Research Article **ISSN: 2395 – 5597**

International Journal of Engineering and Robot Technology

Journal home page: www.ijerobot.com

A STUDY OF ENHANCING SECURITY OF SENSITIVE STATISTICAL INFORMATION USING HYBRID PARADIGMS

D. Rajesh*[1], D. Ramesh[2], D. Giji Kiruba[3]

[*1]Department of Computer Science Engineering, Universal College of Engineering and Technology, Tamilnadu, India.
[2]Department of Computer Science Engineering, James College of Engineering and Technology, Tamilnadu, India.
[3]Department of Electrical and Electronics Engineering, Government Polytechnic College, Nagercoil, Tamilnadu, India.

ABSTRACT
Confidential data of people are often collected, stored, and published by different entities, such as statistical agencies or hospitals, to be analyzed and used by decision makers, politicians, researchers, etc. This leads to new security issues such as compromising the confidentiality of people. So we need the mechanism to protect the date sets and ensuring confidentiality of people. There are many paradigms to protect data sets containing sensitive statistical information have been proposed. The two main paradigms for data set protection are Classical and Synthetic. Recently, the possibility of combining the two paradigms, leading to a hybrid paradigm, has been considered. In this work, the securities of some synthetic and classical methods have analyzed and conclude that they suffer from a high interval disclosure risk. In this paper, the fully hybrid method is proposed to protect the confidentiality of statistical data sets with the goal of reducing its interval disclosure risk.

KEY WORDS
Statistical data sets protection, Synthetic methods and Hybrid methods.

Author of correspondence:

D. Rajesh,
Department of Computer Science Engineering,
Universal College of Engineering and Technology,
Tamilnadu, India.

Email: rajeshd936@gmail.com

INTRODUCTION

The Data collected from people may be confidential. This information used by Decision makers, politicians, researchers, etc. This dissemination of confidential information should ensure, however, that the privacy of people is protected in some way, to be in accordance with current laws and regulations. One approach to achieve some level of privacy in this scenario is the application of some

protection methods to the confidential data, before making them public. The discipline that studies these protection methods is known as Statistical Disclosure Control (SDC)[1].

A suitable protection method is well considered if it achieves a good tradeoff between privacy and statistical utility. Two main paradigms have been proposed to design SDC protection methods. They differ on the kind of original information they perturb. In a statistical data set, we can distinguish between non confidential attributes and confidential attributes, depending on the sensitivity of the information of the attribute. For example, the nationality or age of a citizen is usually considered to be non-confidential attributes, whereas his income or the result of some medical analysis can be considered as confidential attributes. The first paradigm for SDC protection that we denote as classical consists in perturbing the non-confidential attributes only. In this way, the combinations of values which could unambiguously identify an individual disappear. This obfuscation makes it difficult for an intruder to establish relations between the protected data set and external data. Also, as this kind of methods does not modify the confidential attributes, third parties have precise information on confidential data, without knowing to whom this confidential data belongs[7].

The second paradigm that we denote as synthetic consists in building a data model for the confidential attributes from the whole original data set and then randomly generating a new synthetic version of the confidential attributes, constrained by the computed model. This approach preserves the statistical information embedded in the synthetic model but it disregards all the statistics not considered in the model. However, since non confidential attributes are released as they are, an intruder is able to automatically link a protected record with an external database. The security of this paradigm relies, in principle, on the fact that confidential attributes are completely synthetic, instead of a perturbed version of the original confidential values.

Whereas the ways of measuring the statistical utility of a SDC method are quite universal, independent of the paradigm, this is not the case when measuring the privacy level offered by a particular SDC method (one exception is differential privacy[2], that we discuss in detail in Section 5). This is because an attacker trying to obtain some information about the original confidential attributes has access to different kinds of data in each of the two considered paradigms. On the one hand, in the classical paradigm, an attacker has access to the original confidential data but he cannot relate them with concrete individuals because non confidential attributes are modified before their publication. On the other hand, in the synthetic paradigm the attacker knows the original non confidential attributes and, therefore, he can establish relations between the protected records and real individuals, but he cannot obtain the original confidential attributes because they are randomly generated from a statistical model. Combining the two paradigms sounds like a good idea. This would lead to a third paradigm for SDC protection that we denote as hybrid. However, very few have been done in this direction. Very recently in[3] authors show how to obtain a hybrid data set by combining micro aggregation[4], a well-known classical perturbative protection method, with any synthetic data generator. However, the resulting method, called micro hybrid, modifies only the confidential attributes, as in the synthetic paradigm. Indeed, the (implicit) use of micro aggregation is for producing clusters of close records, and then these clusters are the input data for a set of synthetic data generators, that are really in charge of data protection of the confidential attributes. When studying[3], we noticed that the privacy analysis therein is not the correct one for synthetic protection methods. Author's of[3] assume that an intruder has access to all the confidential attributes and then tries to find relations between these confidential attributes and the synthetic ones. This attack is not realistic: if the attacker already knows the confidential information, there is nothing to protect. In contrast, real attacks for the synthetic and hybrid paradigms, specifically interval disclosure attacks, were not considered. Of course, the more security is achieved, the more statistical utility is lost. The SDC method that results from combining MS with our post processing algorithm is clearly hybrid. We test this

method on the same data sets that we employ for the previous experiments. The results show that, in most of the cases, the disclosure risk can be significantly decreased at the cost of a minimum loss in statistical utility.

CLASSICAL PARADIGM

The paradigm for statistical data set protection that we denote as classical is motivated by the fact that information contained in the confidential attributes is typically the most significant or interesting one, from a statistical point of view. For this reason, protection methods in this paradigm do not perturb confidential attributes; only the non-confidential attributes are modified, by some protection method, which does not take into account at all the values of the confidential attributes. Many different protection methods have been proposed and analyzed, including noise addition[9], resampling, etc. In this work, we will use two of these classical methods, rank swapping and micro aggregation that we briefly explain now.

The idea of micro aggregation is to hide an original record inside a group of k protected records, all of them having the same protected non confidential attributes. In this way, the published data set R0 enjoys k-anonymity[11,12]: k protected records have exactly the same probability to correspond to a given original record. To apply a micro aggregation method, groups of k records with close non confidential attributes are formed, and these values are substituted by their centroid. In other words, if is one such group, and centroid of the non-confidential values then we have to achieve minimum information loss, the goal is to find an optimal micro aggregation that minimizes the SSE (i.e., the sum of distances between original records and centroids). Since finding the optimal micro aggregation is in general NP-hard[13], efficient heuristic algorithms like MDAV[4] have been proposed to provide good quality results.

SYNTHETIC PARADIGM

SDC methods in the synthetic paradigm behave the opposite way as those in the classical paradigm: they perturb only the confidential attributes, whereas original non confidential attributes remain unchanged. The new, perturbed values of the confidential attributes are not obtained now by swapping the original confidential values. Instead, the idea is to build up a theoretical/mathematical model from the whole original database R and then replace the confidential part Y with new synthetic values Y 0 which follow the same model as the original ones. In this way, depending on the considered model, some statistics of the original data set can be exactly preserved. For instance, in the IPSO synthetic protection method[6], a linear regression model between original parts X and Y is built up, and new synthetic confidential values Y 0 are randomly generated from X, according to this model. In this way, the mean vector and the covariance matrix of the original data set R are preserved[8]. This idea was extended in[5], so that besides preserving the mean vector and the covariance matrix, the protection method also guarantees similarity of the synthetic confidential values to the original confidential values.

Regarding measures for the privacy risk in this synthetic paradigm, let us first argue that the Linkage Disclosure Risk is not suitable now to measure the real risk of the system in front of real intruders[10]. First of all, if one considers a distance-based record linkage based on the non-confidential attributes, as in the classical paradigm, since these attributes are not modified by synthetic protection methods, each protected record is linked to its original record. However, since confidential attributes have now been changed, we could consider that only in the event that the generated synthetic values coincide with the original confidential attributes there is information disclosure. It is clear that such an approach would yield disclosure risks that would be simply zero, which contradicts the fact that not all synthetic generators provide the same degree of protection. For instance, let us compare a method which simply puts random values in the confidential attributes (high protection but useless data) with a protection method that simply multiplies each confidential value by 1.001. The latter method is clearly unsafe although it's Linkage Disclosure Risk, as previously defined, would be zero.

Another kind of Linkage Disclosure Risk was considered in[3], in which a distance-based record linkage between all the original and protected confidential information was used. However, it is clear that an intruder cannot be assumed to know all the original confidential information of the data set, because in this case, there is no privacy at all. Therefore, even if linkage disclosure may be used as a way to compare different (parameterizations of) synthetic methods, it cannot be considered as a "measure of disclosure risk" (as it was incorrectly done in[3]).

In our opinion, this argument is neither correct nor formal. For instance, as we have explained above, the synthetic method MS guarantees similarity between the synthetic confidential values and the original confidential values. If an intruder is not able to obtain the exact value of the income of a citizen, but he is able to obtain a very good approximation of this income, then it is quite evident that the privacy of this confidential attribute has been seriously damaged. Therefore, it is clear that some kind of "interval disclosure risk (IDR)" must be considered and analyzed. This is what we do in this paper, starting by the definitions of both absolute and relative interval disclosure risks. Before that, we introduce the hybrid paradigm, because some kind of interval disclosure risk will be a suitable risk measure also for hybrid protection methods.

HYBRID PARADIGM

What happens if one combines the two previous paradigms? That is, one can apply a classical protection method to non-confidential attributes; apply some synthetic methods to obtain synthetic confidential attributes and finally publish the protected data set. This sounds as a potentially good idea, but it has apparently received very few attentions from the SDC community; may be the reason is that researchers have believed that the information loss produced by such a combination could be very high, or that it would be difficult to define a good measure for the disclosure risk for this "hybrid" paradigm.

Recently, Domingo-Ferrer and Gonza´lez-Nicola´s[3] have partially considered this possibility of combining classical and synthetic techniques when designing a SDC protection method. Their idea is to apply ks-micro aggregation to the non-confidential attributes so that clusters are implicitly formed, and then apply an independent synthetic procedure for each cluster. The model, restricted to each cluster, will be more precise and so the produced synthetic data will be more similar to the original confidential data of the cluster. Since they combine micro aggregation and hybrid techniques, they call their method micro hybrid. However, in their proposal the non-confidential attributes are never modified. Therefore, strictly speaking, MH-ks can be seen as a synthetic protection method where only confidential attributes are modified. The difference with respect to previous synthetic methods is the way how the model is built: now different and independent models are built for different parts of the data set. We keep the expression "hybrid" for SDC methods that modify both the non-confidential attributes (through some classical method) and the confidential attributes (through some synthetic method).

What about privacy risks in the hybrid paradigm? The intruder observes protected records of the form and we assume that he has also obtained original non confidential information xi from an external data source. His goal would be then to link xi with the appropriate protected recording (through a distance-based record linkage process) and hope that the corresponding synthetic information falls inside a small interval centered at confidential original information. In other words, a good measure of disclosure risk for the hybrid scenario is a combination of both the Linkage Disclosure Risk and the Interval Disclosure Risk. The resulting measure is the one that we introduce in the next section.

SIMPLE TECHNIQES FOR HYBRID PROTECTION

We propose here some simple techniques that can really be classified as hybrid, because values of the non-confidential attributes are modified by applying some classical technique (in our case, micro aggregation) and original values of the confidential

attributes are replaced by synthetically generated ones.

The first proposed technique that we call the natural hybrid method that results from the ideas in[3]. That is, micro aggregation is first used to construct clusters in the non-confidential attributes, and the synthetic data generator MS is applied to each resulting cluster, independently, to generate the new confidential values. After that, the original non confidential attributes are modified by applying micro aggregation, where k may be equal or different to ks. In this way, we ensure k-anonymity for the non-confidential attributes, which makes the linkage + interval disclosure risk decrease[14].

The second proposed technique, that we call conceptually even simpler: the synthetic method MS is applied to the confidential attributes of the whole data set, as usual, and then k-micro aggregation is independently applied to the non-confidential attributes. Again, k-anonymity holds and one would expect a decrease in the linkage + interval disclosure risk, with respect to applying only MS. Namely, even if an intruder can link an original non confidential vector xi with the correct cluster in the protected data set, maybe most of the (synthetic) confidential values in this cluster are far from the original confidential values yi. In some way, this approach has the same goal as p-sensitivity diversity, or t-closeness ensuring a minimum level of protection for confidential attributes[15].

CONCLUSION

In this work, we have analyzed the security offered by some synthetic SDC protection methods recently proposed in the literature. In conclusion, we believe more care should be taken when proposing new synthetic and hybrid SDC protection methods, regarding the possible risks of disclosure. In this sense, we expect the new definition of linkage + interval disclosure risk that we propose in this work will help future researchers. Furthermore, the proposed post processing algorithm can be though as a general and useful technique that can be applied after the execution of any (existing or future) synthetic protection method, with the goal of decreasing disclosure risks while maintaining statistical utility. We have chosen rank swapping and micro aggregation to implement our post processing techniques, because they are popular, simple, and also known to provide a good trade-off between privacy and utility. But other classical protection methods could be used instead, such as noise addition, re sampling, etc. We leave this option as a possible line for future research.

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I would like to express my gratitude towards my parents and member of Universal College of Engineering and Technology for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

CONFLICT OF INTEREST

We declare that we have no conflict of interest.

BIBLIOGRAPHY

1. Willenborg L and De Waal T. Elements of Statistical Diclosure Control: Lecture Notes in Statistics, *Springer,* 155, 2001.
2. Dwork C. "Differential Privacy," *Proc. Int'l Conf. Automata, Languages and Programming (ICALP),* 2006, 1-12.
3. Domingo Ferrer J and Gonza´lez-Nicola´s U. "Hybrid Micro data Using Micro aggregation," *Information Sciences,* 180(15), 2010, 2834-2844.
4. Domingo-Ferrer J and Mateo-Sanz J M. "Practical Data-Oriented Micro aggregation for Statistical Disclosure Control," *IEEE Trans. Knowledge and Data Eng.,* 14(1), 2002, 189-201.

5. Muralidhar K and Sarathy R. "Generating Sufficiency-Based Non-Synthetic Perturbed Data," *Trans. Data Privacy*, 1(1), 2008, 17-33.
6. Burridge J. "Information Preserving Statistical Obfuscation," *Statistics and Computing*, 13, 2003, 321-327.
7. Moore R. "Controlled Data Swapping Techniques for Masking Public Use Micro data Sets," *U.S. Census*, 1996.
8. Domingo-Ferrer J and Torra V. "Disclosure Control Methods and Information Loss for Micro data," Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, *North Holland*, 2001, 91-110.
9. Kim J. "A Method for Limiting Disclosure in Micro data Based on Random Noise and Transformation," *Proc. ASA Section on Survey Research Methodology*, 1986, 303-308.
10. Nin J, Herranz J and Torra V. "Rethinking Rank Swapping to Decrease Disclosure Risk," *Data and Knowledge Eng.*, 64(1), 2008, 346-364.
11. Sweeney L. "k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty Fuzziness Knowledge-Based Systems*, 10(5), 2002, 557-570.
12. Sweeney L. "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *Int'l J. Uncertainty Fuzziness Knowledge-Based Systems*, 10(5), 2002, 571-588.
13. Oganian A and Domingo-Ferrer J. "On the Complexity of Optimal Micro aggregation for Statistical Disclosure Control, "*Statistical J. United Nations Economic Commission for Europe*, 18(4), 2000, 345-354.
14. Winkler W E. "Matching and Record Linkage," *Business Survey Methods, Wiley*, 1995, 355-384.
15. Agrawal R and Srikant R. "Privacy-Preserving Data Mining", *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD)*, 2000, 439-450.

Please cite this article in press as: D. Rajesh *et al.* A Study of Enhancing Security of Sensitive Statistical Information Using Hybrid Paradigms, *International Journal of Engineering and Robot Technology*, 1(2), 2014, 56 - 61.