Bull. Aust. Math. Soc. **104** (2021), 346–348
doi:10.1017/S0004972721000289

INVESTIGATING ATTRIBUTE RISKS AND CONSTRUCTING LINKAGE ERROR MODELS FOR PROBABILISTICALLY-LINKED DATA

Y. MA

(Received 12 March 2021; first published online 19 April 2021)

2020 *Mathematics subject classification*: primary 68P27; secondary 60H40.

Keywords and phrases: data privacy, record linkage, official statistics, noise multiplication masking, remote system, linkage error model, microdata, tabular data, statistical disclosure limitation.

The thesis presents my PhD research achievements in two areas of official statistics. The first area is data privacy. This is concerned with protecting sensitive information of data respondents in public release data. The second area is record linkage. This is concerned with combining records located in different data sources but belonging to the same population unit.

The majority of the thesis concerns data privacy. A statistical agency can release two types of data: microdata and tabular data. Regardless of the form of data, a statistical agency needs to guarantee that sensitive information of data respondents is not disclosed to data intruders. There are two types of disclosure risks that need to be controlled. The first is reidentification risk. Reidentification disclosure occurs if a data intruder correctly associates a record with the corresponding population unit. The second is attribute risk. Attribute disclosure occurs if a data intruder learns new sensitive information about a target population unit. The thesis mainly concerns attribute risks.

To reduce disclosure risks, a common practice used by a statistical agency is to apply a statistical disclosure limitation (SDL) method to original data before data release. An SDL method is a way to alter the original data systematically so that the risks of disclosure are reduced. On the other hand, the process will also reduce the value of the data as an analytical resource, which can be described as data utility. This means that statistical information that one can infer from the altered data is less accurate than what one can infer from the original data. To study an SDL method, both its impact on data utility and data privacy should be considered.

Thesis submitted to the University of Wollongong in March 2019; degree approved on 18 September 2020; supervisors Yan-Xia Lin, James Chipperfield and Pavel Krivitsky.

© 2021 Australian Mathematical Publishing Association Inc.

https://doi.org/10.1017/S0004972721000289 Published online by Cambridge University Press

Noise multiplication masking is an emerging SDL method applicable to both microdata and tabular data. The masking method works by multiplying each observation with a random noise drawn from an underlying noise-generating variable to create a noise-multiplied version of the original data. Statistical estimates can be inferred from the noise-multiplied data using estimators specifically derived for noise-multiplied data. The distribution of the underlying noise-generating variable plays a crucial role in balancing utility–risk trade-offs. The noise multiplication masking method is advocated by many researchers (see, for example, [4]).

There is extensive research on the noise multiplication masking method from the data utility point of view (see, for example, [3]). However, there is inadequate research on attribute risks associated with the masking method. Understanding attribute risks of the masking method can help a statistical agency to select the distribution of an underlying noise-generating variable wisely before data masking takes place. Parts of my research attempt to understand attribute risks of the masking method (see [6, 7]). In this thesis, the following research outcomes are presented:

(1) a potential attacking strategy, namely 'correlation-attack', is discussed which can easily cause attribute disclosure;
(2) a measure for quantifying the average attribute risk of a noise distribution against several attacking strategies is derived using an optimisation model;
(3) the fact that some data intruders might have knowledge about the minimum or maximum of the original data is noted and the consequences related to attribute risks are discussed.

In addition to the topics related to the noise multiplication masking method, one chapter of the thesis discusses tabular data protection in a remote system. In a remote system, data are stored in a remote server. Data users cannot access these data directly. Instead, data users can interact with the data by sending queries. A query could be any particular statistical information a data user wants to know about a particular set of data, such as the aggregate total of the data. The remote system responds to a query by returning an output. When querying on aggregate totals is allowed in a remote system, a differencing attack can cause attribute disclosure. A differencing attack can happen if a data intruder is able to send two queries of aggregate totals, and the contributor values of the two queries are the same except that a target contributor value is excluded in one of the queries. Attribute risks from differencing attacks can be reduced if an output perturbation algorithm is used. In the chapter, we introduce an innovative output perturbation algorithm against differencing attack strategies (see also [5]). The performance of the algorithm is compared with an algorithm in [8].

Besides data privacy, we also present our research achievements on record linkage in one chapter. Record linkage is needed when a statistical agency has multiple microdata files with overlapping information and wants to increase the depth and dimension of the data. The statistical agency might attempt to link these files using a record linkage technique, such as a probabilistic linkage algorithm (see, for example, [2]). A probabilistic linkage algorithm might link records incorrectly, causing

https://doi.org/10.1017/S0004972721000289 Published online by Cambridge University Press

linkage errors in linked data. Analysing linked data with linkage errors will lead to biased statistical estimates for many population parameters. A linkage error model is a matrix which estimates the structure of linkage errors in linked data. A linkage error model can be used to offset biases of statistical estimates. However, constructing a linkage error model seems to be nontrivial. Chambers [1] proposes an exchangeable linkage error (ELE) model which imposes a strong assumption on linkage error structure. Moreover, the ELE model might still lead to biased statistical estimates. In the thesis, I propose an alternative linkage error model called the conditional linkage error (CLE) model. The CLE model does not require the strong assumption made by the ELE model, and it has been shown that the model works well in many situations where the ELE model fails to correct estimation biases. The performance of the two linkage error models is compared via simulations.

References

[1] R. Chambers, 'Regression analysis of probability-linked data', *Statisphere Off. Stat. Res.* **4** (2009).
[2] J. O. Chipperfield and R. L. Chambers, 'Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data', *J. Off. Stat.* **31**(3) (2015), 397–414.
[3] Y. X. Lin, 'Density approximant based on noise multiplied data', in: *Privacy in Statistical Databases*, Lecture Notes in Computer Science, 8744 (Springer, Cham, 2014), 89–104.
[4] Y. X. Lin and P. Wise, 'Estimation of regression parameters from noise multiplied data', *J. Priv. Confid.* **4** (2012), 61–94.
[5] Y. Ma, Y. X. Lin, J. Chipperfield, J. Newman and V. Leaver, 'A new algorithm for protecting aggregate business microdata via a remote system', in: *Privacy in Statistical Databases 2016,* Lecture Notes in Computer Science, 9867 (eds. J. Domingo-Ferrer and M. Pejić-Bach) (Springer, Cham 2016), 210–221.
[6] Y. Ma, Y. X. Lin, P. N. Krivitsky and B. Wakefield, 'Quantifying the protection level of a noise candidate for noise multiplication masking scheme', in: *Privacy in Statistical Databases 2018,* Lecture Notes in Computer Science, 11126 (eds. J. Domingo-Ferrer and F. Montes) (Springer, Cham, 2018), 279–293.
[7] Y. Ma, Y. X. Lin and R. Sarathy, 'The vulnerability of multiplicative noise protection to correlation-attacks on continuous microdata', *Sankhyā Ser. B* **82** (2020), 305–327.
[8] G. Thompson, S. Broadfoot and D. Elazar, 'Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Australian Bureau of Statistics', in: *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality* (UNECE, Ottawa, 2013).

Y. MA, School of Mathematics and Applied Statistics,
University of Wollongong, Wollongong,
New South Wales 2522, Australia
e-mail: mayue3588@gmail.com

https://doi.org/10.1017/S0004972721000289 Published online by Cambridge University Press