New Directions in Anonymization: Permutation Paradigm, Verifiability and Transparency

Josep Domingo-Ferrer
Universitat Rovira i Virgili, Tarragona, Catalonia - josep.domingo@urv.cat

Krishnamurty Muralidhar
University of Oklahoma, Norman OK, U.S.A. - krishm@ou.edu

Abstract

Current approaches to anonymization of microdata sets are either utility-first (use an anonymization method with suitable utility features, then evaluate the disclosure risk and, if needed, reduce the risk by possibly sacrificing some utility) or privacy-first (enforce a target privacy level via a privacy model, *e.g.* k-anonymity or ε-differential privacy, without regard to utility). The second approach is the only one that offers formal privacy guarantees, but it is seldom used in practice because it produces data releases with no utility guarantees. We address the previous conflict between utility and privacy, by showing how to get privacy guarantees without destroying more utility than necessary. Furthermore, we tackle the following unresolved issues: how to make anonymization verifiable by the data subject (so that she can verify how safe is the record she has contributed), how to get rid of background knowledge assumptions when defining the intruder, and what does transparency of anonymization to the user mean. We present a permutation paradigm of anonymization, whereby any microdata anonymization method is functionally equivalent to permutation plus a residual amount of noise addition. Thus, the privacy offered by a method is the amount of permutation it achieves and this amount can be verified not only by the data protector, but also by the subject contributing each record (subject-verifiability). Furthermore, we define an intruder model that makes no assumption on background knowledge and we show how to determine the right amount of permutation to withstand such an intruder without losing more utility than necessary. Finally, we show that an anonymization method safe against such an intruder can also be safely transparent to any user, which increases the analytical utility of anonymized data.
Keywords: data anonymization; statistical disclosure control; subject verifiability; intruder model; transparency to users.

1. Introduction

Privacy is a fundamental right mentioned in the Universal Declaration of Human Rights. Increasingly, privacy is understood as informational self-determination, that is, the capacity of an individual to determine the disclosure and user of his/her personal data.
Statistical disclosure control (SDC, Hundepool *et al.* (2012)) takes care of respondent/subject privacy by anonymizing three types of outputs: tabular data, interactive databases and microdata files. Microdata files consist of records each of which contains data about one individual subject (person, enterprise, etc.) and the other two types of output can be derived from microdata. Hence, we will focus on microdata. The usual setting in microdata SDC is for a data protector (often the same entity that owns and releases the data) to hold the original data set (with the original responses by the subjects) and modify it to reduce the disclosure risk. There are two approaches for disclosure risk control in SDC:

- *Utility first.* An anonymization method with a heuristic parameter choice and with suitable utility preservation properties[1] is run on the microdata set and, after that, the risk of disclosure is measured. For instance, the risk of re-identification can be estimated empirically by attempting record

[1]It is very difficult, if not impossible, to assess utility preservation for all potential analyses that can be performed on the data. Hence, by utility preservation we mean preservation of some preselected target statistics (for example means, variances, correlations, classifications or even some model fitted to the original data that should be preserved by the anonymized data).

linkage between the original and the anonymized data sets (see Torra and Domingo-Ferrer (2003)), or analytically by using generic measures (*e.g.*, Lambert (1993)) or measures tailored to a specific anonymization method (*e.g.*, Elamir and Skinner (2006) for sampling). If the extant risk is deemed too high, the anonymization method must be re-run with more privacy-stringent parameters and probably with more utility sacrifice.

- *Privacy first.* In this case, a privacy model is enforced with a parameter that guarantees an upper bound on the re-identification disclosure risk and perhaps also on the attribute disclosure risk. Model enforcement is achieved by using a model-specific anonymization method with parameters that derive from the model parameters. Well-known privacy models include ε-differential privacy (Dwork, 2006), ε-indistinguishability (Dwork *et al.*, 2006), k-anonymity (Samarati and Sweeney, 1998) and the extensions of the latter taking care of attribute disclosure, like l-diversity (Machanavajjhala *et al.*, 2007), t-closeness (Li *et al.*, 2007), (n,t)-closeness (Li *et al.*, 2010), crowd-blending privacy (Gehrke *et al.*, 2012) and others. If the utility of the resulting anonymized data is too low, then the privacy model in use should be enforced with a less strict privacy parameter or even replaced by a different privacy model.

Currently, there are several shortcomings in anonymization:

- *Comparability.* Comparing anonymization methods for microdata in terms of utility and disclosure risk is made difficult by the diversity of the principles they rely upon.
- *Verifiability by subjects.* Current anonymization does not favor the subject's information self-determination. The data releaser/protector takes legal responsibility for the release and makes all choices (anonymization method, parameters, privacy and utility levels, etc.). Moreover, the subject cannot *verify* whether her record is getting adequate protection.
- *Intruder's background.* In the utility-first approach and the privacy-first approach based on the k-anonymity family, restrictive assumptions need to be made on the intruder's background knowledge (namely, one assumes that the intruder can only link with external identified data sets through a subset of the attributes, called *key attributes*). In ε-differential privacy, no restrictive assumptions need to be made, but enough perturbation is needed to make presence/absence of any particular record unnoticeable in the anonymized data; as a consequence, data utility is substantially damaged.
- *Transparency to users.* The question here is how much detail shall/can be given to the user on the masking methods and parameters used to anonymize a data release. Clearly, the user derives more inferential utility from increased detail (Cox *et al.*, 2011). Yet, some methods may be vulnerable if too much detail on them is given.

In Section 2, we present a *reverse-mapping* procedure that, for any anonymization method, allows mapping the anonymized attribute values back to the original attribute values, thereby preserving the marginal distributions of original attributes. Based on reverse mapping, we show in Section 3 that any anonymization method is functionally equivalent to a permutation supplemented by some small, residual noise (*permutation paradigm*). Thus, permutation is the essential principle of anonymization, which allows giving simple utility and privacy metrics, as well as comparing methods. In Section 4, we introduce a new privacy model, $(d, \mathbf{v})$-permuted privacy. The subject can verify to what extent the privacy guarantee of the model holds for her record (*subject-verifiability*). In Section 5, we introduce a *maximum-knowledge intruder model*, which avoids assumptions on the intruder's background knowledge. We show how to protect against such a powerful intruder. In Section 6 we make the case for *anonymization transparent to the data user*. Finally, we conclude in Section 7.

2. Reverse-mapping anonymized data

We next recall a reverse-mapping procedure, which we first gave in the conference paper Muralidhar *et al.* (2014) in another context. Let $X = \{x_1, x_2, \cdots, x_n\}$ the values taken by attribute X in the original data set. Let $Y = \{y_1, y_2, \cdots, y_n\}$ represent the anonymized version of X. We make no assumptions about the anonymization method used to generate Y, but we assume that the values in both X and Y can be ranked

in some way[2]; any ties in them are broken randomly. Knowledge of X and Y allows deriving another set of values Z via reverse mapping, as per Algorithm 1.

Algorithm 1 Reverse-mapping conversion

Require: Original attribute $X = \{x_1, x_2, \cdots, x_n\}$
Require: Anonymized attribute $Y = \{y_1, y_2, \cdots, y_n\}$
  for $i = 1$ to n **do**
    Compute $j = \text{Rank}(y_i)$
    Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of X of rank j)
  end for
  return $Z = \{z_1, z_2, \cdots, z_n\}$

Releasing the reverse-mapped attribute Z instead of Y has a number of advantages. By construction, each reverse-mapped attribute preserves the rank correlation between the corresponding anonymized attribute and the rest of attributes in the data set; hence, *reverse mapping does not damage the rank correlation structure of the original data set more than the underlying anonymization method.* In fact, Z incurs less information loss than Y since *Z preserves the marginal distribution of the original attribute X. Disclosure risk can be conveniently measured by the rank order correlation between X and Z* (the higher, the more risk).

3. A permutation paradigm for anonymization

Reverse mapping has the following broader conceptual implication: any anonymization method is *functionally equivalent* to a two-step procedure consisting of a permutation step (mapping the original data set $\mathbf{X}$ to the data set $\mathbf{Z}$ obtained by running the reverse mapping procedure in Algorithm 1 for all attributes) plus a noise addition step (adding the difference between $\mathbf{Z}$ and the anonymized data set $\mathbf{Y}$). Furthermore, the noise addition is necessarily small, say residual, because it cannot entail any rank change ($\mathbf{Z}$ and $\mathbf{Y}$ have identical ranks, by Algorithm 1).
In this light, it seems rather obvious that protection against re-identification via record linkage comes from the permutation step in the above functional equivalence. Thus, *any two anonymization methods can, however different their actual operating principles, be compared in terms of how much permutation they achieve, that is, how much they modify ranks.*

4. Subject-verifiable privacy

We have shown that anonymization basically amounts to permutation. Hence, it seems natural to propose a privacy model focusing on permutation, which we call $(d, \mathbf{v})$-permuted privacy.
Given a non-negative integer d and an m-dimensional vector $\mathbf{v}$ of non-negative real numbers, an anonymized data set with m attributes is said to satisfy $(d, \mathbf{v})$-permuted privacy *with respect to original record* $\mathbf{x}$ if,

1. The permutation distance for $\mathbf{x}$ is at least d in the following sense: given the anonymized attribute values $y_*^1, \cdots, y_*^m$ closest to the respective attribute values of $\mathbf{x}$, no anonymized record $(y_p^1, \cdots, y_p^m)$ exists such that the ranks of y_p^j and y_*^j differ less than d for all $j = 1, \cdots, m$.
2. For $1 \leq j \leq m$, if y_*^j is the value of the anonymized j-th attribute Y^j closest to the value x^j of the j-th attribute of $\mathbf{x}$, and $S^j(d)$ is the set of values of the sorted Y^j whose rank differs no more than d from y_*^j's rank, then the variance of $S^j(d)$ is greater than the j-th component of $\mathbf{v}$.

An anonymized data set is said to satisfy $(d, \mathbf{v})$-permuted privacy if it satisfies $(d, \mathbf{v})$-permuted privacy with respect to all records in the original data set.

[2]For numerical or categorical ordinal attributes, ranking is straightforward. Even for categorical nominal attributes, the ranking assumption is less restrictive than it appears, because semantic distance metrics are available that can be used to rank them (for instance, the marginality distance in Domingo-Ferrer *et al.* (2013) and Soria-Comas *et al.* (2014)).

Obviously, the data protector, who has access to the entire original data set and the entire anonymized data set, can verify as described in this section whether the anonymized data set satisfies $(d, \mathbf{v})$-permuted privacy for any d and $\mathbf{v}$ of his choice. The most interesting feature, however, is that *each subject can check whether $(d, \mathbf{v})$-permuted privacy with respect to her original record is satisfied by the anonymized data set for some d and $\mathbf{v}$ of her choice*. The subject only needs to know her original record and the anonymized data set.

5. Maximum-knowledge intruder model

In cryptography, several different attack scenarios are distinguished depending on the intruder's knowledge: ciphertext-only (the intruder only sees the ciphertext), known-plaintext (the intruder has access to one or more pairs of plaintext-ciphertext), chosen-plaintext (the intruder can choose any plaintext and observe the corresponding ciphertext), chosen-ciphertext (the intruder can choose any ciphertext and observe the corresponding plaintext).

In anonymization, we can equate the original data set to a plaintext and the anonymized data set to a ciphertext. Hence, a ciphertext-only attack would be one in which the intruder has access only to the anonymized data: this class of attacks can be dangerous, as shown by Sweeney *et al.* (2013) for de-identified DNA data, by Narayanan and Shmatikov (2008) for Netflix data and by Barbaro and Zeller (2006) for the AOL data. Even if potentially dangerous, assuming that the intruder only knows the anonymized data can be naïve in some situations. For example, if the intruder is one of the subjects in the data set, he will normally know his own original record.

On the other hand, the strongest attacks in cryptography, namely chosen-plaintext and chosen-ciphertext attacks, assume some interaction between the intruder and the encryption system. Thus, they are not relevant in a non-interactive anonymization setting such as the one we are considering (release of anonymized data sets).

Hence, the strongest attack that anonymization of data sets must face is the known-plaintext attack. In this attack, one might think of an intruder knowing particular original records and their *corresponding* anonymized versions; however, this is unlikely, because anonymization precisely breaks the links between anonymized records and corresponding original records. A more reasonable definition for *a known-plaintext attack in anonymization is one in which the intruder knows the entire original data set (plaintext) and the entire corresponding anonymized data set (ciphertext), his objective being to recreate the* correct linkage *between the original and the anonymized records.*

We observe that our definition of the intruder is stronger than any other prior such definition in the data set anonymization scenario. One of the key issues in modeling the intruder in this context is to define his prior knowledge, including available background knowledge. As mentioned above, we assume that the intruder has maximum knowledge: he knows $\mathbf{X}$ and $\mathbf{Y}$, from which he can recreate $\mathbf{Z}$ by reverse mapping; hence, he only lacks the key, that is, the correct linkage between $\mathbf{X}$ and $\mathbf{Z}$. In particular, assuming knowledge of $\mathbf{X}$ by the intruder eliminates the need to consider the presence/absence of external background knowledge (typically external identified data sets linkable through quasi-identifiers) when evaluating the ability of the intruder to disclose information. In this respect, the intruder's background knowledge is as irrelevant in our intruder model as it is in ε-differential privacy.

From the perspective of our intruder, however, all anonymization procedures are reduced to permutations of the original data. Thus, the best option to guess which original record corresponds to which permuted record is to compute the permutation distance between the reverse-mapped data set $\mathbf{Z}$ and the original data set $\mathbf{X}$. The intruder will take as anonymized record corresponding to a specific original record the anonymized record at shortest permutation distance. If the intruder finds that there are many ties (a tie means that an original record has several anonymization records at the same shortest permutation distance) and/or that a substantial fraction of anonymized records do not correspond to any original record, then the intruder will conclude that his record linkages are not reliable *Note that the data protector can also check this and be reassured that the intruder is not getting good linkages.*

The inability of an intruder to assess the accuracy of re-identification via record linkage is often viewed as providing plausible deniability to the data protector. In other words, even if the intruder boasts the record linkages he has computed, he cannot prove *with certainty* which linkages are correct. Hence, any subject seeing that she has been correctly re-identified by the intruder could be reassured by the data protector that

re-identification has occurred by chance alone without the intruder really being sure about it.
However, an intruder with the knowledge specified in Section 5 can perform the analysis described in this section *to verify how likely it is for his computed record linkages to be correct.* To do this, the intruder simply does:

1. Generate a large random set **A** of values by drawing from the original data **X**. If computationally feasible, **A** should contain all possible combinations of attribute values, so that $\mathbf{X} \subset \mathbf{A}$ and $\mathbf{Z} \subset \mathbf{A}$. If not, **A** should be at least substantially larger than **X** and **Z**.
2. Determine the permutation distances at which matches occur between records in **A** and records in **Z**.
3. If the distribution of the permutation distances for the matches between **A** and **Z** overlaps with the distribution of the permutation distances of the intruder's matches between **X** and **Z**, then the intruder's matches are plausibly random and the intruder has no reason to boast about them.

In fact, *the data protector can also perform the above analysis before releasing the anonymized data, in order to determine the optimum amount of permutation (and hence of utility loss) needed for both distributions to have a sufficient overlap.* In that way, the data protector can make sure with minimum utility loss that any matches by any intruder will be plausibly random, with makes the anonymized data safe.

6. Transparency to users

Nowadays, statistical agencies and other data releasers often refrain from publishing the parameters used in anonymization (variance of the added noise, proximity of swapped values, group size in microaggregation, etc.). The exception is when the privacy-first approach is used (based on a privacy model), in which case the anonymization parameters are explicit and dictated by the model. However, as mentioned above, most real data releases are anonymized under the utility-first approach. Withholding the parameters of anonymization is problematic for at least two reasons:

1. The legitimate user cannot properly evaluate the utility of the anonymized data.
2. Basing disclosure protection on the secrecy of the anonymization algorithm and its parameterization is a poor idea, as it is hard to keep that much information secret and it is better to expose algorithms and parameterizations to public scrutiny to detect any weaknesses in them.

Transparency to the data user means giving to the user all anonymization details (except any random seeds used for pseudo-randomization).
We claim that being transparent to the user is of no consequence to the other stakeholders in the anonymization process. Indeed, our maximum-knowledge intruder can perform the above-described shortest-distance record linkage and match analysis without using any information on the anonymization method. The subject can verify how permuted her data have been regardless of whether the user is given details on the anonymization method. For the above reasons, the data protector does not lose anything by being transparent to the user. Hence, transparency is good for the user, and neutral to the intruder, the subject and the data protector.

7. Conclusions

We have presented a new vision of microdata anonymization as basically consisting in permutation. This simplies comparing methods, and also assessing and trading off utility against disclosure risk. We have proposed a new privacy model capturing the amount of permutation achieved. Then we have defined a maximum-knowledge intruder model, which allows getting rid of assumptions on background knowledge. Furthermore, we have shown how the intruder and the data protector can verify how good are the intruder's matches; in fact, the data protector can use this verification procedure to use the minimum permutation (and hence incur the minimum utility loss) required for the intruder's matches to be plausibly random. Finally, we have argued that transparency to users is good for users and has no consequence for the rest of stakeholders; therefore, transparency to users should be the general rule.

Open research issues and a running numerical example illustrating the concepts and algorithms proposed in the above sections can be found in the preprint Domingo-Ferrer and Muralidhar (2015).

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: Government of Catalonia (ICREA Acadèmia Prize to the first author and grant 2014 SGR 537), Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY"), European Commission (projects FP7 "DwB", FP7 "Inter-Trust" and H2020 "CLARUS"), Templeton World Charity Foundation (grant TWCF0095/AB60 "CO-UTILITY") and Google (Faculty Research Award to the first author). The first author is with the UNESCO Chair in Data Privacy. The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO, the Templeton World Charity Foundation or Google.

References

Barbaro, M., & Zeller, T. (2006) A face is exposed for AOL searcher no. 4417749. New York Times.

Cox, L., Karr, A. F., & Kinney, S. K. (2011) Risk-utility paradigms for statistical disclosure limitation. International Statistical Review, 79(2):160-183.

Domingo-Ferrer, J., & Muralidhar, K. (2015) New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. Preprint available at `http://arxiv.org/abs/1501.04186`

Domingo-Ferrer, J., Sánchez, D., & Rufian-Torrell, G. (2013) Anonymization of nominal data based on semantic marginality. Information Sciences, 242:35-48.

Duncan, G. T. & Pearson, R. W. (1991) Enhancing access to microdata while protecting confidentiality: prospects for the future. Statistical Science, 6(3):219-232.

Dwork, C. (2006) Differential privacy. In ICALP'06, LNCS 4052, Springer, pp. 1-12.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In TCC'06, LNCS 3876, Springer, pp. 265-284.

Elamir, E.A.H., & Skinner, C.J. (2006) Record level measures of disclosure risk for survey microdata. Journal of Official Statistics, 22(3):525-539.

Gehrke, J., Hay, M., Lui, E., & Pass, R. (2012) Crowd-blending privacy. In CRYPTO'12, LNCS 7417, Springer, pp. 479-496.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K., & De Wolf, P.-P. (2012) Statistical Disclosure Control. Wiley.

Lambert, D. (1993) Measures of disclosure risk and harm. Journal of Official Statistics, 9(3):313-331.

Li, N., Li, T., & Venkatasubramanian, S. (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In ICDE'07, pp. 106-115.

Li, N., Li, T., & Venkatasubramanian, S. (2010) Closeness: a new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering, 22(7):943-956.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007) l-Diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, 1(1):3, 2007.

Muralidhar, K., Sarathy, R., & Domingo-Ferrer, J. (2014) Reverse mapping to preserve the marginal distributions of attributes in masked microdata. In PSD'14, LNCS 8744, Springer, pp. 105-106.

Narayanan, A., & Shmatikov, V. (2008) Robust de-anonymization of large data sets. In IEEE Security & Privacy Conference, pp. 111-125.

Samarati, P., & Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep., SRI International.

Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., & Martínez, S. (2014) Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB Journal, 23(5):771-794.

Sweeney, L., Abu, A., & Winn, J. (2013) Identifying participants in the personal genome project by name. Harvard University, Data Privacy Lab. White paper no. 1021-1.

Torra, V., & Domingo-Ferrer, J. (2003) Record linkage methods for multidatabase data mining. In Information Fusion in Data Mining. Springer, pp. 99-130, 2003.