UNECE/Eurostat Work Session on SDC

Empirical Disclosure Risk Assessment of the IPSO Synthetic Data Generators

Josep Domingo-Ferrer[1], Vicenç Torra[2], Josep M. Mateo-Sanz[1]
Francesc Sebé[1]

November, 2005

[1] Universitat Rovira i Virgili, Tarragona; [2] Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), Bellaterra

Introduction

- Information Preserving Statistical Obfuscation (IPSO): family of methods (IPSO-A, IPSO-B, IPSO-C) for numerical synthetic data generation

- This paper reports on empirical work carried out to assess the re-identification risk of each method (worst-case scenarios)

Outline

1. Introduction
2. The IPSO model
3. The Test Dataset
4. Record linkage methods tried
5. Experimental results
6. Conclusions

Introduction

Introduction

- Synthetic microdata generators usually care about preserving a model or some statistics
  $\longrightarrow$ but seldom pay attention to disclosure risk:

- Usual alibi:

  "Since released microdata are synthetic, no real re-identification is possible".

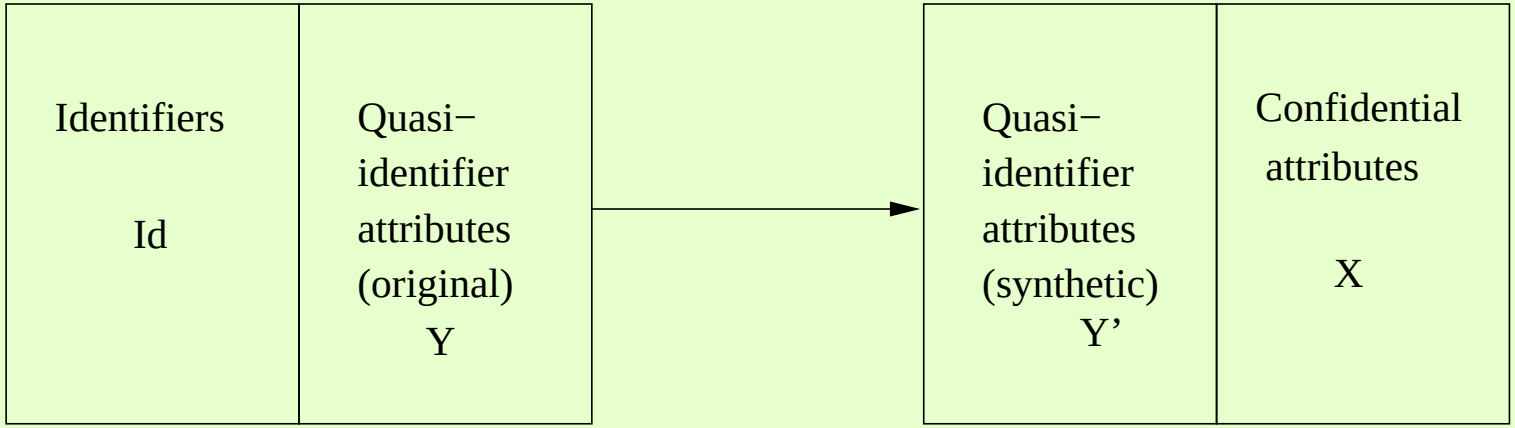
Introduction

- May be reasonable, if synthetic generation is performed on the confidential outcome attributes.

- However, 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.

Introduction: Disclosure model



Introduction

- Focus on IPSO family (IPSO-A, IPSO-B, IPSO-C) for numerical synthetic data generation.
- Run IPSO-A, IPSO-B, IPSO-C on two different datasets.
- Report on the results of record linkage experiments (using different quasi-identifiers and different record linkage methods).

$\longrightarrow$ to give some insight about re-identification which helps data protectors tune their synthetic data generators to make life more difficult for snoopers.

The IPSO method

The IPSO method (informally)

IPSO-A:

- X and Y two sets of attributes
- X: confidential outcome attributes
- Y: 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.

The IPSO method (informally)

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).

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 $\hat{B}_B = \hat{B}$.

IPSO-C: 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.

The test datasets

The test datasets

- Two reference datasets used in the European project CASC:
  - "Census" dataset:
    - 1080 records with 13 numerical attributes (labeled $v1$ to $v13$).
    - Used in CASC and in several other works.
  - "EIA" dataset:
    - 4092 records with 15 attributes.
    - Experiments restricted to the last 10 numerical attributes (labeled $v1$ to $v10$), the other five attributes are categorical.
    - Used in CASC, and in a few other works.

Record linkage methods tried

Record linkage methods tried

- The record linkage methods used fall into two paradigms:

  Record linkage with shared attributes: The external identified dataset $\mathbf{A}$ and the released dataset $\mathbf{B}$ share some attributes which are used for re-identification. Methods 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. Method tried:
  - A new correlation-based record linkage method

Record linkage methods tried

- Distance-based record linkage:
  - Compute distances between records in $\mathbf{A}$ and $\mathbf{B}$.
  - Pairs of records at minimum distance are considered linked pairs.

- Considerations:
  - Distance must be computed based on shared attributes.
  - It depends on the existence of the distance function.
    - A distance is assumed in each attribute V_i: d_{V_i}.
    - Then, assuming equal weight for all attributes, distance between records a and b:

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

Record linkage methods tried

- Considerations (II):
  - Depending on the data type of attributes, different within-attribute distances must be used.
    - *E.g.*, for numerical attributes: Euclidean
  - To avoid scaling problems (equal weight): standardization
    - For numerical data, one can:
      - Standardize each attribute before computing distances (subtracting the attribute mean and dividing by the attribute standard deviation). [DRL1]
      - Compute distances on the unstandardized attributes and standardize distances (subtracting their average and dividing by their standard deviation). [DRL2]

Record linkage methods tried

- Probabilistic record linkage [PRL]
  - PRL assumes that the datasets to be linked share at least one quasi-identifier attribute.
  - Distinguishing features of PRL with respect to DRL1 and DRL2:
    - PRL can work on any data type (numerical or categorical) without any adaptation
    - PRL does not require any assumptions on the relative weight of attributes (no standardization).
  - Its main drawback is its computational burden.

Record linkage methods tried

- Correlation-based record linkage [CRL]
  - A new proposal, that we make for record linkage between numerical datasets without shared attributes.
  - Both datasets $\mathbf{A}$ and $\mathbf{B}$ 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.

Record linkage methods tried

- Correlation-based record linkage: Method
  - Find the pair (i, j) of quasi-identifier attributes in $\mathbf{A}$ and $\mathbf{B}$ with highest correlation.
  - Sort $\mathbf{A}$ by its i-th quasi-identifier attribute, sort $\mathbf{B}$ 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.

    $\rightarrow$ All n records in $\mathbf{A}$ and $\mathbf{B}$ are ranked.
  - Link the k-th record in $\mathbf{A}$ with the k-th record in $\mathbf{B}$, for $k = 1$ to n.

Experimental results

Experimental results

- Experiments:
  - Implemented IPSO-A, IPSO-B and IPSO-C above for generation of partially synthetic data.
  - Applied them to the "Census" and "EIA" datasets to obtain several versions of partially synthetic data.
  - Considered re-identication scenarios with shared and non-shared attributes and tried distance-based, probabilistic and correlation-based record linkage on them.

- We describe this experimental work and the results that were obtained.

"Census": Synthetic data generation

- Looking for worst-case scenarios .... safety of synthetic generators??

  (worst-case: most likely to yield correct re-identifications)

  – Worst case: the snooper uses quasi-identifier attributes which are highly correlated to the remaining attributes in the dataset.

"Census": Synthetic data generation

- Looking for worst-case scenarios .... safety of synthetic generators??

  (worst-case: most likely to yield correct re-identifications)

  - Worst case: the snooper uses quasi-identifier attributes which are highly correlated to the remaining attributes in the dataset.

- Correlations:

  - Use correlations between its 13 attributes to compute a dendrogram.
  - Follow the dendrogram to select quasi-identifier attributes and confidential attributes.

    $\longrightarrow$ 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 $S2$.

"Census": Synthetic data generation

Scen-ario	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	$v4, v7$ $v12, v13$	$v1, v2, v3$ $v5, v6$ $v8, v9, v10, v11$	$v7, v13$	$v1, v2, v3$ $v5, v6$ $v8, v9, v10, v11$

"Census": Synthetic data generation

We took the quasi-identifier attributes in datasets $\mathbf{B}$ in the table and used methods IPSO-A, IPSO-B and IPSO-C on them.

$\longrightarrow$ 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.

"Census": Results (notation)**

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

- 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 $S1$).
- $S2$ 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 $S2$).

"Census": Results

Re-identification experiments: "Census" / IPSO-A[1]

Quasi-identifier in external $\mathbf{A}$	Quasi-identifier in released $\mathbf{B}$	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%)

[1] $A, B, C \rightarrow$ IPSO-A/B/C; $S1, S2$ multivariate multiple regression (4 / 9 attr.)

"Census": Results

Re-identification experiments: "Census" / IPSO-C

Quasi-identifier in external $\mathbf{A}$	Quasi-identifier in released $\mathbf{B}$	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%)

"Census": Results (the table)**

- The table shows the quasi-identifiers used in each experiment, which are subsets of those specified in the Table on scenarios.

- Quasi-identifiers 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 $\mathbf{A}$ and $\mathbf{B}$ 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.

"Census": Results

- Comments:
  - 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.

"EIA": Synthetic data generation

- Looking for worst-case scenarios...
  - Correlation-based dendrogram (10 numerical attributes $v1, \cdots, v10$).
  - Dendrogram to select quasi-identifier attributes and confidential attributes.
  - A single scenario (choice of confidential attributes X) was defined:

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	$v1, v2, v7, v8, v9$	$v3, v4, v5, v6, v10$	$v2, v8, v9$	$v3, v4, v5, v6, v10$

"EIA": Results

Re-identification experiments: "EIA" / IPSO-A, B and C

Quasi-identifier in external $\mathbf{A}$	Quasi-identifier in released $\mathbf{B}$	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%)

Conclusions

Conclusions

- IPSO-C is the safest method (less re-identifications).
  - IPSO-C preserves more regression statistics (vs. IPSO-A, IPSO-B).
  - However, at a closer look ...
    - The individual values generated by IPSO-C for the quasi-identifier attributes are more different from the original values

      (computing average Euclidean distance between original records and records generated by the three IPSO methods $\rightarrow$ largest average distance: IPSO-C).
    - In order to preserve more statistics, IPSO-C resorts to "injecting" more perturbation at the record level than IPSO-A and IPSO-B.

Conclusions

- Number of independent confidential attributes X.
  - S1 ("Census"), the multivariate multiple regression model uses only four confidential attributes X as independent variables.
  - S2, nine confidential attributes X are used.

  $\rightarrow$ The synthetic quasi-identifier attributes Y in Scenario S1 are generated based on less X attributes than in Scenario S2.

- By focusing on identical quasi-identifiers across both scenarios S1 and S2 (that is, $(v7, v12)$ and $(v4, v7, v12, v13)$):
  - For IPSO-A and IPSO-B: DRL1, DRL2 and PRL re-identify more when the regression model has been fitted on few independent attributes. CRL works better when the regression model has been fitted on a greater number of independent attributes.
  - For IPSO-C: exactly the opposite behavior.

Conclusions

- Influence of the quasi-identifier length:
  - Longer quasi-identifier not always better.

- Results:
  - IPSO-A: $(v7, v12)$ best subset for re-identifications.
    $\rightarrow v7, v12$ good representatives of the other quasi-identifier attrs.:
    $v7$ 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).
    $\rightarrow v7, v12$ 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.

Conclusions

- Influence of the quasi-identifier length:
  - Lessons learned:
    1. If a snooper can find via cross-correlation matrix a few quasi-identifier attributes that are highly correlated to 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).

Conclusions

- Performance:
  - Similar overall performance of DRL1, DRL2 and PRL in terms of the number of re-identifications
  - 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.

Conclusions

- On the influence of the dataset size:
  - Different size: "Census", 1080 records; "EIA", 4092 records
  - Percentage of re-identifications lower for "EIA" dataset
  - However, the *absolute number of re-identifications* is not lower in "EIA" when a sufficiently long quasi-identifier is used.
    - E.g., EIA, for quasi-identifier $(v1, v2, v7, v8, v9)$ and shared attributes: 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.

Conclusions

- Only numerical attributes have been considered

- To deal with categorical attributes:
  - 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.