Title: Data privacy preserving scheme using generalised linear models

Author: Min Cherng Lee, Robin Mitra, Emmanuel Lazaridis, An Chow Lai, Yong Kheng Goh, Wun-She Yap

PII: S0167-4048(16)30180-8
DOI: http://dx.doi.org/doi: 10.1016/j.cose.2016.12.009
Reference: COSE 1081

To appear in: *Computers & Security*

Please cite this article as: Min Cherng Lee, Robin Mitra, Emmanuel Lazaridis, An Chow Lai, Yong Kheng Goh, Wun-She Yap, Data privacy preserving scheme using generalised linear models, *Computers & Security* (2016), http://dx.doi.org/doi: 10.1016/j.cose.2016.12.009.

Data Privacy Preserving Scheme using Generalised Linear Models[☆]

Min Cherng Lee[a], Robin Mitra[b], Emmanuel Lazaridis[c], An Chow Lai[a], Yong Kheng Goh[a], Wun-She Yap[a]

[a]*Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Malaysia*

{leemc,laiac,gohyk,yapws}@utar.edu.my

[b]*Southampton Statistical Sciences Research Institute, University of Southampton, United Kingdom*

R.Mitra@soton.ac.uk

[c]*National Institute for Cardiovascular Outcomes Research, University College London, United Kingdom*

emmanuel@lazaridis.eu

[☆]An earlier version of this work [8] appeared in ACISP 2016 as an invited paper.

Abstract

When releasing data for public use, statistical agencies seek to reduce the risk of disclosure, while preserving the utility of the release data. Commonly used approaches (such as adding random noises, top coding variables and swapping data values) will distort the relationships in the original data. To preserve the utility and reduce the risk of disclosure for the released data, we consider the synthetic data approach in this paper where we release multiply imputed partially synthetic data sets comprising original data values, and with values at high disclosure risk being replaced by synthetic values. To generate such synthetic data, we introduce a new variant of factored regression model proposed by Lee and Mitra in 2016. In addition, we take a step forward to propose a new algorithm in identifying the original data that need to be replaced with synthetic data. More importantly, the algorithm that can identify the original data with high disclosure risk can be applied on other existing statistical disclosure control schemes. By using our proposed scheme, data privacy can be preserved since it is difficult to identify the individual under the scenario that the released synthetic data are not entirely similar with the original data. Besides, valid inference about the data can be made using simple combining rules, which take the uncertainty due to the presence of synthetic values. To evaluate the performance of our