Transparency and microaggregation

Vicenç Torra*

* School of Informatics, University of Skövde, Skövde, 54128 Sweden
E-mail: vtorra@his.se

Abstract. Transparency has an important effect on disclosure risk. In general, masking methods have to be evaluated taking into account that intruders can use all available information to attack the data. When the masking method as well as their parameters are disclosed, this information can also be used by an intruder. In this talk we will review results on the effects of transparency in disclosure risk assessment for microdata giving special emphasis to microaggregation.

1 Introduction

The dissemination of databases and microdata to researchers for their analysis needs to avoid the disclosure of sensitive information. Statistical disclosure control and data privacy have developed masking methods to avoid this disclosure. See e.g. [5, 7] for details.

Masking methods ρ transform a database X into a database X' in a way that the analytical properties of X and X' are similar (data utility is high) but where disclosure risk is acceptable. That is, $X' = \rho(X)$ is such that the properties of X and X' are similar.

Different masking methods exist. They can be classified into perturbative methods, non-perturbative methods, and synthetic data generators. Perturbative methods intend to avoid disclosure adding some noise to the original data. Non-perturbative methods avoid disclosure reducing the level of detail of the data. Synthetic data generators [4] use the original data to make models and then replace the original data by artificial data generated by these models. Disclosure is avoided because data is artificial. In reality, none of these methods avoid disclosure completely, and disclosure risk has still to be considered and evaluated.

There exist several ways to measure disclosure risk, based on different assumptions on the possible type of attacks an intruder may consider. A usual classification is to distinguish between identity disclosure and attribute disclosure.

- Identity disclosure. Given the masked file X', we have identity disclosure when an intruder is able to link a record x' in X' to the entity that has supplied the data. E.g., in a data from a hospital, we link a record of a patient with the patient.

- Attribute disclosure. Given the masked file X', we have attribute disclosure when we can increase our knowledge on a given attribute for a particular entity. E.g., in a data from a hospital, we can increase our knowledge about the illness of a certain patient of the hospital.

Identity and attribute disclosure are independent, but it is natural to have attribute disclosure when we have identity disclosure.

Disclosure risk measures have been defined for both types of disclosure. In the particular case of identity disclosure, we have mainly two measures: uniqueness and record linkage.

In this paper we focus on measuring disclosure risk using record linkage. We discuss that the transparency principle affects the application of record linkage algorithms when evaluating disclosure risk. We focus on data protected with microaggregation.

The structure of the paper is as follows. In Section 2 we review record linkage as a way to measure identity disclosure risk. In Section 3 we review the transparency principle. In Section 4 we discuss the transparency attack. In section 5 we discuss an algorithm for fuzzy microaggregation, which is a version of microaggregation that avoids the transparency principle. The paper finishes with some conclusions and lines for future work.

2 The proportion of linked records as a measure of identity disclosure

The number of reidentifications (linked records) between a masked data file and the data file of an intruder has been used as a measure of identity disclosure risk. See e.g. [2, 21, 17]. For this purpose, record linkage algorithms, or in general, data matching algorithms [1] can be used. This approach was formalized in [19], and used in [18] in a few different scenarios of data privacy.

In this scenario, we consider an intruder with a given file Y and the protected data file X'. We usually presume that Y is a subset of X. In other words, we presume that the intruder has information on the same variables being published and about the same individuals whose information is being published.

Other scenarios have also been considered with intruders having other type of information. E.g., information on other variables and/or other individuals.

Then, we have identity disclosure when the intruder is able to successfully link one of his records in Y with one in X' and both correspond to the same individual. The number of correctly linked records, or the proportion of correcly linked records, for the file Y is a measure of the risk.

Naturally, the more effective the record linkage is, the larger the number of correct links. Because of that, we are interested in developing and applying the best record linkage algorithms.

In order to be effective, record linkage algorithms need to use all available information. The worst case scenario [13, 19] is an upper bound of the risk, when the most advanced record linkage is used taking into account all the available information.

3 Transparency principle

The term *transparency principle* was first formulated as such in statistical disclosure control by A. Karr in [8]. Karr defined transparency as "the release of information about processes and even parameters used to alter data".

Transparency has a positive effect when data is published, as researchers can exploit the released information to increase the accuracy of their results. For example, if we mask a single variable data file X by means of noise addition $\rho(X) = X + \epsilon$ where ϵ is such that $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given parameter k, then we know that

$$Var(X') = Var(X) + kVar(X) = (1 + k)Var(X).$$

Naturally, researchers can use the information on noise addition, the distribution of the noise ϵ and, in particular, the parameter k to compute the real value of $Var(X)$.

However, the information about the method and the parameters can also be used by an intruder to attack protected data. Attacks on masked data using this type of information were considered before Karr's definition above.

For example, Winkler [20] discusses and attacks single-ranking microaggregation using re-identification. The approach is based on the definition of a distance function such that records that may have been microaggregated together have a distance equal to zero. Nin et al. [13] proposed a different formalization to deal with the transparency principle. We review this approach below.

4 Transparency attacks using record linkage

When we use a reidentification method to evalute the worst case scenario, we need to take into account the masking method and its parameters, if this information is available to the intruder. This information can improve the reidentification process. Note that for some of the masking methods we know for sure that some possible linkages are impossible. That is, we know that not all records in X' can be the masked record of a given record x.

Therefore, we have that the record linkage cannot be applied to any pair of records, but only on appropriate pairs of records. The formalization follows.

Definition 4.1 *Let X be the original database where each record x is described in terms of attributes $\mathbf{V} = (V_1, \ldots, V_s)$. Let X' be the masked database represented with the same attributes. Here, $V_i(x)$ represents the value of the ith attribute for x and $V_i(x')$ be the value of the masked record x'.*

Then, for each $x \in X$, the intruder can build the set of records in X' that can correspond to the masked version of x. Let $B_j(x)$ denote the set of masked records associated to x when we only consider the jth variable. Then, for the record x, the masked record x_ℓ corresponding to x is in the intersection of $B_j(x)$. That is,

$$x_\ell \in \cap_j B_j(x). \tag{1}$$

Essentially, according to this formulation the transparency attacks can be seen as an intersection attack. Each attribute reduces the set of alternatives (the set of possible records).

For illustration, let us consider the case of univariate optimal microaggregation of a file with not repeated values (this assumption is to simplify the discussion). In this case, we have that for a given x optimal microaggregation can only mask $V_j(x)$ into two of the values in $V_j(X')$. They are $v_n = min_r\{V_r(X')|V_r(X') \geq V_j(X')$ and $v_x = max_r\{V_r(X')|V_r(X') \leq V_j(X')$. This permits to define $B_j(x)$. If we apply this for each attribute V_j, we can then apply Equation 1 for each set B_j.

The definition above shows that the larger the number of attributes, the smaller the set. In any case, the set can never be empty. Note that we are sure that the masked record is in the intersection. In addition, if the intersection set is a singleton, it is clear that this is the masked set and we have a match (a correct reidentification). Note that this is different than a linkage in record linkage where the intruder is not sure if the link is correct or not.

This transparency attack has been applied to univariate and multivariate microaggregation and to rank swapping. In particular, [14] is an effective attack for univariate optimal microaggregation. Experimental results are given that show the improvement of performance of record linkage when we take into account that we have protected the set using univariate optimal microaggregation. [12] discusses the case of multivariate microaggregation. [13] focuses on rank swapping, presenting an effective attack for this masking method.

5 Avoiding transparency attacks: the case of fuzzy microaggregation

When data is published under the transparency principle, masking methods need to be still effective when the intruder is aware of the protection method and the parameters used. In [13] a new rank swapping method resistant to transparency attacks was proposed. We will give an overview of the approach introduced in [16] for microaggregation that is resistant to the transparency attack. It is a type of fuzzy microaggregation.

The first fuzzy microaggregation was proposed in [3]. [16] is a simpler approach that permits a broad range of protections against identity disclosure.

The approach is based on fuzzy clustering [6, 9], which permits elements to belong (with partial membership) to different clusters at the same time. A cluster representative is selected for each cluster. Then, we use the fuzzy memberships to build a probability distribution, and then use this probability distribution to assign randomly each element to one of the clusters. Records are replaced by the cluster representative.

The main difference with standard microaggregation is that we do not necessarily assign a record to the one that minimizes the objective function. Therefore this information can not be used by the intruder.

The procedure requires three parameters. The first one is k and corresponds to the minimum number of elements in the clusters, as in standard microaggregation. In our approach, the minimum number of elements is not a strict constraint. Therefore, k elements are not always guaranteed in a cluster. The parameter is used to determine the number of clusters c, a parameter that the fuzzy c-means requires. In particular, we set $c = (|X|/k)$. Then, we need two values m_1 and m_2 which correspond to the parameter m in fuzzy c-means. Recall that m in fuzzy c-means corresponds to the degree of fuzziness in the solution of the fuzzy clustering. With $m = 1$ we have crisp clusters, and the larger the m, the fuzzier the solution.

The algorithm follows. In this algorithm, fcm corresponds to the application of the fuzzy c-means algorithm. Fuzzy c-means requires the data set to be clustered (X in our case), the number of clusters (the parameter c), and the degree of fuzziness (our parameter m_1). In our case the fuzzy c-means algorithm only returns the c cluster centers. In general, this fuzzy clustering algorithm also returns the membership values but we do not use them as we compute them using m_2.

Algorithm FCM-based microaggregation (X, k, m_1, m_2)

Step 1. $c = (|X|/k)$

Step 2. $V = fcm(X, c, m_1)$

Step 3. Recompute the membership of records x in X to cluster centers v_i in V using m_2:

$$\mu_i(x) = \Big(\sum_{j=1}^{c} \Big(\frac{||x - v_i||^2}{||x - v_j||^2} \Big)^{\frac{1}{m_2 - 1}} \Big)^{-1}$$

Step 4. Use $\mu_i(x)$ as a probability distribution to replace x by one of the v_i.

Fuzzy c-means is equivalent to k-means algorithm when $m_1 = 1$. This means that the optimal solution is one with c different clusters. With a large value of m_1 fuzzy c-means tends to locate all cluster centers in the mean position of X. That is $v_i = v_j = mean(X)$ for all i, j. With respect to m_2, we have that for large m_2, the definition of μ is such that the probability distribution is uniform. In contrast, for $m_2 = 1$, we have that there is only one cluster i with $\mu_i(x) = 1$ for all x.

Therefore, for $m_1 = m_2 = \infty$ (or values large enough as e.g. $m_1 = m_2 > 4$) we have that all records will be replaced by the mean value of X.

When $m_1 = 1$ and $m_2 = \infty$, we have c different cluster centers and a uniform distribution for each record. This implies that the protected file satisfies k-anonymity if we are applying microaggregation in a multivariate way.

On the other hand, when we have $c = |X|$ and $m_1 = m_2 = 1$ we have that we have as many clusters as original records. So, if the clustering algorithm properly assigns one cluster to each record, the masked file will be equal to the original one.

So, the masked file ranges between the original file and a k-anonymous one. The algorithm can also produce a file with $|X|$ copies of the mean of the set X. Therefore, different parameterizations permit different levels of risk and utility.

6 Conclusions

In this paper we have discussed the effect of the transparency principle in masking methods. We have reviewed the analysis of identity disclosure through record linkage under this principle.

We have given an overview of a fuzzy microaggregation method based on fuzzy c-means. We have discussed some of its properties. The application of this fuzzy microaggregation algorithm, as well as a more detailed description of the algorithm, can be found in [16].

As future work we plan to further study the application of this algorithm, and study how to find a good parameterization for a given dataset.

The fuzzy microaggregation algorithm proposed here is based on fuzzy c-means. Other fuzzy clustering algorithms could be used in the algorithm described above. As future work, we will consider the comparison between different clustering algorithms. In addition, we will consider the application of our approach to some of the new algorithms for microaggregation (as e.g. [10, 11]).

References

[1] Christen, P. (2012) Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Springer.

[2] Domingo-Ferrer, J. and Torra, V. (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, North-Holland, 111-134.

[3] Domingo-Ferrer, J., Torra, V. (2002) Towards fuzzy c-means based microaggregation, in P.Grzegorzewski, O. Hryniewicz, M. A. Gil (Eds), Soft Methods in Probability and Statistics, 289-294.

[4] Drechsler, J. (2011) Synthetic datasets for statistical disclosure control: Theory and implementation, Springer.

[5] Duncan, G. T., Elliot, M., Salazar, J. J. (2011) Statistical confidentiality, Springer.

[6] Höppner, F., Klawonn, F., Kruse, R., Runkler, T. (1999) Fuzzy cluster analysis, Wiley.

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

[8] Karr, A. F. (2009) The role of transparency in statistical disclosure limitation, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.

[9] Miyamoto, S. (1999) Introduction to fuzzy clustering (in Japanese), Ed. Morikita, Japan.

[10] Mortazavi, R., Jalili, S. (2014) Fast data-oriented microaggregation algorithm for large numerical datasets, Knowl.-Based Syst., 67 195-205.

[11] Mortazavi, R., Jalili, S. (2015) Preference-based anonymization of numerical datasets by multi-objective microaggregation, Information fusion 25 85-104.

[12] Nin, J., Herranz, J., Torra, V. (2008) On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering 67 399-412.

[13] Nin, J., Herranz, J., Torra, V. (2008) Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, 64:1 346-364.

[14] Nin, J., Torra, V. (2009) Analysis of the Univariate Microaggregation Disclosure Risk, New Generation Computing 27 177-194.

[15] Stokes, K., Torra, V. (2012) Reidentification and k-anonymity: a model for disclosure risk in graphs, Soft Computing 16:10 1657-1670.

[16] Torra, V. (2015) A fuzzy microaggregation algorithm using fuzzy c-means, Proc. CCIA 2015, IOS Press.

[17] Torra, V., Abowd, J. M., Domingo-Ferrer, J. (2006) Using Mahalanobis Distance-Based Record Linkage for Disclosure Risk Assessment, Lecture Notes in Computer Science 4302 233-242.

[18] Torra, V., Stokes, K. (2012) A formalization of record linkage and its application to data protection, Int. J. of Unc. Fuzziness and Knowledge Based Systems, 20:6 907-919.

[19] Torra, V., Stokes, K. (2013) A formalization of re-identification in terms of compatible probabilities, arXiv:1301.5022.

[20] Winkler, W. E. (2002) Single-ranking micro-aggregation and re-identification, Research Report Series #2002-08, U.S.Bureau of the Census.

[21] Winkler, W. E. (2004) Re-identification methods for masked microdata, PSD 2004, Lecture Notes in Computer Science 3050 216-230.