’ own and are not necessarily shared by UNESCO.

References

1. Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M. 2003. *Reference data sets to test and compare SDC methods for the protection of numerical microdata*. Deliverable of the EU IST-2000-25069 "CASC" project. http://neon.vb.cbs.nl/casc/
2. Domingo-Ferrer, J., Oganian, A., Torres, A., Mateo-Sanz, J. M. 2002. "On the security of microaggregation with individual ranking: analytical attacks." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5):477-491.
3. Domingo-Ferrer, J., Sánchez, D., and Rufian-Torrell, G. 2013. "Anonymization of nominal data using semantic marginality." *Information Sciences* 242: 35-48.
4. Domingo-Ferrer, J., and Torra, V. 2001. "A quantitative comparison of disclosure control methods for microdata." In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. North-Holland, pp. 111-134.

5. Domingo-Ferrer, J., and Torra, V. 2003. "Disclosure risk assessment in statistical disclosure control of microdata via advanced record linkage." *Statistics and Computing* 13(4):343-354.
6. Domingo-Ferrer, J., and Torra, V. 2005. "Ordinal, continuous and heterogeneous k-anonymity through microaggregation." *Data Mining and Knowledge Discovery* 11(2):195-212.
7. Fellegi, I., and Sunter, A. B. 1969. "A theory for record linkage." *Journal of the American Statistical Association* 64(328):1183-1210.
8. Gouweleeuw, J. M., Kooiman, P., De Wolf, P.-P. 1998. "Post randomisation for statistical disclosure control: theory and implementation." *Journal of Official Statistics* 14(4):463-478.
9. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and De Wolf, P.-P. 2012. *Statistical Disclosure Control*. Wiley.
10. Jaro, M. A. 1989. "Advances in record linkage methodology as applied to matching the 1985 census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414-420.
11. Mateo-Sanz, J. M., Sebé, F., and Domingo-Ferrer, J. 2004. "Outlier protection in continuous microdata masking." In *Privacy in Statistical Databases-PSD 2004*. LNCS 3050, Springer, pp. 201-2015.
12. Moore, R.A. 1996. *Controlled Data Swapping for Masking Public Use Microdata Sets.* Washington DC: Research report series (RR96/04), Statistical Research Division, US Census Bureau.
13. Nin, J., Herranz, J., and Torra, V. 2008. "Rethinking rank swapping to decrease disclosure risk." *Data & Knowledge Engineering* 64(1):346-364.
14. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., and Martínez, S. 2014. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity." *VLDB Journal* 23(5):771-794.
15. Torra, V., and Domingo-Ferrer, J. 2003. "Record linkage methods for multidatabase data mining." In *Information Fusion in Data Mining* (ed. V. Torra). Springer, pp. 99-130.
16. Winkler, W. E. 1995. "Matching and record linkage," In *Business Survey Methods*. Wiley, pp. 355-384.
17. Winkler, W. E. 2004. "Masking and re-identification methods for public-use microdata: overview and research problems." In *Privacy in Statistical Databases-PSD 2004*. LNCS 3050, Springer, pp. 231-246.