Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Helsinki, Finland, 5 to 7 October 2015)

Topic (i): Identifying and measuring risk of statistical disclosure

Microdata masking as permutation

Krishnamurty Muralidhar[*] and Josep Domingo-Ferrer[**]

[*] Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman OK 73019-4007, USA, e-mail krishm@ou.edu

[**] UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, e-mail josep.domingo@urv.cat

Abstract: Current research on microdata masking has considerable diversity of thought, borrowing principles from computer science, statistics, and other areas to develop, modify, and enhance masking methods for microdata. But this diversity also makes it difficult to evaluate and select the best method to achieve an optimum trade-off between analytical validity and disclosure prevention. In this paper, we define a cryptology-inspired maximum-knowledge adversary. We then show that, for this adversary, any microdata masking method is functionally equivalent to a permutation. Based on the maximum-knowledge adversary and the permutation model, we propose subject-verifiable privacy measures at the data set level and at the record level. We believe that this general model allows comparing different microdata masking methods without limiting the diversity of thought in microdata masking research.

1 Introduction

Statistical disclosure limitation research has a rich history in aiding data administrators by developing tools and techniques that allow them to release microdata while preventing disclosure of identity and value (Hundepool et al. 2012). In developing these techniques SDL researchers have drawn from research in other areas of study, including computer science, mathematics, statistics, and others. SDL researchers are also a diverse group representing computer science, statistics, business, law, public administration, and healthcare, to name a few. This diversity of fields and interests is not surprising since data exist in practically every endeavor and sharing such data is vital to advance scientific knowledge.

While this diversity is to be celebrated, it has also led to one particularly undesirable consequence – a lack of consensus on the evaluation of microdata masking methods. This is not very surprising. Researchers from different areas place different levels of importance on different criteria. Computer science researchers may place a greater

emphasis on preserving privacy, while social scientists or medical researchers (who wish to use the data) may place a greater emphasis on preserving utility. This has led to two different approaches: utility-first (the method and/or the parameters of the method are chosen so as to satisfy some pre-specified level of utility) and privacy-first (the method and/or the parameters are chosen so as to satisfy some pre-specified level of privacy).

Unfortunately, neither utility-first nor privacy-first models are capable of describing the diverse data masking methods that have been proposed. Perhaps more importantly, the definitions of utility and privacy are not independent of the metrics used to evaluate the performance of a particular method. For example, if we choose utility-first and wish to preserve a certain level of relationship among the variables, then additive perturbation with noise variance that is derived from the specified utility model may be selected. Similarly, if we choose privacy-first under the *k*-anonymity privacy model (Samarati and Sweeney, 1998), then by definition, privacy is assured when *k*-anonymity is satisfied. But this precludes comparing across masking methods that employ different approaches using different metrics that may be applicable. It is likely that some readers would by now be thinking of differential privacy (Dwork 2006); however, differential privacy, by definition, is predicated on the notion of "response to a query" and is not necessarily the most appropriate privacy model for microdata masking.

In addition, the key concern with releasing microdata is the re-identification of the individuals about whom data has been released. Almost all recent "disclosure" scenarios focus on re-identification of individuals, such as re-identification of DNA records (Sweeney et al. 2013), Governor William Weld (Sweeney, 2002), Netflix (Narayanan and Shmatikov, 2008), and AOL (Barbaro and Zeller, 2006). Yet we have no formal privacy definition of this type of attack.

The objective of this paper is to provide a new, general model of microdata masking that is capable of describing practically all masking methods, and a privacy model focusing on re-identification risk. In addition, the model that we describe is independent of the method, the parameters of the masking method, or the metrics used to evaluate privacy and utility.

Section 2 describes the permutation model of microdata masking. Section 3 describes our maximum-knowledge, purely malicious adversary. The advantages of the permutation model and adversary model to compare the performance of methods are highlighted in Section 4. Section 5 contains an empirical illustration of those advantages. Conclusions are summarized in Section 6.

2 The permutation model of microdata masking

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 **X**. Let $Y = \{y_1, y_2, \cdots, y_n\}$ represent the anonymized version of X in the anonymized data set **Y**. 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 (beyond numerical or ordinal values, rankings can also be computed for nominal values, e.g. based on marginality as per Domingo-Ferrer et al. (2013)); any ties in the ranking are broken randomly. Knowledge of X and Y allows deriving another set of values Z via reverse mapping, as per the following algorithm.

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 $Z = \{z_1, z_2, \cdots, z_n\}$

Algorithm 1 can be independently run for each attribute of the original data set **X** and corresponding attribute of **Y**. In this way, a data set **Z** is obtained. Since we make no assumptions about the masking method used to generate **Y** from **X**, conceptually, *any microdata masking method is functionally equivalent to doing the following: i) permute the original data set* **X** *to obtain* **Z***; ii) add some noise to* **Z** *to obtain* **Y**. The noise used to transform **Z** into **Y** is necessarily small (residual) because it cannot change any rank: note that, by the construction of Algorithm 1, the ranks of corresponding values of **Z** and **Y** are the same. It is important to note that the functional equivalence described does not imply any actual change in the masking method: we are simply saying that the way the method transforms **X** into **Y** could be exactly mimicked by first permuting **X** and then adding residual noise, that is, **X** → **Z** → **Y**. Thus, microdata masking methods can be described in terms of two components, the permutation component and the residual noise component.

The above approach can be used to describe any microdata masking method with one exception: the fully synthetic multiple imputation approach proposed by Rubin (1993). Since there exists no one-to-one correspondence between the original and masked data in this case, reverse mapping cannot be performed. However, in practice, fully synthetic multiple imputation is not frequently used, and partially synthetic multiple imputation is the preferred approach. Hence we do not consider this to be a serious limitation of our approach.

3 The adversarial model

One of the complexities of microdata masking is the definition of the adversary. It is practically impossible to distinguish between a legitimate user of the data and an adversary who wishes to misuse the data. This gives rise to the question of how much knowledge the user/adversary has. To avoid this problem, in this study, we assume a maximum-knowledge adversary as specified below.

Our adversary is inspired in cryptology, where different attack scenarios are distinguished depending on the adversary's knowledge: ciphertext-only (the adversary only sees the ciphertext), known-plaintext (the adversary has access to one or more pairs of plaintext-ciphertext), chosen-plaintext (the adversary can choose any plaintext

and observe the corresponding ciphertext), chosen-ciphertext (the adversary can choose any ciphertext and observe the corresponding plaintext). In microdata masking, we can equate the plaintext to the original data and the ciphertext to the masked data. The ciphertext-only attack is a naïve one, as it assumes that the adversary knows nothing. Protecting only against this type of attack when masking microdata could easily result in disclosure as shown by the DNA, Netflix and AOL attacks mentioned above. In contrast, the chosen-plaintext and chosen-ciphertext attack assume that the adversary has the ability to interact with the encryption method. But in microdata masking, no such interaction is possible. Hence, the most appropriate attack scenario for microdata masking is the known-plaintext attack.

We define the known-plaintext attack for microdata masking as one in which the adversary knows the entire original (plaintext) and masked (ciphertext) data set (but not their linkage), his objective being to recreate the correct linkage *between the original and the masked records*. Since this adversary already knows the original data set (maximum-knowledge), he gains no additional knowledge from the release of the masked microdata set. The intent of the adversary is purely malicious – he is interested in embarrassing the data administrator by identifying the correspondence between the original and masked microdata.

Note that our adversary has the capability to perform reverse mapping as specified by Algorithm 1. Given a masked microdata set **Y**, our adversary can compute $\mathbf{Y} \rightarrow \mathbf{Z}$, thereby eliminating the residual noise component of the masking mechanism. Thus, from the perspective of our adversary, *the only masking that is relevant is the permutation component.*

The adversary that we describe here is not new in the microdata masking literature. Most record linkage procedures assume just such an adversary (see the extensive work by William Winkler (2004). However, other researchers argue that assuming such an adversary represents the worst-case scenario (Skinner, 2008). While this may be true, assuming such an adversary also facilitates the comparison of the masking techniques using a common benchmark. Without such an assumption, such a comparison would be difficult, if not impossible.

Finally, in microdata masking, it is often argued that it is difficult (or even impossible) to distinguish between a legitimate user and an adversary – every user has to be treated as a potential adversary. But while a legitimate user learns something from the masked microdata that she did not know before, our purely malicious adversary has nothing to gain from the released microdata. In this he clearly differs from a legitimate user.

4 Advantages of the permutation model

4.1 Comparing the performance of a masking method on different data sets

Traditionally, the level of protection offered by a masking method is measured by the values of the method's parameter. This approach has several problems. First and foremost, the same parameter value used on two different data sets may yield completely different results. For the purposes of illustration, consider the multiplicative noise method. Assume that two data sets are to be masked using this

method with a 10% change parameter (the multiplicative factor is selected from Uniform(0.9, 1.1)). The following table provides the results for two different data sets.

Table 1. Levels of protection (permutation) offered by multiplicative noise with 10% parameter to two different data sets

Observation	Data set 1	Data set 2	Masking value	Masked data set 1	Masked data set 2	Rank of masked data set 1	Rank of masked data set 2
1	1	1001	1.091	1.091	1092.361	1	7
2	11	1011	0.907	9.981	917.356	2	1
3	21	1021	1.004	21.077	1024.742	3	4
4	31	1031	1.003	31.088	1033.931	4	5
5	41	1041	1.090	44.676	1134.341	5	10
6	51	1051	1.051	53.601	1104.594	6	8
7	61	1061	0.929	56.677	985.814	7	2
8	71	1071	0.985	69.964	1055.371	8	6
9	81	1081	0.912	73.901	986.258	9	3
10	91	1091	1.015	92.334	1106.988	10	9

The second data set is a simple transformation of the first data set with a value of 1000 being added to each observation. Yet, when the same masking approach is applied to this data set, it results in a considerably different level of masking. The observations in the first data set, while modified, retain the original rankings. However, the rankings of the observations in the second data set are quite different from the original rankings. Even to the naïve observer, it is evident that the multiplicative method with 10% modification results in much higher protection for the second data set than for the first data set. The permutation approach would allow for a more accurate comparison of the effectiveness of the masking methods.

More generally, the distributional characteristics of the data have a significant impact on the level of protection offered by a particular method. For some methods (such as additive noise), changes in the magnitude of the microdata may have no impact on the masking, while for other methods (such as multiplicative noise), a change in the magnitude may have a considerable impact on the level of masking. The example in Table 1 provides a simple illustration of this issue.

As shown in Table 1, this does not present a problem for the permutation model, which is able to correctly assess the extent of permutation. More importantly, for a data set of a given size, the extent of the permutation is independent of the distributional characteristics of the data set.

4.2 Comparing different masking methods on the same data set

We have shown above that the value of the masking parameter is not a good measure of the protection provided by a method, because protection is data-dependent. Here we argue that the parameter values of different methods give no clue to compare the protection these methods offer for a given data set. This is largely due to parameters having different semantics for different methods. For example, the parameter for

additive noise is usually related to the variance of the microdata; the parameter for multiplicative noise is usually specified as a percentage; the parameter for micro-aggregation is usually specified as a number of records; and finally, synthetic data approaches either have no parameters or they use a model specifying a relationship among attributes. The problem with these different forms of specification is that the data administrator is at a loss when comparing methods. How does specifying additive noise variance to be 10% of the variance of the original data set compare with specifying multiplicative noise drawn from a uniform distribution set as $\pm$10% of the original values? Will specifying 10% level for both additive and multiplicative masking result in the same level of protection? What would be the comparable level of protection with micro-aggregation and synthetic data approaches? Thus, while the parameters enable implementing a particular method, they do not allow the administrator to effectively compare methods.

Alternatively, studies have suggested that the variance of the difference between the original and masked data can be used as the level of privacy of the method (Adam and Wortmann, 1989; Muralidhar and Sarathy, 1999). In this model, higher variance represents higher privacy (and lower disclosure risk). This measure is usually presented as the variance of the difference between the original and masked values (Variance of (X – Y)). For the additive noise method, this approach represents a natural measure of privacy since a larger noise variance naturally means greater masking of the original values. For other methods, however, it is not clear that this approach necessarily represents a good comparative measure. Even for additive noise, there is no clear understanding of the relationship between the variance of the noise added and the level of protection afforded by the method.

In order for the variance of the noise to represent a good measure of privacy, it would be necessary for the disclosure resulting from all methods with the same level of noise variance to be exactly the same. But this has been shown to be incorrect in previous studies. For example, adding correlated noise has been shown to have different disclosure risk characteristics compared to adding uncorrelated noise, even though, at the individual variable level, the variance of the noise is exactly the same. Similar differences have been observed for synthetic data methods compared to additive noise methods. Hence, using noise variance to compare across methods has limited value.

Regardless of the underlying masking mechanism, the permutation model overcomes the above difficulties and allows comparing methods based on the extent to which the masked values have been permuted with respect to the original values. As observed earlier, the data administrator does not have to modify the masking mechanism in any way to assess the level of permutation.

Another frequently used approach to compare the protection that several masking methods afford for a given data set is to use an empirical disclosure risk assessment (e.g. based on record linkage) and vary the parameters of the different methods to achieve approximately the same level of disclosure. The problem with this approach is that it is predicated upon the selection of the disclosure risk assessment measure. Hence, while a particular masking method may be the best based on a particular measure of disclosure risk, there is no guarantee that it stays the best for a different measure. This raises the question as to which disclosure risk assessment measure

should be used. Currently, there is no particular risk assessment measure that has been shown to be superior to all other measures.

With the permutation model, the data administrator is free to choose any combination of masking method and disclosure risk measure. The permutation model simply allows the data administrator to set a benchmark in terms of ranks to compare their performance.

4.3 Auxiliary information

Assessing the auxiliary information available to the adversary is felt as necessary for the data administrator to provide adequate protection. It is argued that an adversary with a very high level of knowledge, using auxiliary information, would be able to compromise the data set. This is an issue that needs to be addressed when comparing the performance of masking methods. The purely malicious adversary we assume in our permutation model eliminates auxiliary information as an issue to evaluate masking mechanisms. Given that the adversary is assumed to have knowledge of the entire data set, no auxiliary information could enable the adversary to improve his predictive ability regarding the linkage between the original and masked data records.

5 An empirical illustration

In this section, we provide a simple empirical illustration of the application of the permutation model to compare masking mechanisms. For this purpose, we choose a data set with three different attributes and 1000 observations. The characteristics of the attributes are provided in Table 2.

Table 2. Statistical characteristics of the attributes in the data set used in the empirical illustration

Attribute	Mean	Variance	Minimum	Maximum
1	1	1	0	10
2	1000	25	980	1020
3	50	25	30	70

We use two different approaches for masking the data: (1) additive noise variance with mean 0 and variance equal to 10% of the variance of the data, generated from a normal distribution, and (2) multiplicative method with the multiplicative value generated from a uniform distribution in the range (0.90, 1.10) (that is, the masked value is between 90% and 110% of the original value). The original data were masked using each of these approaches. The masked data were reverse-mapped and the level of permutation was computed.

5.1 Protection offered by multiplicative noise

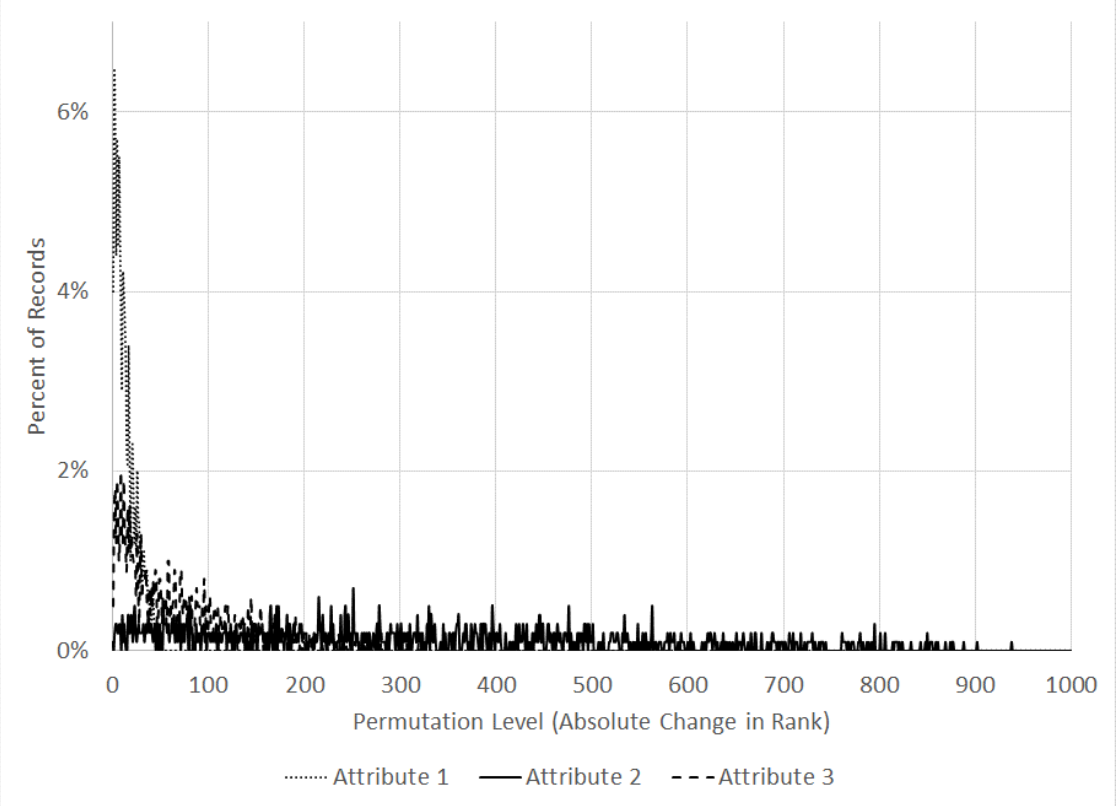

Fig. 1. Permutation level achieved by multiplicative noise for the three attributes

Figure 1 displays the permutation level achieved by multiplicative noise for the three attributes. Since the masking parameter was held constant at 10% (the masked value being in the range 90%-110% of the original value) across all three attributes, we would expect that all three attributes would be masked to the same level. But Figure 1 shows that this is not the case. Attribute 1 has a relatively small magnitude (ranging between 0 and 20). Consequently, the masked values are relatively unchanged by the multiplicative method. Hence, the level of permutation is very small and no record changes in rank by more than 50. By contrast, attribute 2 has a relatively large magnitude and small range. This results in very large change in the ranks of attribute 2 as shown in Figure 1. The change in rank for some records is as large as 900. The results for attribute 3 is different from that of either attribute 1 or 2. In terms of permutation level, attribute 3 falls between attributes 1 and 2 since the data magnitude of attribute 3 falls in between the magnitudes of attributes 1 and 2. Thus, relying on the masking parameter in this scenario would result in wildly varying levels of protection for the three attributes. Using the permutation model allows the data administrator to select the level of masking by comparing the permutation level across the different attributes. The permutation levels for all three levels of the additive method can be easily compared, as in Figure 1.

5.2 Comparing methods

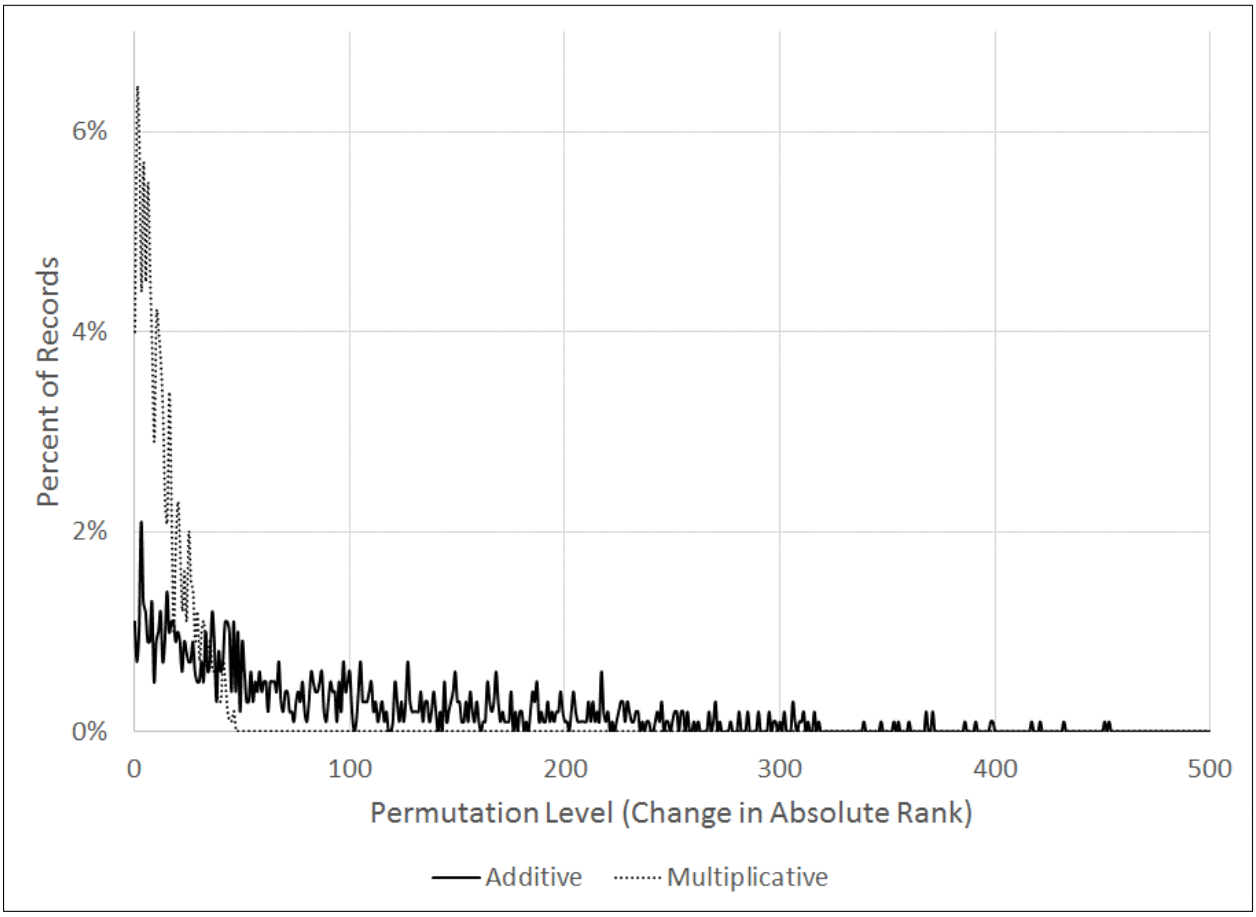

Fig. 2. Permutation level achieved by additive and multiplicative noise for attribute 1

Figure 2 gives a comparison of the permutation level by the additive and multiplicative methods for attribute 1. The figure indicates that the level of permutation for the additive method (with change in absolute rank ranging from 0 to 450) is much higher than the one for the multiplicative method (with change in absolute rank ranging from 0 to 50). As discussed earlier, the magnitude of the values of attribute 1 is relatively small, which results in small changes due to the multiplicative method. With the additive method, the level of noise is independent of the magnitude of the values and consequently the permutation level is much higher. In order to effectively compare the two methods, the data administrator must modify the masking parameters of one of the methods so that the permutation levels of both methods are comparable. In order for the multiplicative method to be comparable to the additive method in terms of utility preservation, it would be necessary to specify a much larger multiplicative factor than the current range of 90%-110%.

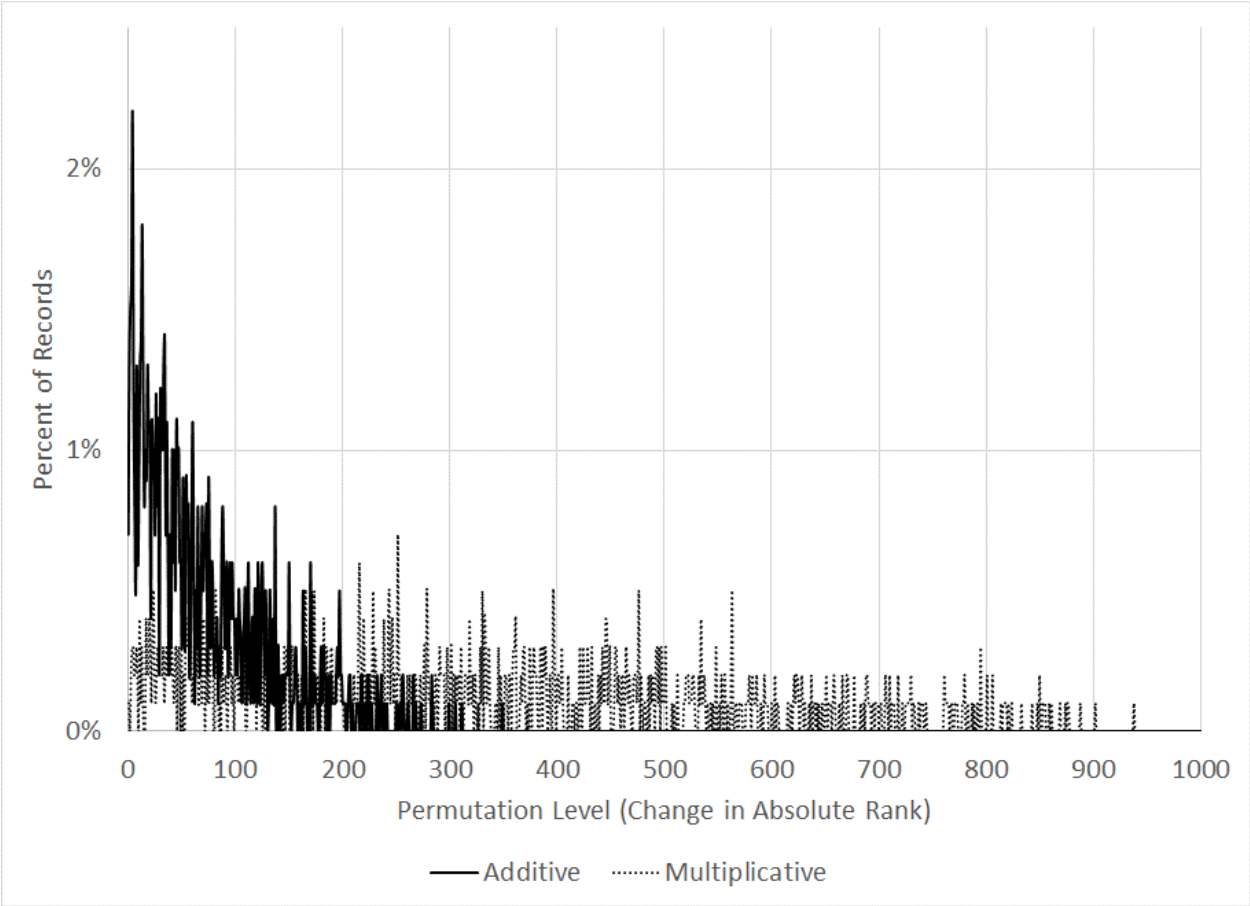

Fig. 3. Permutation level achieved by additive and multiplicative noise for attribute 2

Figure 3 shows the same comparison for attribute 2. Not surprisingly, the situation is the opposite of the one observed for attribute 1: here, multiplicative noise results in a much higher level of permutation than the additive method. In this case, it would be necessary for the data administrator to reduce the masking level of the multiplicative method in order to make it comparable to the additive method in terms of utility preservation.

6 Conclusions

The objective of this study was to describe a general model of microdata masking that is capable of describing the wide variety of methods that are available in the literature. The permutation model allows the data administrator to perform meaningful comparisons across different methods independently of their nature, their parameters, the distributional characteristics of the data, and/or the choice of the disclosure risk measure.

Acknowledgments

The following funding sources are gratefully acknowledged: Government of Catalonia (ICREA Acadèmia Prize to the second author and grant 2014 SGR 537), Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY"), European Commission (project H2020 RIA-644024 "CLARUS"), Templeton World Charity

Foundation (grant TWCF0095/AB60 “CO-UTILITY”). The second 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 or the Templeton World Charity Foundation.

References

Adam, N. R. & Wortmann, J.C. (1989) Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4):515-556.

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

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

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

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

Muralidhar, K. & Sarathy, R. (1999). Security of random data perturbation methods. ACM Transactions on Database Systems, 24(4):487-493.

Muralidhar, K., Sarathy, R. & Domingo-Ferrer, J. (2013) Reverse mapping to preserve the marginal distributions of attributes in masked microdata. In *PSD 2014-Privacy in Statistical Databases*, LNCS 8744, Springer, pp. 105-116.

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

Rubin, D. (1993) Discussion on statistical disclosure limitation, *Journal of Official Statistics* 9(2):461-468.

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

Skinner, Chris J. (2008) Assessing disclosure risk for record linkage. In: Domingo-Ferrer, Josep and Saygın, Yücel, (eds.) Privacy in statistical databases: UNESCO chair in data privacy international conference, PSD 2008 Istanbul, Turkey, September 24-26, 2008 proceedings. Lecture notes in computer science,5262 . Springer-Verlag, Berlin, Germany, pp. 166-176.

Sweeney, L. (2002) k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10: 557-570.

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.

Winkler, W.H. (2004) Re-identification Methods for Masked Microdata, US Bureau of the Census (http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf).