Reverse Mapping to Preserve the Marginal Distributions of Attributes in Masked Microdata

Krish Muralidhar[1], Rathindra Sarathy[2], Josep Domingo-Ferrer[3]

[1]Gatton Research Professor, College of Business & Economics, University of Kentucky, Lexington, KY, USA (krishm@uky.edu)
[2]Ardmore Chair and Professor of Information Systems, Spears College of Business, Oklahoma State University, Stillwater, OK, USA (rathin.sarathy@okstate.edu)
[3]UNESCO Chair in Data Privacy, Dept. of Computer Engineering and Maths, Universitat Rovira i Virgili, Tarragona, Catalonia (josep.domingo@urv.cat)

Abstract. In this paper we describe a new procedure that is capable of ensuring that the marginal distributions of attributes in microdata masked with a masking mechanism end up being the same as the marginal distributions of attributes in the original data. We illustrate the application of the new procedure using several commonly used masking mechanisms.

Keywords: Data masking, marginal distribution, reverse mapping

1 Introduction

Releasing masked microdata in place of sensitive data for the purposes of aggregate analysis has received considerable attention in the literature. Hundepool et al [8] provide a comprehensive discussion of the different masking mechanisms that can be used for generating the masked microdata. The masking mechanisms can be classified based on many criteria. For the purposes of this study, we consider in particular classification, namely, whether the marginal distribution of the output (the masked microdata) is exactly the same as the marginal distribution of the input (the original data). Most of the masking mechanisms currently available (such as additive noise, multiplicative noise, and microaggregation) modify the marginal distribution of the attributes they mask. There are only a few mechanisms (data swapping, data shuffling, and log-linear models) capable of preserving the marginal distributions of the masked attributes to be exactly the same as those of the original attributes.

The modification in the marginal distributions of the masked attributes is, in most cases, caused by the masking mechanism modifying the values of attributes in the original data. With some masking procedures, the masked attributes have a completely different marginal distribution compared to the original attributes. Other masking procedures are capable of asymptotically (but not exactly) preserving the marginal distributions of the original attributes. For the purposes of this study, we do not distinguish

between the two, because our objective is that the marginal distribution of the masked data be exactly the same as that of the original data.

The above requirement may seem extremely stringent, but there is a good reason for it. In many cases, users are reluctant to use data that has been modified. When discussing noise addition, a Census Bureau researcher made the observation that "users have found this extremely irritating and unacceptable" [14]. Raghunathan et al [12] describe the reaction of the users in the following manner: "Could we seriously propose spending time analyzing completely 'fake' data?" Privacy experts often argue that such users are misinformed and that the data is not fake; it has been modified to protect the sensitive information and preserves many of the statistical properties of the original data. With very knowledgeable users, this argument is convincing. However, for a vast majority of the users, this argument carries little weight. In their opinion, modifying the data is the equivalent of fake data. It is of little consequence whether the users are right or wrong. If users perceive the modified data as fake and are reluctant to use it, then the very purpose of providing masked microdata becomes pointless.

Data administrators have the option to not inform users that the data has been modified. This may not be possible if the data administrator is required to provide users with information about the fact that the released data is not the true data. This is particularly true for government agencies that release data; they may be required, legally, to inform users that the data has been masked even if they do not provide the details of the masking. Even if not telling users that the data has been masked was allowed, it presents several practical problems. With many masking mechanisms, in order to reach valid inferences using the masked data, it may be necessary to perform additional processing to compensate the bias of the results obtained from analyzing the masked data as if it were the original data [12]. If the users are not informed about the modification, they will not be able to reach proper statistical inferences. In addition, there is a real danger that the fact that the data has been modified becomes known, hence resulting in users distrusting all data from this source.

It is interesting to note that users do not react in the same manner to all masking mechanisms. Government agencies often use data swapping since it is often more acceptable to users, even though data swapping is perturbative (that is, it involves the modification of the original values of a particular record) [4]. However, while data swapping offers the ability to maintain the marginal distribution, it may not satisfy other analytical validity and/or disclosure risk requirements [5, 9, 10]. Hence, there is a need to develop new procedures that are capable of maintaining the marginal distribution while providing the data administrator with the flexibility to choose the appropriate masking mechanism. The objective of this study is to describe a simple, general post-masking procedure that maps the masked values to a permutation of the original values, regardless of the underlying masking mechanism adopted.

2 Data Swapping

Data swapping involves exchanging values of the original attribute between records that lie within the proximity specified by the data administrator [9, 10, 11]. For numerical attributes, the proximity is usually represented by the rank of the value for a particular record. Experience seems to suggest that data swapping is more palatable to

users than other forms of masking. So how can we explain the difference in the responses to data swapping and other methods?

We believe that the primary reason is that even though the attribute values of records have been exchanged, the actual values themselves are unmodified. In other words, data swapping does not involve the modifying the actual values that existed in the original data. Hence, the collection of values of each original attribute remains unmodified, even it may have been permuted among records, that is, data swapping ensures that the marginal distribution of the masked attribute is exactly the same as that of the original attribute. It also allows the data administrator to provide a simple explanation of the masking mechanism as follows:

> In order to protect privacy, the values of some records may have been exchanged with other records.

This is a simple explanation that is understood by most users, which is in contrast to other procedures. We believe that these are keys to user acceptance of data swapping and reluctance (and in some cases, outright rejection) of other masking mechanisms.

Based on the above argument, the solution seems relatively straightforward and simple … always use data swapping. It is true that this is likely to result in high user acceptance and probably explains the popularity of data swapping as a masking mechanism. This is an acceptable solution if user acceptance is the only criterion. This is not the case. When releasing microdata, data administrators have to consider multiple criteria. In addition to user acceptance, data administrators must also ensure that the released masked microdata provides high analytical validity and low disclosure risk.

The relative importance of these three criteria (analytical validity, disclosure risk, and user acceptance) may vary depending on the context. But it is unlikely that analytical validity and disclosure risk can be completely ignored. After all, if disclosure risk is irrelevant, then the data administrator can simply release the original data; and if analytical validity is irrelevant, then the data administrator can simply release randomly sorted data (or encrypted data or no data at all). Either of these situations is extremely unlikely in practice. In most real life situations, the data administrator would have to select a masking mechanism that ensures an adequate level of analytical validity, disclosure risk prevention, and user acceptance.

Unfortunately, while evidence suggests that data swapping ensures user acceptance, it does not always provide the desired levels of analytical validity and disclosure risk prevention. Prior research indicates that data swapping always results in attenuating the relationship between attributes [5, 9, 10]. This attenuation is directly related to the proximity of the swapped data; as the distance between the swapped records increases, the attenuation of the relationship between the attributes also increases. The situation is the opposite if we consider disclosure risk. When the swapped records are in close proximity, there is a very high risk of disclosure. In order to provide adequate disclosure risk prevention, it may be necessary to increase the proximity of the swapped records, which results in poor analytical validity. In evaluating a select group of masking mechanisms, Domingo-Ferrer and Torra [5] found that the information loss characteristics of data swapping fell in the lower half of the 100+ alternative mechanisms evaluated. And in terms of disclosure risk prevention, there is considerable evidence to

suggest data swapping performs poorly [10, 11]. Nin et al [11] also point to the disturbing possibility that complete disclosure of a masked record is also possible, which Domingo-Ferrer and Torra [5] also acknowledge. Thus, while data swapping may have a higher level of user acceptance, compared to other masking mechanisms, it may also have lower analytical validity and/or higher disclosure risk.

The above discussion indicates the need for a masking mechanism that provides superior analytical validity and/or lower disclosure risk but also provides the intuitively appealing characteristic of data swapping: an unmodified marginal distribution. In this study, we describe a simple, general approach by which the values of each attribute of the masked data can be reverse-mapped to a permutation of the original attribute values; in this way, the masked data has the same marginal distributions as the original data. It is important to note that we are not proposing a new masking method. We wish to provide the data administrator with the flexibility to choose a data masking mechanism that suits the particular context.

3 Post-Masking Reverse Mapping

The statement of the problem is straightforward. The data administrator chooses a masking mechanism and uses it to obtain a masked version of the original sensitive data set. Masking can operate on the quasi-identifier attributes to prevent identity disclosure (which would happen if quasi-identifiers allowed an intruder to link a record with an identified record in an external data source), or it can operate on sensitive attributes to prevent attribute disclosure (records linkage is not prevented, but the sensitive attributes are masked) or it can operate on all attributes.

The data administrator now seeks a transformation of the masked data such that each attribute in the transformed masked data (which will be released to the public) has exactly the same marginal distribution as the corresponding attribute in the original data. The data administrator can attempt statistical transformations [13] but such transformations do not guarantee that the transformed data will have exactly the same marginal distributions as the original data.

In this study we propose a simple transformation of the masked attribute values to a permutation of the original attribute values using rank-based reverse mapping. The advantage of this approach is that it maintains the masking (which is necessary to prevent disclosure risk) but also allows preserving the marginal distributions (which facilitates user acceptance). Let $X = \{x_1, x_2, \dots, x_n\}$ be the values taken by attribute X in the original data set. Let $Y = \{y_1, y_2, \dots, y_n\}$ represent the masked version of X. We make no assumptions about the masking mechanism used to generate Y except that it must be possible to link a single record in Y to a single record in X.

The post-masking reverse mapping is performed as follows:

$For\ i\ =\ 1\ to\ n$
$Compute\ j\ =\ Rank(y_i)$
$\quad Set\ z_i\ =\ x_{(j)}\ (where\ x_{(j)}\ is\ the\ value\ of\ X\ of\ rank\ j)$
$Repeat$
$Release\ Z\ =\ \{z_1, z_2, \dots, z_n\}$

It is easy to see that Z has exactly the same values and hence the same marginal distribution as X, but values have been permuted according to the (rank) order of the masked values Y. The extension of the reverse mapping to the multivariate case does not present a problem – simply repeat the process for every masked attribute. Since the objective is to maintain the marginal distribution of the individual attributes, application of the procedure on an attribute-by-attribute basis has no effect on the outcome.

From a statistical perspective, the reverse mapping procedure can be described as the rank based mapping of the cumulative distribution function of the masked values Y back to the cumulative distribution function of the confidential values X. The concept of copulas in statistical analysis is based on a similar transformation [15]. However, copulas are often used to model the joint distribution of attributes with different marginal distributions. The reverse mapping procedure we describe is a univariate one (where each attribute is reverse-mapped individually) and no effort is made to model the joint distribution of the attributes. Hence, the only property that is maintained is that the rank of Z is the same as that of Y and, unlike copulas, no claims regarding the joint distribution of the attributes can be made. The reverse mapping procedure is also similar to the "Normal to Anything" procedure used to generate related random attributes with different marginal distributions [1, 2].

The extent to which the data administrator wishes to provide information additional to Z would depend on the context. The data administrator may also provide information on the process used to mask the data and extent of the changes. The release of this information would depend, however, on the extent to which the release of this information would affect disclosure risk. Note, for example, that if the masking method deterministically generates Y from X, e.g. $Y = f(X)$, and the administrator publishes Y and the (parameterized) masking method $f()$, then any intruder can compute $Y_{permuted} = f(Z)$; since the permutation that transforms $Y_{permuted}$ into Y is the same that transforms Z into X, the intruder can recover the original X (if there are ties in Y, though, the intruder might not be able to unequivocally determine all values $x_1, x_2, \ldots, x_n$).

We now illustrate the application of this procedure using a small data set consisting of 25 observations and two attributes: a non-sensitive attribute S and a sensitive attribute X that is to be masked. As mentioned above, we make no assumptions about the masking mechanism other than that it allows the *data protector* to keep track of which original record corresponds to which masked record.

Table 1 shows that the marginal distribution of Z is the same as that of X, because Z has the same set of values as X, but re-ordered according to the ranks of Y.. As discussed earlier, the specific masking mechanism used to generate Y is irrelevant for the purposes of reverse mapping. Consider the second observation (ID = 2). In this observation, the original attribute X takes value 1122, the masked attribute Y takes value 1023.59, and the reverse mapping yields a value of 1015. The large change between the values of X and Y is reflected in a large change in the corresponding ranks (from 23 in X to 13 in Y). For some observations (ID = 7, 10, 11, 22) the ranks of X and Y are the same.

Table 1. Example illustrating the implementation of reverse mapping

ID	S	Rank of S	X	Rank of X	Y	Rank of Y	Z
1	131	24	1110	22	1190.91	23	1122
2	104	15	1122	23	1023.59	13	1015
3	95	9	912	6	958.10	6	912
4	87	5	1015	13	1110.99	21	1096
5	102	14	1096	21	1074.13	18	1070
6	98	12	1018	14	971.81	10	974
7	90	7	889	3	856.12	3	889
8	97	11	974	10	939.37	5	908
9	108	16	1073	19	1111.88	22	1110
10	119	22	1177	25	1253.93	25	1177
11	100	13	963	9	970.13	9	963
12	109	17	1070	18	1102.79	19	1073
13	96	10	1003	12	1107.92	20	1091
14	89	6	906	4	960.64	7	921
15	110	18	921	7	993.10	12	1003
16	116	21	1144	24	1215.15	24	1144
17	94	8	934	8	974.77	11	979
18	80	3	908	5	881.33	4	906
19	114	20	979	11	969.52	8	934
20	73	2	819	2	786.79	1	780
21	72	1	780	1	817.53	2	819
22	112	19	1035	16	1068.10	16	1035
23	84	4	1032	15	1069.53	17	1065
24	133	25	1065	17	1053.42	15	1032
25	121	23	1091	20	1027.35	14	1018

One potential issue that could pose a problem is observations that have tied ranks, particularly when X is a discrete attribute. Tied ranks could also arise in the masked values Y as a result of the masking mechanism. For example, applying univariate microaggregation or generalization could result in tied masked values. There is a simple solution to this situation. Any ties in the original data and/or the masked data are broken randomly. With large data sets, breaking ties randomly will have little or no impact on the procedure, as long as the number of records sharing a certain value is small compared to the number of records in the data set.

In some cases, releasing a permutation of the original attribute values may still be disclosive: e.g., in a data set on a town's population, releasing the highest value of "Income" reveals the income of the wealthiest person in town, whose identity is easy to guess regardless of the permutation. Hence, statistical agencies may need further measures to prevent disclosure. These include top and bottom coding, rounding, and other similar procedures. These procedures can be implemented on the reverse-mapped attributes exactly as they would for the original attributes.

Finally, the reverse mapping procedure could also be adopted for categorical attributes. Recently, Domingo-Ferrer et al [6] proposed a procedure for numerical mapping of nominal attributes that captures and quantifies their underlying structure (specifically, the hierarchical structure that is often present in nominal classification schemes). This procedure provides the ability to quantify any nominal attribute and, once quantified, the reverse mapping procedure can be applied to such data.

4 Assessing Disclosure Risk and Information Loss

Apart from the fact that the reverse mapping procedure results in the marginal distributions of the original and masked attributes being identical, it also offers another important advantage for data administrators. In Table 1, we have presented the values (Y) resulting from masking the sensitive values (X). Assume that the data administrator wishes to compare the results of this masking procedure with an alternative procedure and assume that the alternative procedure results in a different set of values (Y^*). In most cases, the distribution of Y is different from the distribution of Y^*. This makes evaluation of information loss and disclosure risk difficult, since the data administrator is comparing two different sets of values. However, assume that the values of Y, Y^* have also been reverse-mapped to yield Z, Z^*, respectively. The distributions of Z and Z^* are identical, which allows the data administrator to make a more meaningful comparison of the performance of the two methods.

Consider r_{xz} the rank order correlation between (X, Z). For the purposes of this analysis, and without loss of generality, we will assume r_{xz} to be non-negative and hence $0 \leq r_{xz} \leq 1$. Having $r_{xz} = 0$ implies that Z and X are completely unrelated. For large data sets, the situation $r_{xz} = 0$ can be reached by randomly sorting the values in X to obtain Z. This setting offers the highest level of protection against disclosure risk. Similarly, $r_{xz} = 1$ implies that knowledge of Z results in complete disclosure of X. Thus, r_{xz} represents a simple measure of disclosure risk – the higher the value of r_{xz}, the greater the disclosure risk. It is important to note that the rank order correlation is a superior measure compared to product moment correlation since rank order correlation measures all monotonic relationships (including linear ones), but product moment correlation measures only linear relationships.

Similar to disclosure risk, we can assess the information loss in the relationship between attributes by considering the rank order correlation of the attributes in the released data set compared to the rank order correlation in the original data set. For the illustration provided in Table 1, let r_{sx} and r_{sz} represent the rank order correlation between (S, X) and (S, Z), respectively. The difference between the two rank order correlations represents a simple measure of the information loss resulting from masking the data. When there are multiple non-masked and/or masked attributes in the released data set, the data administrator will consider the pair-wise rank order correlation (1) between all pairs composed of non-masked and a masked attribute, and (2) between all pairs of masked attributes. Such correlations will be compared with the rank order correlations between the corresponding pairs of attributes in the original data set. This provides a simple versatile approach for assessing the information loss in the relationship among attributes.

It should be noted that we do not preclude the use of alternative measures of disclosure risk and information loss measures. Since the reverse-mapped values are the same across all mechanisms, we are comparing "apples versus apples" rather than "apples versus pears."

5 Comparison of Masking Mechanisms

In this section, we apply reverse mapping to the output of several masking mechanisms to illustrate the procedure as well as to provide a simple comparison of releasing the masked data versus releasing data that is reverse-mapped after masking, using the data in Table 1. We consider four alternative approaches for masking the data: (1) Additive noise (ADD) with noise from a normal distribution with mean zero and variance equal to 25% of the variance of X, (2) A single imputation (IMP) using the partially synthetic data procedure suggested by Drechsler et al. [7], generating the masked values Y using the posterior predictive distribution $f(X|S)$, (3) Microaggregation (MIC) with k = 5, and (4) Data swapping (SWP) with rank proximity = 5.

Table 2 provides the original data, reverse-mapped data for all four methods, and the masked data Y for the first three masking procedures (since swapping directly results in Z). The idea of masking mechanisms providing different permutations of the original values is illustrated by the X column and the Z columns of the four different masking mechanisms. The only difference between the four Z columns is the way values are permuted. Evaluation of the masking mechanism can be performed (as we illustrate below) by evaluating only the output (Z) and without any consideration of the parameters of the masking mechanism.

Prior to discussing the performance characteristics of the different masking mechanisms, we briefly address the results of the reverse-mapped microaggregation procedure. In the traditional univariate microaggregation, the values of all records within a single aggregated group are set to the mean of their values. For instance, records {1, 2, 5, 10, 16} form one aggregated group. The mean value of X in this group is 1129.80 and hence the value of all five of these records is set to 1129.80 (see last column MIC Y in Table 2). This presents a problem since there are only five unique values in the entire data set instead of the original 25. This aggregation also results in variance attenuation, that is, the variance of Y (9246.06) is much smaller than the variance of X (10223.51).

When reverse mapping is performed, ties are broken randomly, and the values of X are reverse-mapped based on the values of Y. This is the equivalent of *randomly permuting* the values within each of the aggregated groups and releasing the result. The reverse mapping eliminates the presence of k records with the same values and the subsequent variance attenuation; in fact, it is an alternative to other variance restoration approaches for microaggregation, like [3] (which uses synthetic values rather than a permutation of original values). We believe that the reverse mapping procedure is consistent with the original intent of microaggregation – that values within the aggregated groups should be indistinguishable. Since the values within any aggregated group are completely random, the reverse mapping procedure achieves this objective. This is consistent with the notion of probabilistic k-anonymity [16], which can be attained when multivariate microaggregation with group size k is applied to the set of quasi-identifier

attributes and reverse mapping is then used on the microaggregated values of each quasi-identifier attribute.

Table 2. Results of masking and reverse mapping for all masking mechanisms

ID	S	X	ADD Z	IMP Z	MIC Z	SWP Z	ADD Y	IMP Y	MIC Y
1	131	1110	1073	1177	1122	1073	1074.11	1251.88	1129.80
2	104	1122	1096	921	1144	1096	1086.02	935.33	1129.80
3	95	912	934	1018	963	974	953.62	1005.54	940.80
4	87	1015	921	974	1032	1065	940.20	970.86	1009.40
5	102	1096	1070	1091	1177	1122	1051.91	1033.74	1129.80
6	98	1018	1018	889	1018	979	1027.87	879.13	1009.40
7	90	889	819	934	819	921	863.64	940.48	860.40
8	97	974	1015	908	912	912	1018.39	886.67	940.80
9	108	1073	1144	1070	1091	1110	1119.18	1025.96	1066.80
10	119	1177	1177	1110	1096	1091	1174.44	1071.50	1129.80
11	100	963	1032	1096	974	934	1031.09	1035.20	940.80
12	109	1070	1122	1003	1073	1032	1105.33	984.95	1066.80
13	96	1003	963	963	1003	1035	976.26	944.64	1009.40
14	89	906	889	1035	889	819	875.42	1015.18	860.40
15	110	921	908	1015	921	889	908.60	986.23	940.80
16	116	1144	1091	1073	1110	1144	1081.92	1031.50	1129.80
17	94	934	979	780	934	963	991.27	789.44	940.80
18	80	908	906	906	780	780	886.03	882.32	860.40
19	114	979	1035	1065	979	1018	1031.90	1019.49	1009.40
20	73	819	912	912	906	906	928.69	930.28	860.40
21	72	780	780	819	908	908	688.71	826.58	860.40
22	112	1035	1110	979	1070	1003	1103.82	978.10	1066.80
23	84	1032	974	1032	1015	1070	982.19	1007.62	1009.40
24	133	1065	1065	1144	1035	1015	1045.98	1214.20	1066.80
25	121	1091	1003	1122	1065	1177	1017.15	1104.29	1066.80

Table 3. Disclosure risk measures for the four masking mechanisms

	ADD Z	IMP Z	MIC Z	SWP Z
Rank order correlation between X and	0.860	0.577	0.952	0.892

Table 3 provides a disclosure risk assessment based on the correlation between X and Z for each of the masking methods. Of the four mechanisms, multiple imputation has the lowest correlation with the original attribute and hence provides the lowest disclosure risk. Microaggregated values have the highest correlation with the original attribute and hence, the highest disclosure risk (this is consistent with Domingo-Ferrer and Torra pointing out in [5] that *univariate* microaggregation offers little disclosure protection). Additive noise and swapping have lower correlation (and lower disclosure risk) than microaggregation. The disclosure risk of the imputed sample is significantly lower than that of the other three methods. It is important to note that we made no effort to fine tune the parameters of the masking mechanisms so that they provide comparable disclosure risk. The information in Table 3 allows the data administrator to perform such an analysis (such as increasing the value of k so that the disclosure risk resulting from microaggregation is comparable to the others).

Table 4. Information loss for the four masking mechanisms

	X	ADD Z	IMP Z	MIC Z	SWP Z
Rank order correlation between S and	0.742	0.738	0.710	0.682	0.568

Table 4 provides an assessment of information loss, the rank order correlation between the non-sensitive attribute S and X compared to that between S and each of the masked data. Of the four approaches, additive noise is best at preserving the correlation with the non-masked attribute followed by imputation, microaggregation, and swapping. It is important to note that this is only an illustration and it is possible that a different set of pseudorandom numbers could produce different results. The performance of swapping is rather poor – it results in the highest information loss and also results in high disclosure risk. By contrast, multiple imputation easily yields the best performance – extremely low disclosure risk *and* information loss, which is essentially guaranteed by the underlying model used to generate the masked values. Comparing the results of the masked values and the reverse-mapped values indicates that the reverse mapping process does not have a meaningful impact on information loss. Given that this is a small data set, this result is encouraging.

Table 5 presents the actual change in ranks between the original and masked data. As expected, for both swapping (rank proximity = 5) and microaggregation (k = 5), the change in rank is less than or equal to 5. Multiple imputation results in the highest change in ranks – in one case as high as 16 (record ID = 2). And only imputation has (four) records whose change in ranks is higher than 10. But the results for imputation should not be surprising; the specific purpose of generating the masked value from the posterior predictive distribution is to minimize disclosure risk. The large change in the ranks ensures low disclosure risk. Finally, the objective of this analysis is not to highlight the merits or any particular masking mechanism, but to show that the reverse mapping procedure performs effectively and facilitates an easy comparison of the masking mechanisms.

Table 5. Change in rank for each of the masking mechanisms

Change In Rank	ADD	IMP	MIC	SWP
0	4	1	5	1
1	3	4	8	2
2	3	0	5	4
3	5	5	2	6
4	2	1	5	10
5	4	4	0	2
6 – 10	4	6	0	0
11 – 15	0	3	0	0
16 – 20	0	1	0	0

6 Conclusions and Future Work

The objective of this study was to investigate a new procedure that ensures that the marginal distributions of the masked attributes are the same as those of the corresponding original attributes. We present a post-masking method based on mapping the cumulative distribution function of each masked attribute back to the cumulative distribution function of the corresponding original attribute based on ranks. We refer to this post-masking method as reverse mapping and we have illustrated it using simple examples.

Reverse mapping also allows viewing microdata masking mechanisms in a new framework, namely, one in which the output from *any* masking mechanism applied to an original attribute is simply regarded as a particular permutation of the values of the original attribute. By providing a common ground for all masking mechanisms, we improve the data administrator's ability to perform more meaningful comparisons of masking mechanisms. We also hope that this framework will allow researchers to find potential links between masking mechanisms that may have been considered disparate thus far.

The results presented in this study are illustrative and preliminary. A comprehensive investigation of the impact of reverse mapping on masking is currently being conducted by these authors. Specifically, there are many issues that require further investigation, including : (1) Impact of the size of the data set; (2) Impact of the characteristics of the data set; (3) Incremental disclosure risk (if any) resulting from the reverse mapping; (4) More extensive general analysis of disclosure risk and information loss; (5) Use of reverse mapping with masking methods (e.g. synthetic methods) in which the number of output masked values differs from the number of input original values; and (6) Impact of reverse mapping on privacy models, that is, what privacy models can be satisfied by reverse-mapped masked data (we have already indicated that reverse-mapped microaggregated data can satisfy probabilistic *k*-anonymity).

References

1. Cario, M. C., and B. L. Nelson. 1997. *Modeling and Generating Random Vectors with arbitrary Marginal Distributions and Correlation Matrix*. Evanston, IL: Department of Industrial Engineering and Management Sciences, Northwestern University.
2. Chen, H. 2001. "Initialization for NORTA: Generation of Random Vectors with Specified Marginals and Correlations." *INFORMS Journal on Computing* 13:312-331.
3. Domingo-Ferrer, J. 2009. "Non-Perturbative Masking Methods." In *Encyclopedia of Database Systems*, by L. Liu and M.T. Ozsu, 1912-1913. US: Springer.
4. Domingo-Ferrer, J and U. González-Nicolás. 2010. "Hybrid Microdata Using Microaggregation." *Information Sciences* 180:2834-2844.
5. Domingo-Ferrer, J., and V. Torra. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, by P. Doyle, J.I. Lane, J.J.M. Theeuwes and L.V. Zayatz, 111-133. Amsterdam: Elsevier.
6. Domingo-Ferrer, J., D. Sanchez, and G. Rufian-Torrell. 2013. "Anonymization of Nominal Data Based on Semantic Marginality." *Information Sciences* 242:35-48.
7. Drechsler, J., S. Bender, and S. Rassler. 2008. "Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel." *Transactions on Data Privacy* 1:105-130.
8. Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer, and P-P. de Wolf. 2012. *Statistical Disclosure Control.* West Sussex, United Kingdom: John Wiley & Sons.
9. Moore, R.A. 1996. *Controlled Data Swapping for Masking Public Use Microdata Sets.* Washington DC: Research report series (RR96/04), Statistical Research Division, US Census Bureau.
10. Muralidhar, K, and R. Sarathy. 2006. "Data Shuffling: A New Masking Approach for Numerical Data." *Management Science* 52:658-670.
11. Nin, J., J. Herranz, and V. Torra. 2008. "Rethinking Rank Swapping to Decrease Disclosure Risk." *Data & Knowledge Engineering* 64:346-364.
12. Raghunathan, T.E., J.P Reiter, and D.B. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19:1-16.
13. Sebé, F., J. Domingo-Ferrer, J.M. Mateo-Sanz, and V. Torra. 2002. "Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata." In *Inference Control in Statistical Databases*, by J. (Ed) Domingo-Ferrer, 163-171. Berlin Heidelberg: Springer Verlag.
14. Simpson, G.R. 2001. "The 2000 Count: Bureau Blurs Data to Keep Names Confidential." *Wall Street Journal*, February 14: B1-B2.
15. Sklar, A. 1959. "Fonctions de répartition à n dimensions et leurs marges." *Publications de l'Institut de Statistique de L'Université de Paris* 8:229-231.
16. Soria-Comas, J., and J. Domingo-Ferrer. 2012. "Probabilistic k-anonymity through microaggregation and data swapping." *IEEE International Conference on Fuzzy Systems.* Brisbane, Australia: IEEE. 1-8.