Mitigating the Curse of Dimensionality in Data Anonymization

Jordi Soria-Comas and Josep Domingo-Ferrer

Universitat Rovira i Virgili
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
{jordi.soria, josep.domingo}@urv.cat

Abstract. In general, just suppressing identifiers from released microdata is insufficient for privacy protection. It has been shown that the risk of re-identification increases with the dimensionality of the released records. Hence, sound anonymization procedures are needed to anonymize high-dimensional records. Unfortunately, most privacy models yield very poor utility if enforced on data sets with many attributes. In this paper, we propose a method based on principal component analysis (PCA) to mitigate the curse of dimensionality in anonymization. Our aim is to reduce dimensionality without incurring large utility losses. We instantiate our approach with anonymization based on differential privacy. Empirical work shows that using differential privacy on the PCA-transformed and dimensionality-reduced data set yields less information loss than directly using differential privacy on the original data set.

Keywords: Privacy preserving data publishing, curse of dimensionality, differential privacy

1 Introduction

Under the EU General Data Protection Regulation (GDPR), which is becoming a *de facto* global standard, personally identifiable information (PII) cannot be accumulated for secondary use. Hence, to stay GDPR-compliant, exploratory analytics must take place on anonymized data.

As repeatedly aired in the media [15, 1, 19], just suppressing direct identifiers (names, passport numbers, etc.), let alone replacing them by pseudonyms, is not enough to anonymize a data set. In the case of high-dimensional data even adding a small amount of noise may not be enough to prevent re-identification [11].

In this work we deal with anonymization of microdata with a moderate to large number of attributes. This type of data are more and more common in the big data landscape, in which data sets are constructed by merging several sources and may end up having a great number of attributes.

Direct application of statistical disclosure control (SDC) methods [8] to high-dimensional data is likely to result in high information loss and hence in poor utility, due to the sparsity of such data.

In this work we tackle the problem by investigating a pre-processing transformation aimed at reducing dimensionality before anonymization is carried out. Specifically, we propose principal component analysis (PCA), which could also be replaced by another dimensionality reduction technique, e.g. if data are not numerical. Subjects send their original records to the controller, who gathers the entire original data set and can compute the principal components of it. The dimensionality reduction comes from the controller choosing only those components that carry most information. Anonymization is carried out on this smaller set of components and the result if mapped back to the domain of the original attributes to obtain the anonymized data set.

For the sake of concreteness, we exemplify dimensionality reduction to obtain differentially private data sets. According to our experiments, *using differential privacy on the PCA-transformed and dimensionality-reduced data incurs substially less information loss than directly applying differential privacy on the original data.*

The rest of this paper is organized as follows. In Section 2, we recall some background concepts. In Section 3 we deal with dimensionality reduction. Section 4 exemplifies our approach to generate differentially private data sets. Empirical work on the presented approach is reported in Section 5. Finally, in Section 6 we summarize conclusions and indicate future research lines.

2 Background

2.1 Principal component analysis

PCA is a dimensionality-reduction technique that uses an orthogonal transformation to convert a set of observations of possibly correlated attributes into a set of values of linearly uncorrelated attributes called *principal components*. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest possible variance under the constraint that it is orthogonal to the preceding components.

More formally, consider a data set X containing n records with m attributes, such that the attribute means are zero. The covariance matrix of X is

$$C_X = \frac{1}{n} X^T X.$$

We want to find a new basis $\{p_1, \ldots, p_m\}$ such that: (i) $Var(P_i) \geq Var(P_j)$, for all $i < j$, (ii) $Cov(P_i, P_j) = 0$, for all $i \neq j$ (low redundancy) and (iii) the vectors p_i are orthonormal.

The PCA transformation is attained by means of the change of basis matrix $P = [p_1 \dots p_m]$, where the p_i's are eigenvectors of C_X sorted by descending order of the corresponding eigenvalues. That is, the resulting data set is XP.

Since the principal components are sorted in descending order of the information they contain (the first component has the largest amount of information), dimensionality reduction is very simple. We just have to drop some components starting from the last one (which is the least informative one).

2.2 Differentially private data sets

Definition 1 (ϵ-differential privacy) *A randomized function κ gives ϵ-differential privacy (ϵ-DP, [6]) if, for all data sets X_1 and X_2 that differ in one record (a.k.a. neighbor data sets), and all $S \subset Range(\kappa)$, we have*

$$\Pr(\kappa(X_1) \in S) \leq \exp(\epsilon) \Pr(\kappa(X_2) \in S).$$

Given a query function f, the goal in DP is to find a randomized function κ_f that satisfies ϵ-DP and approximates f as closely as possible. For the case of numerical queries κ_f can be obtained via noise addition; that is $\kappa_f() = f() + N$, where N has been properly adjusted. Adding Laplace distributed noise whose scale has been adjusted to the global sensitivity of the query f is, probably, the most common approach (other alternatives include [16, 10, 12]). The global sensitivity or L_1-sensitivity, Δf, of a function $f : \mathcal{D}^n \to \mathbb{R}^d$ is the maximum variation of f between neighbor data sets:

$$\Delta f = \max_{d(X_1, X_2)=1} \|f(X_1) - f(X_2)\|_1 .$$

Proposition 2 *Let $f : \mathcal{D}^n \to \mathbb{R}^d$ be a function. The mechanism $\kappa_f(X) = f(X) + (N_1, \dots, N_d)$, where N_i are drawn i.i.d. from a Laplace$(0, \Delta f/\epsilon)$ distribution , is ϵ-DP.*

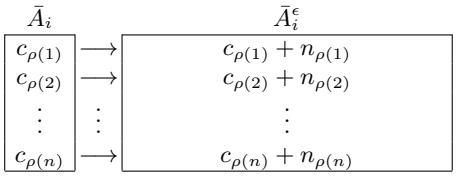
The above original definition of DP was intended for interactive queries to a database. If we want to obtain a DP data set, a straightforward procedure would be to add noise to the results of queries about individual attribute or record values. However, this is not feasible, because the sensitivity of individual records is large, and according to Proposition 2 a lot of noise would be needed and the results would be very inaccurate. Better methods to produce DP microdata sets can be found in the literature, based on histograms [7], microaggregation [17], Bayesian networks [20] or synthetic data [9, 14]. Although using histograms is especially problematic for high-dimensional data (due to sparsity), the rest of the above approaches also yield increasingly inaccurate results as the dimensionality of the original data set increases. In the rest of this paper, we will use the microaggregation-based procedure [17], that we next recall.

Microaggregation [4, 5] is a well-known SDC technique that works in two steps: i) split the original data set into clusters of at least k records following a criterion of maximum within-cluster similarity; ii) replace the records in each

cluster by the cluster centroid. In [18], the focus is moved from the original data set X to its microaggregated version $\bar{X}$. After $\bar{X}$ is generated, X is dropped and the goal becomes to protect $\bar{X}$ via DP. Since $\bar{X}$ contains less information than X, this change of focus does not increase the risk of disclosure: that is, $\bar{X}_\epsilon$ (an ϵ-DP version of $\bar{X}$) entails less disclosure risk than X_ϵ (an ϵ-DP version of X). Additionally, $\bar{X}$ is less sensitive to changes in one individual than X. In particular, when centroids are computed as arithmetic averages, the sensitivity of the centroid of a cluster C is a factor $1/|C|$ of the sensitivity of the original records.

Let X be the original data set (with attributes $A_1, \ldots, A_m$) and $\bar{X}$ be a microaggregated version of X (with attributes $\bar{A}_1, \ldots, \bar{A}_m$). Among the possibilities described in [17], we use the individual ranking microaggregation approach: each attribute is microaggregated separately from the rest. The DP data set $\bar{X}_\epsilon$ is generated by combining ϵ_i-DP versions of each of the attributes: by the sequential composition property of DP, if we have an ϵ_i-DP version of each attribute $\bar{A}_i$, the combination is $(\epsilon_1 + \ldots + \epsilon_m)$-DP.

To obtain the ϵ_i-DP version of attribute $\bar{A}_i$ we proceed as illustrated in Figure 1. Each centroid is masked with an appropriate amount of noise that depends on the cardinality $|C_j|$ of the corresponding cluster, on the sensitivity ΔA_i of the original attribute, A_i, and on the ϵ_i assigned to the attribute. For each cluster C_j of A_i values, we sample a noise n_j from a $Laplace(0, \Delta A_i/(|C_j| \times \epsilon_i))$ distribution and add it to the corresponding centroid c_j, to obtain $c_j + n_j$. It is important to remark that each occurrence of c_j is replaced by the same masked value $c_j + n_j$.



where $\rho(r)$ =number of record r's cluster
$n_j = Laplace(0, \Delta A_i/(|C_j| \times \epsilon_i))$

Fig. 1. Procedure to generate an $(\epsilon_1 + \ldots + \epsilon_m)$-DP data set by combining an ϵ_i-DP version of the attribute A_i , for $i = 1, \ldots, m$.

3 PCA-based dimensionality reduction

Given a data set X, PCA applies a change of basis to X. Let X^{PC} be the resulting data set and $\{p_1, \ldots, p_m\}$ be the new basis (the principal components). A change of basis is a bijective transformation: it alters the representation of the

information but does not alter the amount of information. Thus, both X and X^{PC} have the same risk of disclosure.

PCA concentrates the information in the first components. That is, the amount of information contained in p_i is greater than in p_j for any $i < j$. As explained in Section 2, we can take advantage of this property to reduce the dimensionality of the data set. This is done by dropping some of the last principal components (the ones with less information). Let $X_{PC'}$ be the data set generated from X_{PC} by dropping components $p_{l+1}, \ldots, p_m$. This transformation is not injective: as the result of dropping components, we reduce the amount of information in the data set. Thus, $X^{PC'}$ incurs no more risk of disclosure than X.

After generating an anonymized data set $Y^{PC'}$ by applying an appropriate statistical disclosure control method $\mathcal{G}$ to $X^{PC'}$, we would like to release the protected data set in terms of the original basis. That is, we want to generate an anonymized data set Y in the same basis as X. To do so, we need to apply the inverse PCA transformation. In other words, we need to change the basis from $\{p_1, \ldots, p_l\}$ to $\{x_1, \ldots, x_m\}$.

Algorithm 1 presents a formal description of the previously described steps.

Algorithm 1 Reducing data dimensionality via PCA in anonymization

Require:
X : original data set
$\mathcal{G}$: Anonymization algorithm

Let $X^{PC} :=$ PCA representation of X
Let $X^{PC'} :=$ projection of X^{PC} over the first l principal components
Let $Y^{PC'} := \mathcal{G}(X^{PC'})$
Let $Y :=$ result of applying the inverse PCA transformation to $Y^{PC'}$
Return Y

4 Example: differential privacy anonymization on PCA-reduced dimensionality data

We illustrate the PCA-based dimensionality reduction when the anonymization algorithm $\mathcal{G}$ in Algorithm 1 intends to reach differential privacy. Algorithm 1 entails a pre-processing of the original data set (PCA representation and projection over the first l principal components) and a post-processing of the DP data set (inverse PCA transformation).

It is well established that post-processing DP results preserves the same level of DP: if an output satisfies ϵ-DP, any transformation of it that adds no new information offers at least the same level of privacy. However, the PCA pre-processing step should be performed in a DP way. In [3] it is described how to attain DP in PCA.

In this section, we slightly depart from strict DP. We claim that, in our specific case, we can use a non-DP PCA transformation without degrading the privacy of subjects with respect to DP. If using a non-DP PCA transformation is safe, it should be preferred over DP PCA because it is simpler and yields more accurate data.

Our claim is based on the following observation: data sets X and X^{PC} are equivalent (they contain exactly the same information) and, hence, the privacy risk of X_ϵ (an ϵ-DP version of X) is equivalent to the privacy risk of X_ϵ^{PC} (an ϵ-DP version of X^{PC}). After generating X^{PC}, we drop some of the components to reduce the dimensionality of the data and obtain $X^{PC'}$. As the amount of information in $X^{PC'}$ is smaller than in X^{PC}, the privacy risk of $X^{PC'}$ is smaller than the privacy risk of X^{PC}, and thus than the privacy risk of X.

Even though publishing the principal components would entail privacy leakage for subjects (as principal components are based on the original data set X), we assert that the way PCA is used in Algorithm 1 does not increase the privacy leakage associated with the anonymization algorithm $\mathcal{G}$, because:

- Algorithm 1 does not disclose the data in their principal components representation: the PCA transformation is done in the first step and it is undone in the last step.
- The anonymized data Y returned for publication are computed based on an anonymized version $Y^{PC'}$ of the principal components coefficients. In the case of using DP as $\mathcal{G}$, Y is computed based on coefficients $Y^{PC'} := X_\epsilon^{PC'}$ that are DP; thus Y does not leak more private information than the one DP leaks about the coefficients.

5 Empirical work

In this section we present experimental work on the example given in Section 4 above. To evaluate the DP anonymization on PCA-reduced dimensionality data, we compared the accuracy of the DP data set generated using the microaggregation-based approach described in Section 2.2, with and without the proposed PCA-based dimensionality reduction methodology. In particular, we compared against the accuracy of DP data sets obtained in [13].

The empirical evaluation was performed on the Census data set [2], which is a widely used reference numerical data set and was also employed in [13]. Census contains 13 numerical attributes and 1080 records. For the sake of comparability with [13], we defined the domain of each attribute to range between 0 and 1.5 times the maximum attribute value in the data set. The difference between the bounds of the domain of each attribute A_i determines the sensitivity of that attribute (ΔA_i) and, as detailed above, determines the amount of Laplace noise to be added to microaggregated outputs. Since the Laplace distribution is unbounded, for consistency we bounded noise-added outputs to the domain ranges define above.

We used the same information loss measure as in [13], namely the relative error (RE). RE is measured as the absolute difference between each original

attribute value a and its masked version a' divided by the original value. A sanity bound was included in the denominator to mitigate the effect of very small values. We used the same sanity bound as in [13]: for an attribute A it is $|\max(Dom(A)) - \min(Dom(A))|/100$. Thus,

$$RE_A(a, a') = \frac{|a - a'|}{\max\{a,\, sanity_bound_A\}}.$$

When using PCA, the amount of dimensionality reduction is the number of discarded principal components. First, we examined the accuracy of the generated DP data set in terms of RE as we changed the amount of dimensionality reduction. In these experiments, we took $\epsilon = 1$ as the target DP level, which is smaller than the ϵ values usually chosen when generating DP data sets —in spite of DP not offering any effective privacy guarantee for large ϵ, it is not uncommon to see $\epsilon = 10$ or larger in this context. In Figure 2 we report the average RE (y-axis) over the remaining l principal components for different microaggregation cluster sizes (x-axis); each curve represents a different number l of remaining principal components.

RE is the result of three errors: the dimensionality reduction error, the microaggregation error, and the noise introduced by DP. We focused on analyzing the interplay between the error due to dimensionality reduction and the other types of errors.

The effect of dimensionality reduction is twofold. On the one side, the fewer the components that remain, the greater the error due to dimensionality reduction. On the other side, the fewer the components, the larger the share of ϵ we can assign to each component, which reduces the error due to DP. This is clear by looking at the curves with one and two components in Figure 2. For small values of k, the total error in the one-component curve is smaller. The reason is that, for a small k, the error due to DP dominates and having multiple components magnifies it. In contrast, for larger k, the error due to DP becomes less significant and it is preferable to decrease the error due to dimensionality reduction, even if that implies increasing the error due to DP.

Even if the total RE is a combination of opposing factors, it becomes apparent in Figure 2 that preserving a large number of components is, in general, not a good compromise. To emphasize this fact and show the real value of the proposed method we compared against the results in [13], which do not use dimensionality reduction. Figure 3 shows the RE of the method described in [13] when dealing with the 13 attributes of the Census data set. We observe in Figure 3 that for DP with $\epsilon = 1$ the the best RE is substantially greater than 1 (actually around 4) because the scale is logarithmic. Using PCA dimensionality reduction with 3 principal components we reached an RE of 0.5. Even though RE=0.5 may still be too much, the improvement is very significant. To reduce the RE to more acceptable rates, we could slightly increase ϵ (e.g. $\epsilon = 2$). Indeed, despite DP becomes meaningless if a large ϵ is used, greater values are common in DP data set generation. In contrast, increasing ϵ is not very useful in [13]: even using $\epsilon = 10$ (too large to offer any privacy guarantees) yields results worse than those obtained for $\epsilon = 1$ with our approach.

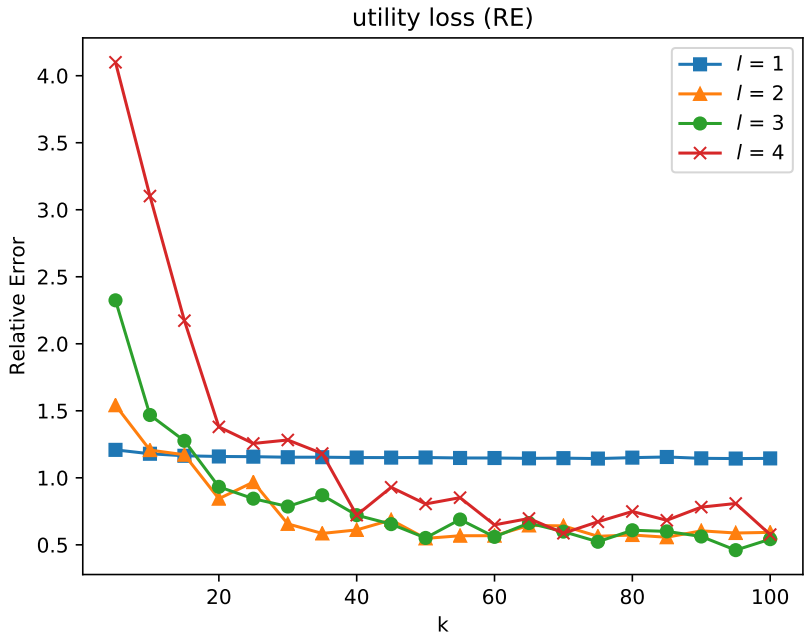


Fig. 2. Average RE (y-axis) of the DP data set for $\epsilon = 1$ in terms the cluster size (x-axis). Each curve corresponds to a different number l of remaining principal components.

6 Conclusions and future research

In the context of privacy-preserving data publishing, attaining privacy guarantees at the individual level while preserving a reasonable data utility is only feasible for low-dimensional data sets. We have presented an approach for dimensionality reduction that is aimed at improving the loss of information caused by anonymization.

We leverage principal components decomposition: subjects send their original records to the controller, who gathers the entire original data set and computes the principal components of it. The dimensionality reduction comes from the controller choosing only those components that carry most information.

We have exemplified our approach for the specific case of DP data set generation. We have shown that pre-processing data (to transform the original data set into its principal components) does not increase the risk of disclosure. Experimental results are based on the DP data set generation mechanism of [17], and *they show a very significant information loss reduction in the DP data set.*

Future research lines include the following:

– Amplify empirical work by working on data sets with more attributes and more records.

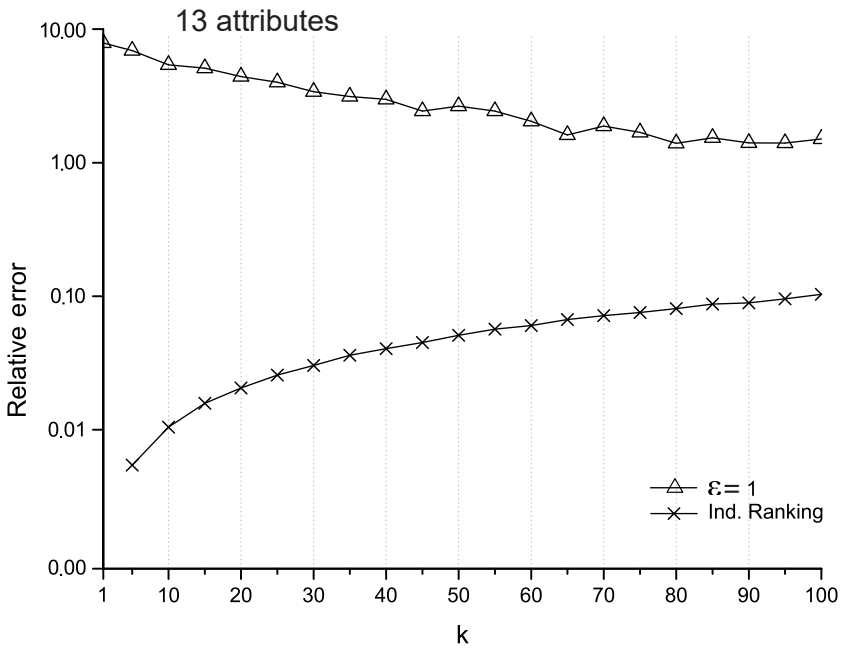


Fig. 3. Average RE (y-axis) of the DP data set in terms the cluster size (x-axis) of the DP data set generation method described in [13] and of the baseline microaggregation algorithm.

- Extend the PCA-based dimensionality reduction for categorical attributes. This will require resorting to dimensionality reduction techniques for categorical data, such as correspondence analysis or others.
- Study the case in which the controller is not trusted. In this case, the controller does not gather the entire original data set. Therefore, dimensionality reduction cannot be achieved via PCA as explained in this paper. Possible alternatives are to use randomized response to allow clustering of attributes, in such a way that attributes in different clusters are weakly dependent or nearly independent. This would make it possible to separately deal with each cluster of attributes, which would provide dimensionality reduction.

Acknowledgment and Disclaimer

We thank Fadi Hassan for help with the empirical work. Partial support to this work has been received from the European Commission (project H2020-700540 "CANVAS"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705), and from the Spanish Government (project RTI2018-095094-B-C21 "Consent"). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and are not necessarily shared by UNESCO.

References

1. M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. *New York Times*, Aug. 9, 2006.
2. R. Brand, J. Domingo-Ferrer and J. M. Mateo-Sanz. *Reference data sets to test and compare SDC methods for the protection of numerical microdata*. Deliverable of the CASC project (IST-2000-25069), 2002.
3. K. Chaudhuri, A. D. Sarwate and K. Sinha. A near-optimal algorithm for differentially-private principal components. *J. Mach. Learn. Res.*, 14(1):2905-2943, 2013.
4. J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189-201, 2002.
5. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery* 11(2):195-212, 2005.
6. C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference-TCC 2006*, pp. 265-284. Springer, 2006.
7. M. Hardt, K. Ligett and F. McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems-NIPS 2012*, pp. 2339-2347, 2012.
8. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.
9. A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering-ICDE 2008*, pp. 277-286. IEEE, 2008.
10. F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science- FOCS 2007*, pp. 94-103. IEEE Computer Society, 2007.
11. A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pp. 111-125. IEEE, 2008.
12. K. Nissim, S. Raskhodnikova and A. Smith. Smooth sensitivity and sampling in private data analysis. In *39th Annual ACM Symposium on Theory of Computing-STOC 2007*, pp. 75-84. ACM, 2007.
13. D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas. Utility-preserving differentially private data releases via individual ranking microaggregation. *Information Fusion*, 30:1-14, 2016.
14. J. Snoke and A. Slavković. p-MSE mechanism: differentially private synthetic data with maximal distributional similarity. In *Privacy in Statistical Databases-PSD 2018*, pp. 138-159. Springer, 2018.
15. O. Solon. 'Data is a fingerprint': why you aren't as anonymous as you think online. *The Guardian*, Jul. 13, 2018.
16. J. Soria-Comas and J. Domingo-Ferrer. Optimal data-independent noise for differential privacy. *Information Sciences*, 250:200–214, 2013.
17. J. Soria-Comas and J. Domingo-Ferrer. Differentially private data publishing via optimal univariate microaggregation and record perturbation. *Knowledge-Based Systems* 125:13-23, 2018.

18. J. Soria-Comas and J. Domingo-Ferrer. Differentially private data sets based on microaggregation and record perturbation. In *Modeling Decisions for Artificial Intelligence-MDAI 2017*, pp. 119-131. Springer, 2017.
19. L. Sweeney. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000.
20. J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava and X. Xiao. PrivBayes: private data release via Bayesian networks. *ACM Transactions on Database Systems* 42(4), Art. 25, 2017.