WP. 28
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (iv): Access to business microdata for analysis

EMPIRICAL DISCLOSURE RISK ASSESSMENT OF THE IPSO SYNTHETIC DATA GENERATORS

Invited Paper

Submitted by the Rovira I Virgili University of Tarragona and IIIA-CSIC, Spain[1]

[1] Prepared by Josep Domingo-Ferrer, Vicenç Torra, Josep M. Mateo-Sanz and Francesc Sebé.

Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators

Josep Domingo-Ferrer*, Vicenç Torra**, Josep M. Mateo-Sanz*, Francesc Sebé*

* Rovira i Virgili University of Tarragona, Dept. of Computer Engineering and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, ({josep.domingo,josepmaria.mateo,francesc.sebe}@urv.net)

** IIIA-CSIC, Campus UAB, E-08193 Bellaterra, Catalonia, (vtorra@iiia.csic.es)

Abstract. Information Preserving Statistical Obfuscation (IPSO) is a family of three methods IPSO-A, IPSO-B, IPSO-C for numerical synthetic data generation designed by Burridge in 2003. This paper reports on empirical work carried out to assess the re-identification risk of each method in different worst-case disclosure scenarios, with different datasets and using different record linkage methods. The conclusions of this study give some insight on how IPSO and other synthetic data generators can be tuned to minimize re-identification risk. Further, we discuss how a similar analysis could be conducted for synthetic generators of categorical data.

1 Introduction

Synthetic microdata generators usually care about preserving a model or some statistics, but they seldom pay attention to disclosure risk. The usual alibi is to argue that, since released microdata are synthetic, no real re-identification is possible. While this may be reasonable if synthetic generation is performed on the confidential outcome attributes, it is an unrealistic assumption if synthetic data generation is performed on the quasi-identifier attributes. In the latter case, re-identification can indeed happen if a snooper is able to link an external identified data source with some record in the released dataset using the quasi-identifier attributes: coming up with a correct pair (identifier, confidential attributes) is indeed a re-identification.

The disclosure model we work with is depicted in Figure 1. We assume that the released dataset on the right-hand side of the figure consists of confidential attributes X and non-confidential quasi-identifier attributes Y'; quasi-identifier attributes Y' have been masked using a partially synthetic data generation method. A snooper has obtained the external identified dataset on the left-hand side of the figure, which consists of one or several identifier attributes Id and several quasi-identifier attributes Y. Attributes Y are original and are not necessarily the same as attributes Y' in the released dataset. The snooper attempts to link records in the released dataset with records in the external identified dataset. Linkage is done by matching quasi-identifier attributes Y and Y'. The snooper's goal is to pair identifier values with confidential attribute values (*e.g.* to pair citizens' names with health conditions).

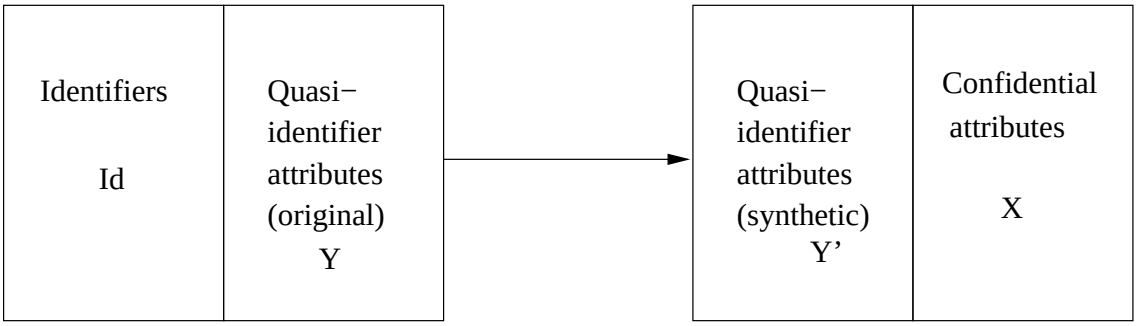

Figure 1: Re-identification scenario. Quasi-identifiers Y and Y' can have shared attributes or not

1.1 Contribution and plan of this paper

For the sake of concreteness, this paper focuses on a particular family of synthetic data generators, namely Information Preserving Statistical Obfuscation (IPSO, Burridge (2003)). IPSO is a family of three methods IPSO-A, IPSO-B, IPSO-C for numerical synthetic data generation which preserve, to a varying extent, a multivariate multiple regression model taking confidential attributes as independent variables and quasi-identifier attributes as dependent variables.

We have run IPSO-A, IPSO-B and IPSO-C on two different datasets and we report on the results of record linkage experiments on those datasets using different quasi-identifiers and different record linkage methods. In particular we consider the case where no quasi-identifier attributes are shared between the released dataset and the external identified source. The purpose of this study is to give some insight about re-identification which helps data protectors tune their synthetic data generators to make life more difficult for snoopers. We also discuss extensions of our study for synthetic generators of categorical data.

Section 2 briefly recalls IPSO-A, IPSO-B and IPSO-C. Section 3 describes the two datasets used. Record linkage methods employed in our analysis are explained in Section 4. Experimental results are given in Section 5. Conclusions and extensions are listed in Section 6.

2 The IPSO methods

Three variants of a procedure called Information Preserving Statistical Obfuscation (IPSO) are proposed in Burridge (2003). The basic form of IPSO will be called here IPSO-A. Informally, suppose two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. A multiple regression of Y on X is computed and fitted Y'_A attributes are computed. Finally, attributes X and Y'_A are released by IPSO-A in place of X and Y.

In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of

regression coefficients.

Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete dataset (y, x). If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). The second IPSO method, IPSO-B, modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved. This is done by the third IPSO method, IPSO-C.

3 The test datasets

We have used two reference datasets (Brand, et al. 2002) used in the European project CASC:

1. The "Census" dataset contains 1080 records with 13 numerical attributes labeled $v1$ to $v13$. This dataset was used in CASC and in several other works (Domingo-Ferrer, et al. 2001, Dandekar, et al. 2002, Yancey, et al. 2002, Laszlo & Mukherjee 2005, Domingo-Ferrer & Torra 2005, Domingo-Ferrer, et al. 2005).

2. The "EIA" dataset contains 4092 records with 15 attributes. The first five attributes are categorical and will not be used. We restrict to the last 10 numerical attributes, which will be labeled $v1$ to $v10$. This dataset was used in CASC, in Dandekar et al. (2002), Domingo-Ferrer et al. (2005) and partially in (Laszlo & Mukherjee 2005) (an undocumented subset of 1080 records from "EIA", called "Creta" dataset, was used in the latter paper).

4 Record linkage methods tried

The record linkage methods used fall into two paradigms:

- *Record linkage with shared attributes.* We assume that the external identified dataset **A** and the released dataset **B** share some attributes which are used for re-identification. Two methods corresponding to this approach have been tried:
  - Distance-based record linkage
  - Probabilistic record linkage
- *Record linkage without shared attributes.* No common attributes between the external identified dataset and the released dataset are assumed. A new correlation-based record linkage method has been designed and tried here.

We describe distance-based record linkage, probabilistic record linkage and correlation-based record linkage in the sections below. More details on distance-based and probabilistic record linkage can be found in Torra & Domingo-Ferrer (2003).

4.1 Distance-based record linkage

This approach, originally described in Tendick (1992) and Fuller (1993), consists of computing distances between records in **A** and **B**. Then, pairs of records at minimum distance are considered linked pairs. Of course, the distance between a pair of records must be computed based on shared attributes between those records, so that this approach does not work without shared attributes between the external data source and the released dataset.

Naturally, the application of this method depends on the existence of the distance function. Thus, a distance is assumed in each attribute V_i. We denote this distance by d_{V_i}. Assuming equal weight for all attributes, a record-level distance between records a and b can be constructed as:

$$d(a,b) = \sum_{i=1}^{n} d_{V_i}(V_i^A(a), V_i^B(b))$$

Depending on the data type of attributes, different within-attribute distances must be used. For numerical attributes, the Euclidean distance is a reasonable choice. See Domingo-Ferrer & Torra (2001) and Domingo-Ferrer & Torra (2002) on distances for categorical attributes. Whatever the distance and attribute type, one should use some kind of standardization to avoid scaling problems and give equal weight to attributes when combining them. For numerical data, one can

- Standardize each attribute before computing distances (this is done by subtracting the attribute mean and dividing by the attribute standard deviation). This type of distance-based record linkage will be called DRL1 in what follows.
- Compute distances on the unstandardized attributes and standardize distances by subtracting their average and dividing by their standard deviation. This approach will be called DRL2 in what follows.

4.2 Probabilistic record linkage

Probabilistic record linkage, called PRL in what follows, is described in Fellegi & Sunter (1969), Jaro (1989) and Winkler (1995). See the above mentioned references for details. Like distance-based record linkage, PRL assumes that the datasets to be linked share at least one quasi-identifier attribute.

The distinguishing features of PRL with respect to DRL1 and DRL2 are that: i) PRL can work on any data type (numerical or categorical) without any adaptation; ii) PRL does not require any assumptions on the relative weight of attributes (in particular, it requires no standardization). Its main drawback is its computational burden.

4.3 Correlation-based record linkage

This is a new proposal, called CRL in what follows, that we make for record linkage between numerical datasets without shared attributes. We assume that both datasets **A** and **B** have their own numerical quasi-identifier attributes. We also assume that both datasets consist of n records corresponding to the same set of individual respondents.

The method finds the pair (i, j) of quasi-identifier attributes in **A** and **B** with highest correlation. Then **A** is sorted by its i-th quasi-identifier attribute and **B** is sorted by its j-th quasi-identifier attribute. If there remain subsets of records with equal rank in either dataset, find the pair of attributes with the second highest correlation and use them to decide the ordering within those subsets of records. This process can be iterated until no two records in either dataset have the same rank or we have used all quasi-identifier attributes; in the latter case, use a random ordering for any remaining records with equal rank. At the end of this process, all n records in **A** and **B** are ranked. The final step is to link the k-th record in **A** with the k-th record in **B**, for $k = 1$ to n.

5 Experimental results

We implemented IPSO-A, IPSO-B and IPSO-C above for generation of partially synthetic data. We then applied them to the "Census" and "EIA" datasets to obtain several versions of partially synthetic data. Next, we considered re-identication scenarios with shared and non-shared attributes and tried distance-based, probabilistic and correlation-based record linkage on them. This section describes in detail this experimental work and the results that were obtained.

5.1 Results on "Census"

We took the "Census" dataset and used the correlations between its 13 attributes to compute a dendrogram. We followed the dendrogram rather than the semantics of attributes in "Census" to select quasi-identifier attributes and confidential attributes. The rationale of this is that we were looking for worst-case scenarios to test the safety of the synthetic generators IPSO-A, IPSO-B and IPSO-C: the worst case (most likely to yield correct re-identifications) happens when the snooper uses quasi-identifier attributes which are highly correlated to the remaining attributes in the dataset. Thus, we chose quasi-identifier attributes with central positions in the dendrogram; this strategy led us to two different choices of confidential outcome attributes X and quasi-identifier attributes Y which gave two different scenarios $S1$ and $S2$. Table 1 summarizes the attributes in each dataset for each scenario.

We then took the quasi-identifier attributes in datasets **B** in Table 1 and used methods IPSO-A, IPSO-B and IPSO-C on them. In other words, we fitted a multivariate multiple regression model to them by taking as independent attributes the confidential attributes X and as dependent attributes the quasi-identifier attributes Y.

We first explain the notation used in the tables of results in this section:

Table 1: Splittings of "Census" into datasets **A** and **B** and attributes per dataset. In individual experiments, several subsets of quasi-identifier attributes Y were considered

Scenario	Data set	Shared attributes		Non-shared attributes	
		Quasi-id. Y	Conf. attr. X	Quasi-id. Y	Conf. attr. X
S1	**A**	$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$		$v3, v4, v9, v12$	
	B	$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v2, v5, v8, v10$	$v1, v6, v7$ $v11, v13$	$v2, v5, v8, v10$
S2	**A**	$v4, v7, v12, v13$		$v4, v12$	
	B	$v4, v7, v12, v13$	$v1, v2, v3, v5, v6$ $v8, v9, v10, v11$	$v7, v13$	$v1, v2, v3, v5, v6$ $v8, v9, v10, v11$

- A, B, C as a subscript denote that the attribute was generated using IPSO-A, IPSO-B or IPSO-C, respectively; no subscript means that the attribute is original.
- $S1$ as a superscript means that this attribute was obtained by fitting a multivariate multiple regression model taking as independent attributes four confidential attributes X (specifically, $v2, v5, v8, v10$, see scenario $S1$ in Table 1).
- $S2$ as a superscript means that this attribute was obtained by fitting a multivariate multiple regression model taking as independent attributes nine confidential attributes X (specifically, $v1, v2, v3, v5, v6, v8, v9, v10, v11$, see scenario $S2$ in Table 1).

Table 2 shows the results of record linkage experiments between the "Census" dataset and a partially synthetic version of it generated using IPSO-A. The table shows only the quasi-identifiers used in each experiment, which are subsets of those specified in Table 1.

Quasi-identifiers in Table 2 were selected using the cross-correlation matrix between the original quasi-identifier attributes and the quasi-identifier attributes generated using method IPSO-A. The rationale of our quasi-identifier choices is that at least some of the quasi-identifiers in datasets **A** and **B** should be highly correlated. Note that this strategy in quasi-identifier selection can be followed by a real snooper, since he can compute the cross-correlation matrix between the external identified dataset and the released, partially synthetic datasets.

The results for IPSO-B were very similar to those for IPSO-A, and will not be reported here for the sake of brevity. The results for IPSO-C are different and are shown in Table 3.

It can be observed that, for the same quasi-identifier attributes, method IPSO-C results in less re-identifications than methods IPSO-A and IPSO-B. Since, IPSO-C preserves more statistics than the other two methods, it is clearly the best choice.

Table 2: Re-identification experiments using dataset "Census" and method IPSO-A. Results in number of correct re-identifications over an overall number of 1080 records. Percentage of correct re-identifications between parentheses. DRL1: attribute-standardizing implementation of distance-based record linkage (DRL); DRL2: distance-standardizing implementation of DRL; PRL: probabilistic record linkage; CRL: correlation-based record linkage

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v7, v12$	$v7_A^{S1}, v12_A^{S1}$	144 (13.3%)	144 (13.3%)	144 (13.3%)	7 (0.6%)
$v4, v7, v11, v12$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}$	85 (7.8%)	82 (7.5%)	68 (6.2%)	7 (0.6%)
$v4, v7, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v12_A^{S1}, v13_A^{S1}$	104 (9.6%)	106 (9.8%)	116 (10.7%)	7 (0.6%)
$v4, v7, v11, v12, v13$	$v4_A^{S1}, v7_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$	79 (7.3%)	80 (7.4%)	85 (7.8%)	7 (0.6%)
$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v1_A^{S1}, v3_A^{S1}, v4_A^{S1}, v6_A^{S1}, v7_A^{S1}$ $v9_A^{S1}, v11_A^{S1}, v12_A^{S1}, v13_A^{S1}$	36 (3.3%)	31 (2.8%)	82 (7.2%)	7 (0.6%)
$v7, v12$	$v7_A^{S2}, v12_A^{S2}$	79 (7.3%)	79 (7.3%)	79 (7.3%)	40 (3.7%)
$v4, v13$	$v4_A^{S2}, v13_A^{S2}$	50 (4.6%)	50 (4.6%)	50 (4.6%)	5 (0.4%)
$v7, v12, v13$	$v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$	82 (7.5%)	81 (7.5%)	85 (7.8%)	40 (3.7%)
$v4, v7, v12, v13$	$v4_A^{S2}, v7_A^{S2}, v12_A^{S2}, v13_A^{S2}$	85 (7.8%)	86 (7.9%)	93 (8.6%)	40 (3.7%)
$v4$	$v7_A^{S1}$	N/A	N/A	N/A	7 (0.6%)
$v7$	$v4_A^{S1}$	N/A	N/A	N/A	4 (0.3%)
$v4, v12$	$v7_A^{S1}, v13_A^{S1}$	N/A	N/A	N/A	37 (3.4%)
$v3, v4, v9, v12$	$v1_A^{S1}, v6_A^{S1}, v7_A^{S1}, v11_A^{S1}, v13_A^{S1}$	N/A	N/A	N/A	37 (3.4%)
$v1, v6, v7, v11, v13$	$v3_A^{S1}, v4_A^{S1}, v9_A^{S1}, v12_A^{S1}$	N/A	N/A	N/A	4 (0.3%)
$v4, v12$	$v7_A^{S2}, v13_A^{S2}$	N/A	N/A	N/A	43 (3.9%)
$v7, v13$	$v4_A^{S2}, v12_A^{S2}$	N/A	N/A	N/A	8 (0.7%)

5.2 Results on "EIA"

We took the "EIA" dataset and computed a correlation-based dendrogram of its 10 numerical attributes $v1, \cdots, v10$. Like for "Census", we used the "EIA" dendrogram rather than the semantics of "EIA" attributes to select quasi-identifier attributes and confidential attributes. A single scenario (choice of confidential attributes X) was defined. Table 4 summarizes the quasi-identifiers considered in each dataset for the paradigms with shared and non-shared attributes.

We then took the quasi-identifier attributes in dataset **B** in Table 4 and used methods IPSO-A, IPSO-B, IPSO-C on them. In other words, we fitted a multivariate multiple regression model to **B** by taking as independent attributes the confidential attributes X and as dependent attributes the quasi-identifier attributes Y. The notation in Table 5 below is the same used in the analogous tables for the "Census" dataset, except that no scenario superscript is used. The table shows the results of record linkage experiments between the "Census" dataset and partially synthetic versions of it generated using IPSO-A, IPSO-B and IPSO-C. Only the quasi-identifiers used in each experiment are listed, which are subsets of those specified in Table 4.

Quasi-identifiers in Table 5 were selected using the cross-correlation matrix between the original quasi-identifier attributes and the quasi-identifier attributes gener-

Table 3: Re-identification experiments using dataset "Census" and method IPSO-C. Results in number of correct re-identifications over an overall number of 1080 records.

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v7, v12$	$v7_C^{S1}, v12_C^{S1}$	32 (2.9%)	32 (2.9%)	32 (2.9%)	13 (1.2%)
$v4, v7, v11, v12$	$v4_C^{S1}, v7_C^{S1}, v11_C^{S1}, v12_C^{S1}$	39 (3.6%)	39 (3.6%)	36 (3.3%)	13 (1.2%)
$v4, v7, v12, v13$	$v4_C^{S1}, v7_C^{S1}, v12_C^{S1}, v13_C^{S1}$	35 (3.2%)	35 (3.2%)	33 (3.0%)	13 (1.2%)
$v4, v7, v11, v12, v13$	$v4_C^{S1}, v7_C^{S1}, v11_C^{S1}, v12_C^{S1}, v13_C^{S1}$	40 (3.7%)	40 (3.7%)	43 (3.9%)	13 (1.2%)
$v1, v3, v4, v6, v7$ $v9, v11, v12, v13$	$v1_C^{S1}, v3_C^{S1}, v4_C^{S1}, v6_C^{S1}, v7_C^{S1}$ $v9_C^{S1}, v11_C^{S1}, v12_C^{S1}, v13_C^{S1}$	19 (1.7%)	19 (1.7%)	50 (4.6%)	13 (1.2%)
$v7, v12$	$v7_C^{S2}, v12_C^{S2}$	42 (3.9%)	42 (3.9%)	42 (3.9%)	12 (1.1%)
$v4, v13$	$v4_C^{S2}, v13_C^{S2}$	17 (1.6%)	17 (1.5%)	17 (1.5%)	6 (0.5%)
$v7, v12, v13$	$v7_C^{S2}, v12_C^{S2}, v13_C^{S2}$	31 (2.8%)	31 (2.8%)	36 (3.3%)	12 (1.1%)
$v4, v7, v12, v13$	$v4_C^{S2}, v7_C^{S2}, v12_C^{S2}, v13_C^{S2}$	26 (2.4%)	26 (2.4%)	33 (3.0%)	12 (1.1%)
$v4$	$v7_C^{S1}$	N/A	N/A	N/A	10 (0.9%)
$v7$	$v4_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v4, v12$	$v7_C^{S1}, v13_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v3, v4, v9, v12$	$v1_C^{S1}, v6_C^{S1}, v7_C^{S1}, v11_C^{S1}, v13_C^{S1}$	N/A	N/A	N/A	3 (0.3%)
$v1, v6, v7, v11, v13$	$v3_C^{S1}, v4_C^{S1}, v9_C^{S1}, v12_C^{S1}$	N/A	N/A	N/A	18 (1.7%)
$v4, v12$	$v7_C^{S2}, v13_C^{S2}$	N/A	N/A	N/A	6 (0.5%)
$v7, v13$	$v4_C^{S2}, v12_C^{S2}$	N/A	N/A	N/A	10 (0.9%)

Table 4: Splittings of "EIA" into datasets **A** and **B** and attributes per dataset

Data set	Shared attributes		Non-shared attributes	
	Quasi-id. Y	Conf. attr. X	Quasi-id. Y	Conf. attr. X
A	$v1, v2, v7, v8, v9$		$v1, v7$	
B	$v1, v2, v7, v8, v9$	$v3, v4, v5, v6, v10$	$v2, v8, v9$	$v3, v4, v5, v6, v10$

ated using methods IPSO-A, IPSO-B, IPSO-C. The rationale of our quasi-identifier choices is that at least some of the quasi-identifiers in datasets **A** and **B** should be highly correlated. Note that this strategy in quasi-identifier selection can be followed by a real snooper, since he can compute the cross-correlation matrix between the external identified dataset and the released, partially synthetic datasets.

6 Conclusions and extensions

It can be seen that, among the methods tried, IPSO-C is the safest one, in that it is the one allowing less re-identifications. Apparently, this is perfect, because IPSO-C also preserves more regression statistics that IPSO-A and IPSO-B. However, at a closer look, it can be seen that the individual values generated by IPSO-C for the quasi-identifier attributes are more different from the original values than in the case of IPSO-A and IPSO-B. This can easily be seen by computing the average

Table 5: Re-identification experiments using dataset "EIA" and methods IPSO-A, IPSO-B and IPSO-C. Results in number of correct re-identifications over an overall number of 4092 records.

Quasi-identifier in external **A**	Quasi-identifier in released **B**	DRL1	DRL2	PRL	CRL
$v1$	$v1_A$	10 (0.2%)	10 (0.2%)	10 (0.2%)	32 (0.8%)
$v1, v7, v8$	$v1_A, v7_A, v8_A$	23 (0.5%)	24 (0.5%)	11 (0.2%)	30 (0.7%)
$v1, v2, v7, v8, v9$	$v1_A, v2_A, v7_A, v8_A, v9_A$	186 (4.5%)	171 (4.1%)	189 (4.6%)	46 (1.1%)
$v1$	$v9_A$	N/A	N/A	N/A	9 (0.2%)
$v1, v7$	$v2_A, v8_A, v9_A$	N/A	N/A	N/A	7 (0.2%)
$v2, v8, v9$	$v1_A, v7_A$	N/A	N/A	N/A	6 (0.1%)
$v1$	$v1_B$	10 (0.2%)	10 (0.2%)	10 (0.2%)	26 (0.6%)
$v1, v7, v8$	$v1_B, v7_B, v8_B$	23 (0.6%)	24 (0.5%)	11 (0.2%)	25 (0.6%)
$v1, v2, v7, v8, v9$	$v1_B, v2_B, v7_B, v8_B, v9_B$	187 (4.6%)	171 (4.1%)	189 (4.6%)	47 (1.1%)
$v1$	$v9_B$	N/A	N/A	N/A	9 (0.2%)
$v1, v7$	$v2_B, v8_B, v9_B$	N/A	N/A	N/A	10 (0.2%)
$v2, v8, v9$	$v1_B, v7_B$	N/A	N/A	N/A	8 (0.2%)
$v1$	$v1_C$	7 (0.2%)	7 (0.2%)	7 (0.2%)	8 (0.2%)
$v1, v7, v8$	$v1_C, v7_C, v8_C$	10 (0.2%)	10 (0.2%)	6 (0.1%)	9 (0.2%)
$v1, v2, v7, v8, v9$	$v1_C, v2_C, v7_C, v8_C, v9_C$	42 (1.0%)	42 (1.0%)	71 (1.7%)	28 (0.7%)
$v1$	$v9_C$	N/A	N/A	N/A	7 (0.2%)
$v1, v7$	$v2_C, v8_C, v9_B$	N/A	N/A	N/A	6 (0.1%)
$v2, v8, v9$	$v1_C, v7_C$	N/A	N/A	N/A	5 (0.1%)

Euclidean distance between original records and records generated by the three IPSO methods; the largest average distance is between original and IPSO-C records. The explanation of the above is that, in order to preserve more statistics, IPSO-C resorts to "injecting" more perturbation at the record level than IPSO-A and IPSO-B.

We now examine the influence of the number of independent confidential attributes X. In Scenario S1 ("Census" dataset, Table 1), the multivariate multiple regression model uses only four confidential attributes X as independent variables. In Scenario S2, nine confidential attributes X are used. In fact, the X in Scenario S1 are a subset of the X in Scenario S2. Thus, the synthetic quasi-identifier attributes Y in Scenario S1 are generated based on less X attributes than in Scenario S2. Surprising enough, the differences between both scenarios as to the number of re-identifications are less straightforward than one would expect (see Tables 2 and 3). By focusing on identical quasi-identifiers across both scenarios S1 and S2 (that is, $(v7, v12)$ and $(v4, v7, v12, v13)$) we can see that, for IPSO-A and IPSO-B, distance-based and probabilistic record linkage re-identify more when the regression model has been fitted on few independent attributes. For those two methods, correlation-based record linkage works better when the regression model has been fitted on a greater number of independent attributes. IPSO-C displays exactly the opposite behavior: more DRL1, DRL2 and PRL re-identifications and less CRL

re-identifications are obtained when there are more independent attributes.

Another important point to be analyzed is the influence of the quasi-identifier length. A longer quasi-identifier does not necessarily result in more re-identifications. Indeed, it can be seen in Table 2 than more re-identifications are obtained with $(v7, v12)$ than with longer quasi-identifiers also including $v7$ and $v12$. The reason is that, as it can be checked in the cross-correlation matrix between the original quasi-identifier and the quasi-identifier generated by IPSO-A, it turns out that $v7$ and $v12$ are good representatives of the other quasi-identifier attributes: $v7$ is highly correlated with $v4_A$ (0.9778), $v6_A$ (0.9807) and $v7_A$ (0.9812); $v12$ is highly correlated with $v3_A$ (0.9509), $v11_A$ (0.9788), $v12_A$ (0.9793) and $v13_A$ (0.9792). Thus $v7$ and $v12$ complement each other in sort of "covering" nearly all quasi-identifier attributes generated by IPSO-A (only $v1_A$ and $v9_A$ stay "uncovered"). This is no surprise, given the central position that $v7$ and $v12$ hold in the dendrogram of the "Census" dataset. Thus, the lessons learned are:

1. If a snooper can find via cross-correlation matrix a few quasi-identifier attributes that are highly correlated to the all partially synthetic quasi-identifier attributes, she should use only those few attributes for re-identification; using longer quasi-identifiers will only add noise and reduce the number of successful re-identifications.

2. *The data protector should generate partially synthetic microdata in such a way that no such small set of original quasi-identifier attributes are highly correlated to all synthetic quasi-identifier attributes.* In doing so, the data protector will force potential snoopers to use longer quasi-identifiers, which makes life more difficult for them (more external identified information required).

We can also compare the performance of the record linkage methods used. It seems that the overall performance of DRL1, DRL2 and PRL in terms of the number of re-identifications is similar. Nonetheless, while both distance-based methods DRL1 and DRL2 stay similar for any quasi-identifier length, probabilistic record linkage PRL seems to clearly outperform DRL1 and DRL2 for longer quasi-identifiers. Correlation-based record linkage (CRL) behaves clearly worse than PRL, DRL1 and DRL2 and should not be used in the shared-attributes paradigm. However, it is the only method among those considered that is still applicable without shared attributes.

Finally, a few words on the influence of the dataset size. We used two datasets with differents sizes ("Census", 1080 records; "EIA", 4092 records) to attempt an assessment of the influence of the dataset size on the number of re-identifications. By comparing Table 5 with Tables 2 and 3, we see that the percentage of re-identifications is lower for the larger "EIA" dataset, as one would expect. However, the *absolute number of re-identifications* is not lower in "EIA" when a sufficiently long quasi-identifier is used. In fact for quasi-identifier $(v1, v2, v7, v8, v9)$ and shared attributes, we obtain between 170 and 190 re-identifications for IPSO-A and IPSO-B, and between 40 and 70 for IPSO-C, which is more than the number of re-identifications we obtained when using the "Census" dataset.

Only numerical attributes have been considered in this work. To deal with categorical quasi-identifier attributes one would need:

- To use methods which, unlike IPSO-A, IPSO-B and IPSO-C, are appropriate for generation of categorical synthetic microdata.
- To use distance-based record linkage with ordinal or nominal distances rather than the Euclidean distance.
- To use Spearman's rank correlations instead of Pearson's correlations to adapt correlation-based record linkage to ordinal attributes (for nominal attributes there is no obvious adaptation).

Probabilistic record linkage is the only record linkage method among those used that can directly work on categorical data without any adaptation.

Acknowledgments

This work was partly funded by Cornell University under contracts no. 47632-10042 and 47632-10043, by the Catalan government under project 2002 SGR 00170 and by the Spanish government under project SEG2004-04352-C04-01/03 "PROPRIETAS".

References

R. Brand, et al. (2002). 'Reference data sets to test and compare SDC methods for protection of numerical microdata'. European Project IST-2000-25069 CASC, http://neon.vb.cbs.nl/casc.

J. Burridge (2003). 'Information preserving statistical obfuscation'. *Statistics and Computing* **13**:321–327.

R. Dandekar, et al. (2002). 'LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection'. In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, vol. 2316 of *LNCS*, pp. 153–162, Berlin Heidelberg. Springer.

J. Domingo-Ferrer, et al. (2001). 'Comparing SDC methods for microdata on the basis of information loss and disclosure risk'. In *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pp. 807–826, Luxemburg. Eurostat.

J. Domingo-Ferrer, et al. (2005). 'A polynomial-time approximation to optimal multivariate microaggregation'. *submitted manuscript* .

J. Domingo-Ferrer & V. Torra (2001). 'A quantitative comparison of disclosure control methods for microdata'. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, & L. Zayatz (eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111–134, Amsterdam. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

J. Domingo-Ferrer & V. Torra (2002). 'Validating distance-based record linkage with probabilistic record linkage'. In F. T. M. T. Escrig & E. Golobardes (eds.), *Topics in Artificial Intelligence*, vol. 2504 of *LNCS*, pp. 207–215, Berlin Heidelberg. Springer.

J. Domingo-Ferrer & V. Torra (2005). 'Ordinal, continuous and heterogenerous k-anonymity through microaggregation'. *Data Mining and Knowledge Discovery* **11**(2). (to appear).

I. P. Fellegi & A. B. Sunter (1969). 'A theory for record linkage'. *Journal of the American Statistical Association* **64**(328):1183–1210.

W. A. Fuller (1993). 'Masking procedures for microdata disclosure limitation'. *Journal of Official Statistics* **9**:383–406.

M. A. Jaro (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.

M. Laszlo & S. Mukherjee (2005). 'Minimum spanning tree partitioning algorithm for microaggregation'. *IEEE Transactions on Knowledge and Data Engineering* **17**(7):902–911.

P. Tendick (1992). 'Assessing the effectiveness of the noise addition method of preserving confidentiality in the multivariate normal case'. *Journal of Statistical Planning and Inference* **31**:273–282.

V. Torra & J. Domingo-Ferrer (2003). 'Record linkage methods for multidatabase data mining'. In V. Torra (ed.), *Information Fusion in Data Mining*, pp. 101–132, Germany. Springer.

W. E. Winkler (1995). 'Advanced methods for record linkage'. In *Proc. of the American Statistical Association Section on Survey Research Methods*, pp. 467–472. ASA.

W. E. Yancey, et al. (2002). 'Disclosure risk assessment in perturbative microdata protection'. In J. Domingo-Ferrer (ed.), *Inference Control in Statistical Databases*, vol. 2316 of *LNCS*, pp. 135–152, Berlin Heidelberg. Springer.