On the Effects of Data Protection on Multi-database Data-Driven Models

Lili Jiang and Vicenç Torra

Department of Computing Sciences, Umeå University, Umeå, Sweden
ljiang@cs.umu.se, vtorra@ieee.org

Abstract. This paper analyses the effects of masking mechanism for privacy preservation in data-driven models (regression) with respect to database integration. Especially two data masking methods (microaggregation and rank swapping) are applied on two public datasets to evaluate the linear regression model in terms of privacy protection and prediction performance. Our preliminary experimental results show that both methods achieve a good trade-off of privacy protection and information loss. We also show that for some experiments although data integration produces some incorrect links, the linear regression model is still comparable, with respect to prediction error, to the one inferred from the original data.

Keywords: Data protection · Masking methods · Reidentification · Microaggregation · Rank swapping · Multidatabase integration

1 Introduction

Data protection mechanisms for databases are usually implemented by means of applying a distortion into the database. Masking methods are the mechanisms to produce such distortion. In short, given a database X , a masking method ρ produces $X' = \rho(X)$ that is a sanitized version of X . This X' corresponds to a distorted version of X so that the sensitive information in X cannot be inferred, and at the same time the analysis we obtain from X' are similar to those we obtain from X .

Privacy models are computational definitions of privacy. Different privacy models exist taking into account the type of object being released, the type of disclosure under consideration, etc. Differential privacy [3], k -anonymity [10, 11], privacy from reidentification [6, 14] are some of these privacy models. When we are considering a database release (database publishing, database sanitization), k -anonymity and privacy from reidentification are two of the main models. They focus on identity disclosure. That is, we intend to avoid that intruders find a particular person in a database. Then, a database is safe against reidentification when it is not possible (or only possible to some extent) to identify a person in the published database. k -Anonymity has a similar purpose. That is, avoid

reidentification and finding particular individual's data in a database. Nevertheless, the definition is different. k -Anonymity [10] requires that for each record in the database there are at least $k-1$ other records that are indistinguishable to it. In this way, there will be always confusion on which was the right link. Then, the goal of a masking method ρ is to produce $X' = \rho(X)$ that is compliant with one of these definitions.

Differential privacy is an alternative privacy model that focuses on the inferences from queries or functions when applied to a database. Then, we have $y = f(X)$ and we have that y satisfies differential privacy when an addition or a deletion of a record from X will not change much the result y . Local differential privacy is a variation of differential privacy that applies to individual records. Differential privacy can tackle some vulnerabilities from k -Anonymity when sensitive values in an equivalence class lack diversity or the intruder has background knowledge.

Since 2000, a significant amount of research has been done in the field of data privacy [5, 13] about methods for databases. Some methods exist that provide a good trade-off between privacy and information loss. That is, research has been done to find methods that distort the database enough to avoid disclosure in some extent and at the same time keeping some of the interesting properties of the data for potential future usage. Interesting properties includes some statistics but also building models through machine and statistical learning.

Nowadays, there is increasing interest in database integration in order to build data-driven models. That is, for applying machine and statistical learning to large datasets in terms of both number of records and number of variables. Further virtual data integration (data federation) [17, 18] have been explored, where data is accessed and virtually integrated in real-time across distributed data sources without copying or otherwise moving data from its system of records. The effects of masking into data integration is not well known. It is understood that masking will modify a database in a way that linkage between databases will not be possible. In contrast, masking has been proven not to be always a big obstacle for the correct application of machine learning algorithms. There are results that show that for some databases, masked data is still useful to build data-driven models.

In this paper we present a preliminary work on the analysis of the effects of masking with respect to database integration. We analyse the effects of two data masking strategies on databases. We show that while the number of correct linkages between two masked databases drop very quickly with respect to the amount of protection, the quality of data-driven models does not degrade so quickly.

The structure of this paper is as follows. In Sect. 2 we review some concepts that are needed in the rest of this work. In Sect. 3 we introduce our approach and in Sect. 4 we present our results. The paper finishes with some conclusions and directions for future research.

2 Preliminaries

In this paper we will consider two data protection mechanisms: microaggregation and rank swapping. They both permit to transform a database X into a database X' with some level of protection. Here, protection is against reidentification. We have selected these two masking methods because microaggregation and rank swapping have been proved to be two of the most effective masking methods against reidentification. See e.g. [1, 2].

Microaggregation consists of building small clusters with the original datafile, compute the centroids of these clusters, and then replace the original data by the corresponding centroids. The clusters are all enforced to have at least k clusters. The number of records k is the privacy level. Small clusters represent a small perturbation, while large k imply large privacy guarantees.

When a database contains several attributes, and all these attributes are microaggregated together, the final file satisfies k -anonymity. Recall that a file satisfies k -anonymity when for each record there are $k - 1$ other records with the same value. This will be the case of a microaggregated file when masking all the files at the same time.

There is a polynomial algorithm for microaggregation when we consider a single attribute [4]. Nevertheless, when more attributes are microaggregated together heuristic algorithms are used, as the problem is NP. See [9]. In this work we have used MDAV, one of these heuristic algorithms, and, more particularly, the implementation provided by the `sdMicro` package in R. See [12] for details.

Rank swapping is a masking method that is applied attribute wise. For a given attribute, a value is swapped by another one also present in the file that is within a range. For example, consider that for an attribute V_1 we have in the file to be masked the following values (1, 2, 4, 7, 9, 11, 22, 23, 34, 37). Then, if we consider a parameter $s = 2$ we can swap a value for another value situated either to two positions in the right or two on the left. For example, we can swap 9 with 4, 7, 11, or 22. Only one swap is allowed for each value.

Instead of giving an absolute number of positions (as $s = 2$ above) we may consider giving a percentage of positions in the file (say p).

The larger the p , the larger the distortion, and, thus, the larger the protection. In contrast, the smaller the p , the smaller the distortion, and, thus, the better the quality of the data, but also the larger the risk. Here risk is understood as a identity disclosure risk. In other words, we use the risk of reidentification.

In this paper we will not go into the details of computing disclosure risk for the files protected. See e.g. [7, 8], for discussions on risk for microaggregation and rank swapping.

3 Evaluating Data-Driven Models with Data-Integration

Our approach to evaluate the effects of data protection for data-driven models when they are applied to data-integration consists of the following steps.

- Partition a database DB_0 horizontally in test and training. Let DB be the training part. Let DB_t be the testing part.
- Take the database DB and partition it vertically into two databases DB_1 and DB_2 sharing some attributes. Let nC be the number of attributes that both databases share.
- Mask independently using a masking method ρ the two databases DB_1 and DB_2 producing $DB'_1 = \rho(DB_1)$ and $DB'_2 = \rho(DB_2)$.
- Integrate DB'_1 and DB'_2 using the nC common attributes. Let DB' be the resulting database. That is, $DB' = \text{integrate}(DB'_1, DB'_2)$ where *integrate* is an integration mechanism for databases.
- Compute a data-driven model for DB and the same data-driven model for DB' . Let us call them $m(DB)$ and $m(DB')$.
- Evaluate the models $m(DB)$ and $m(DB')$ using the test database DB_t .

In order to make this process concrete, some steps need further explanation. We will describe them below.

Database integration has been done applying distance-based record linkage. That is, for each record r_1 in DB'_1 we compute the distance to each record r_2 in DB'_2 and we select the most similar one. That is $r'(r_1) = \arg \min_{r_2 \in DB'_2} d(r_1, r_2)$. We use an Euclidean distance for d that compares the common attributes in both databases DB'_1 and DB'_2 .

As DB'_1 and DB'_2 both proceed from the same database DB through its partition and the process of masking the two parts, we can evaluate in what extent the database is correctly integrated. That is, we can count how many times $r'(r_1)$ is the correct link in DB'_2 for r_1 . As we will discuss in Sect. 4 the number of correct links drops very quickly with respect to the data protection level. That is, most of the links are incorrect even with a low protection.

The other steps that need to be described are the masking methods, and the computation of the model. In relation to masking, we apply the same method to both DB_1 and DB_2 , and the methods are microaggregation (using MDAV) and rank swapping. We have explained these methods in Sect. 2. Then, in order to build the data-driven model we use a simple linear regression model. Comparison of the model is based on their prediction quality. More particularly, we compute the sum of squared errors of both $m(DB)$ and $m(DB')$ and compare them. The comparison is possible because we have saved some records of the original database for testing. DB_t has been kept unused from the original database DB_0 .

4 Experiments and Results

In this section we first present the experiments performed, describing the datasets and giving details so that the experiments can be reproduced by those interested. Therefore, the description includes attributes used in each step, as well as the parameters of the masking methods. Then, we describe the results obtained.

4.1 Setting

We have applied our approach to two different datasets. They are

- CASC: This dataset consists of 1080 records and 13 numerical attributes. It has been extensively used to evaluate masking methods in data privacy. It was created in the EU project CASC, and it is provided by the `sdcMicro` package in R. We have used the version supplied by this `sdcMicro` package. See e.g. [5] for a description and for other uses of this dataset.
- Concrete Compressive Strength. This is a dataset consisting of 1030 records and 9 numerical attributes. It is provided by the UCI repository. We have selected this dataset because it is of small size, all data is numerical and it has been used in several works to study and compare several regression models, including linear regression. See e.g. [15, 16].

The first step consists of partitioning the database into test and training sets. We have used 80% records for training and 20% for testing.

The second step is about the vertical partitioning of the databases. This is about selecting some attributes for building the first database DB_1 and some attributes for building the second one DB_2 .

The attributes in the CASC file are AFNLWGT, AGI, EMCONTRB, FED-TAX, PTOTVAL, STATETAX, TAXINC, POTHVAL, INTVAL, PEARVAL, FICA, WSALVAL, ERNVAL. We have splitted them in two databases considering $nC = 1, 2, 3, 4, 5, 6, 7, 8$ different sets of common variables. This corresponds to eight different pairs of databases. For the first pair, DB_1 includes attributes 1–7 in the list above, and DB_2 includes attributes 7–13 in the list above. The following pairs have been built as follows. The first databases include attributes 1–7, 1–8, 1–8, 1–9, 1–9, 1–10, 1–10, respectively. The second databases include, respectively, attributes 6–13, 6–13, 5–13, 5–13, 4–13, 4–13, 3–13.

The file related to the concrete problem has been used to generate 4 pairs of databases. The process of partitioning is similar to the case of the CASC file. In this case, we have $nC = 1, 2, 3, 4$ common attributes. The first databases consist of attributes 1–5, 1–6, 1–6, 1–7, respectively. The second databases consist of attributes 5–9, 5–9, 5–9, 5–9. For example, the first pair of databases DB_1 and DB_2 will be defined as follows: DB_1 contains the first 5 attributes, and DB_2 contains from attribute 5 to attribute 9.

Once data is partitioned, we have protected the two resulting databases using two different masking methods, each one with different parameterizations. We have used microaggregation (using MDAV as the microaggregation algorithm) and considering $k = 3, 4, 5, 6, 8, 10, 12, 15$, and rank swapping considering $p = 0.001, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1$. In both cases, the larger the parameter, the larger the distortion. We have used the R package `sdcMicro` to protect the datasets.

As we have explained above, we have applied distance-based record linkage for database integration. We have used the Euclidean distance using the common attributes. We have used our own implementation of distance-based record linkage.

Finally, we have computed a model for both the original training dataset DB and the masked and integrated database DB' . We have used a linear model for one of the attributes. All other attributes were used as independent variables of the model. For the CASC dataset we have used the first variable as the dependent variable. For the Concrete dataset, we have used the last variable in the file as the dependent variable. In this second dataset, this variable is the one used as dependent in previous research. We have used the function lm in R for building the models.

In order to analyse the results of the models, we compute the sum of squared errors between the prediction of the linear model and the true values. To compute this error we use the test set that, as we have explained above, consists of 20% of the records of the original files.

In addition, in order to evaluate in what extent the database integration is good, once databases DB_1 and DB_2 have been masked into DB'_1 and DB'_2 , we count the number of records that are correctly linked when we build DB' from them.

It is relevant to note that our approach contains some steps based on randomization. In particular, the partition of the original databases into training and testing is based on a uniform distribution. Then, rank swapping also uses a random element to determine how elements are swapped. Because of that, for each combination (partition, masking method, parameter) we have applied our approach 5 times and study the averages of these executions.

4.2 Results

On important element to take into account in our setting is the integration of the two databases DB'_1 and DB'_2 and whether the records in one database are correctly linked with the other database. Our experiments show that the number of correct links drops when the protection increases. More particularly, the number of correct links becomes very small very quickly except for the case of CASC dataset masked using rank swapping, where a larger distortion is needed for the same effect when the number of attributes is relatively large (when we are using more than 4 attributes in the linkage). This reduction on the number of correct links depends on the number of common attributes nC . The larger the number of common attributes, the larger the reidentifications. In Fig. 1, we display the number of correct links for the Concrete dataset (top figures) and CASC dataset (bottom figures), when the number of common attributes nC range from 1 to 6 and protection increases for both microaggregation and rank swapping (k or p , respectively, as above). The figures also include the number of correct links in the case of no protection at all. This is the first point on the left of each figure.

We can observe in the figures that the number of correct links when data is masked using rank swapping is always larger than the number of correct links when data is masked using microaggregation. This is so because in microaggregation we are masking all attributes at the same time. This produces a file that satisfies k -anonymity for a given k , and, thus, probability of correct linkage

Fig. 1. Number of correct links for the Concrete dataset (top) and the CASC dataset (bottom) when data is protected using microaggregation (left) and rank swapping (right). Number of correct links decrease when protection is increased (i.e., k in microaggregation or p in rank swapping increase). Different curves correspond to different number of attributes in the reidentification (circles mean only one variable in reidentification). Number of links increase when the number of attributes increase (from 1 to 6 or 8).

becomes $1/k$. In contrast, in rank swapping each attribute is masked independently. When several attributes are considered in the linkage, noise of different attributes are independent, and then some records may have a larger probability of being correctly linked.

Figure 2 represents the mean squared error for the concrete dataset for the last attribute. Figures represent mean values after 10 runs. Top figures correspond to data protected using rank swapping and bottom figures correspond to data protected using microaggregation. From right to left (and then top to bottom) we have different micro number of attributes (from one to six). We can see that there is a trend of increasing error when we increase protection, and that the error is somehow smaller when the number of attributes used in the linkage increase.

Nevertheless if we compare these results about the error with the ones related to the number of correct links, we see that even the number of correct links can be very low, the error is not increasing so fast. We observe that there is still some quality in the models even when the files are not linked correctly.

Fig. 2. Error of the models (mean squared error) for the Concrete dataset when data has been masked using rank swapping and microaggregation with different levels of protection (parameters p and k described in the text). From left to right and top to bottom number of common attributes in the integration process equal to $nC = 1, 2, 3, 4, 5$, and 6 .

Fig. 3. Error of the models (mean squared error) for the CASC dataset when data has been masked using rank swapping and microaggregation with different levels of protection (parameters p and k). First attribute used as the dependent attribute. From left to right and top to bottom number of common attributes in the integration process equal to $nC = 1, 2, 3, 4, 5, 6, 7$, and 8 .

Figure 3 provides the results for the CASC dataset. They are the results of mean squared error for the linear model of the first attribute. Figures on top refer to data masked using rank swapping and the figures in the bottom refer to

Fig. 4. Error of the models (mean squared error) for the CASC dataset when data has been masked using rank swapping and microaggregation with different levels of protection. Attribute 13th used as the dependent attribute. From left to right and top to bottom number of common attributes in the integration process equal to $nC = 1, 2, 3, 4, 5, 6, 7$, and 8.

data masked using microaggregation. Then, figures correspond to different sets of common attributes. More particularly, from left to right and top to bottom we have $nC = 1, 2, 3, 4, 5, 6, 7, 8$.

Figure 4 also correspond to the CASC dataset for the last attribute (13th attribute).

These figures for the CASC dataset show in some cases this same trend of larger error for stronger protection, but this trend is not so clear in most of the figures. The clearer cases correspond to microaggregated files for models built with the first variable. Note that the results are the average of 5 runs for the CASC dataset.

5 Conclusions and Future Work

In this paper, we presented experimental results on how masking methods (microaggregation and rank swapping) protect privacy in respect of data integration. We applied both masking methods on two datasets and further evaluated how a data-driven model (a linear regression data model) behaves in respect of database integration with different privacy protection extents. Especially we have experimented with different number of common attributes between these two databases in terms of prediction performance on record linkage in data integration. We concluded, based on our preliminary results, that while the number of correct linkages between two masked databases drop very quickly with respect to the amount of protection, the quality of data-driven models does not degrade so quickly.

These results are in line of previous research in data privacy which shows that some data masking and, thus, data protection, can be achieved with low or even no cost for machine and statistical learning applications. See e.g. the discussion in [13]. The reason for this behavior need additional research.

The results in this paper suggested several additional interesting directions for future work. Firstly, we plan to extend the data integration scenario with privacy protection for more than two databases, which is critically important in big data era. Secondly, we intend to investigate semantic based data integration for privacy protection, which will have a hybrid consideration of string match and semantic match common attributes. Thirdly, we will further evaluate the privacy protection mechanisms in more machine learning prediction models (e.g., random forest, support vector machine, deep neural networks).

Acknowledgements. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Partial support by Vetenskapsrådet project: “Disclosure risk and transparency in big data privacy” (VR 2016-03346, 2017-2020) is also acknowledged.

References

1. Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Pre-proceedings of ETK-NTTS 2001, vol. 2, pp. 807–826. Eurostat (2001)
2. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L. (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–134. North-Holland (2001)
3. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
4. Hansen, S., Mukherjee, S.: A polynomial algorithm for optimal univariate microaggregation. *IEEE Trans. Know. Data Eng.* **15**(4), 1043–1044 (2003)
5. Hundepool, A., et al.: Statistical Disclosure Control. Wiley, Hoboken (2012)
6. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *J. Am. Stat. Assoc.* **84**(406), 414–420 (1989)
7. Nin, J., Herranz, J., Torra, V.: Rethinking rank swapping to decrease disclosure risk. *Data Knowl. Eng.* **64**(1), 346–364 (2007)
8. Nin, J., Herranz, J., Torra, V.: On the disclosure risk of multivariate microaggregation. *Data Knowl. Eng.* **67**, 399–412 (2008)
9. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control, statistical. *J. U. N. Econ. Comm. Eur.* **18**(4), 345–354 (2000)
10. Samarati, P.: Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001)
11. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k -anonymity and its enforcement through generalization and suppression. SRI Intl. Technical report (1998)
12. Templ, M.: Statistical disclosure control for microdata using the r-package *sdcmicro*. *Trans. Data Priv.* **1**(2), 67–85 (2008)
13. Torra, V.: Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer, Heidelberg (2017)
14. Winkler, W.E.: Re-identification methods for masked microdata. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 216–230. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25955-8_17
15. Yeh, I.-C.: Modeling of strength of high performance concrete using artificial neural networks. *Cem. Concr. Res.* **28**(12), 1797–1808 (1998)
16. Yeh, I.-C.: Analysis of strength of concrete using design of experiments and neural networks. *J. Mater. Civ. Eng. ASCE* **18**(4), 597–604 (2006)
17. Leymann, F.: A practitioners approach to data federation. In 4. In: Workshop Föderierte Datenbanken, Berlin, Germany. CEUR-WS/Vol.25 (1999)
18. Vu, X.S., Mlouk, A., Elmroth, E., Jiang, L.: Graph-based interactive data federation system for heterogeneous data retrieval and analytics. *WWW* **19**(5), 3595–3599 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.