NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

UNIVERSITY OF
WOLLONGONG
AUSTRALIA

Workshop on Advances in Statistical Disclosure Limitation

Wednesday 5 February 2020

SMART Building 6, Room 105, University of Wollongong Campus

Programme

9.00am	**Registration** **Welcome to the workshop (Yan-Xia Lin)**
9.15am – 10.45am	**Short course on Data Privacy and Statistical Disclosure Limitation by Anna Oganyan** (National Centre for Health Statistics)
10.45am – 11.00am:	**Morning Break**
Session 1	**Chairperson: Pauline O'Shaughnessy, UOW**
11.00am – 11.30am:	**Speaker 1 - Bradley Wakefield,** University of Wollongong: *Synthetic data generation – the moment-based approach*
11.30am – 12.00pm:	**Speaker 2 – Edwin Lu,** ABS: *Unit Risk Measure: Nearest-Neighbour Algorithm*
12.00pm – 12.30pm:	**Speaker 3 – Anna Oganyan,** National Centre for Health Statistics: *Ideas for dissemination of genetic data - synthetic approach*
12.30pm – 1.30pm:	**Working lunch session with round table discussion** **Organiser: Yan-Xia Lin, UOW and Chris Mann, ABS**

Workshop on Advances in Statistical Disclosure Limitation

Programme

Session 2	**Chairperson: Bradley Wakefield, UOW**
1.30pm – 2.00pm:	**Speaker 4 – Ming Ding,** Data61: *Performance Analysis and Optimization in Privacy-Preserving Federated Learning*
2.00pm – 2.30pm:	**Speaker 5 – James Bailie,** ABS: *Designing formal privacy mechanisms for the p% rule*
2.30pm – 3.00pm:	**Speaker 6 – Goran Lesaja,** Georgia Southern University: *Controlled Tabular Adjustment Problem: Theory, Models and Computations*
3.00pm – 3.15pm:	**Afternoon Break**
Session 3	**Chairperson: Joseph Chien, ABS**
3.15pm – 3.45pm:	**Speaker 7 – Anna Oganyan,** National Centre for Health Statistics: *SDL of high-dimensional data – clustering approach*
3.45pm – 4.15pm	**Speaker 8 – Pauline O’Shaughnessy,** University of Wollongong: *A general framework for measuring disclosure*
4.15pm:	**Finish**
6.00pm	**Informal Dinner at Samaras Restaurant** **Corner of Market and Corrimal St, Wollongong**

Workshop on Advances in Statistical Disclosure Limitation

Titles and Abstracts

Wednesday 5 February 2020

SMART Building 6, Room 105, University of Wollongong Campus

Short Course on Data Privacy and Statistical Disclosure Limitation

Keynote Speaker: Dr Anna Oganyan, National Centre for Health Statistics, USA

Abstract: In this talk I will define key concepts of microdata protection and describe some relevant Statistical Disclosure Limitation (SDL) methods. I will start with basic definitions and describe a structure of a microdata file. Before releasing such data to the public, statistical agencies have an obligation by law to protect the confidentiality of the respondents/data providers and at the same time they strive to release a product that would satisfy the ever growing demands of potential data users. So, the goal of microdata protection is two-fold: minimize the risk of disclosure of respondents' confidential information and maximize the utility of the released data for the user. The key issue here is that these goals are conflicting goals. To decrease the disclosure risk, data protector typically has to perturb microdata in some way, which often leads to decreased utility of the resultant data to the user. On the other hand, the efforts to improve the utility of the perturbed protected data can lead to the increased disclosure risk. So, a trade-off between data utility and disclosure risk is the main issue of SDL practice. This is why a decision about how to define and measure data utility and disclosure risk should be among the first steps in the process of microdata protection. This will help not only to better understand, but to choose and compare the existing SDL methods and also to develop the most appropriate protection strategy for a particular scenario of data release. I will give examples of such definitions and discuss their advantages and disadvantages. After that, I will present several SDL methods suitable for the protection of microdata and discuss their effectiveness based on the proposed metrics of utility and disclosure risk.

10.45am – 11.00am: Morning Break

Speaker 1 - Bradley Wakefield

University of Wollongong bnw722@uowmail.edu.au

Synthetic data generation – the moment-based approach

Abstract: The multivariate moment's problem and its application to the estimation of joint density functions are often considered highly impracticable in modern-day analysis. Although many results exist within the framework of this issue, it is often an undesirable method to be used in real-world applications due to the computational complexity and intractability of higher-order moments.

This talk will demonstrate the application of the multivariate moment based method in estimating a joint density for the purpose of generating privacy preserving synthetic data. The talk investigates the varying statistical properties of such an approach to synthetic data and assess the privacy protections such data has relative to other statistical disclosure control methods. Importantly, we note that in order to investigate the privacy implications of any disclosure of information, we must first understand what information is destroyed and what uncertainty is introduced. By doing so, we can avoid excessive perturbation and optimise statistical utility.

Speaker 2 – Edwin Lu

Australian Bureau of Statistics

Unit Risk Measure: Nearest-Neighbour Algorithm

Abstract: The Australian Bureau of Statistics (ABS) currently adopts the Five Safes Framework for managing disclosure risk while maximising the utility of public sector data. Under this framework, expanded confidentialised unit record files (expanded CURFs) and detailed microdata files are made accessible through the ABS DataLab environment, which has statistical analysis tools like R and SAS that enable complex modelling and analysis. The ABS DataLab imposes moderate to strong controls under four dimensions of the five safes. In addition to the controls of the DataLab, a unit record risk measure based on the nearest-neighbour algorithm was developed to help mitigate the risk of spontaneous recognition – the risk that, in the ordinary course of analysis, a user of the data unintentionally identifies a person to which a record in the data belongs. This presentation will provide some context around spontaneous recognition risk in the ABS DataLab, then describe the nearest-neighbour algorithm that has been used to detect spontaneous recognition risks as well as the treatments applied.

Speaker 3 – Anna Oganyan

National Centre for Health Statistics annaoganyan7@gmail.com

Ideas for dissemination of genetic data - synthetic approach

Abstract: Individual-level genetic data, such as Single-Nucleotide Polymorphism (SNP) data is usually considered to be of a sensitive nature and access to it is often restricted through the controlled environment such as Research Data Centers (RDC) at government institutions. Access to an RDC can be a lengthy process. Therefore, the availability of a version of genetic data which has undergone statistical disclosure limitation alternation to protect the privacy of the respondents may be beneficial to researchers who cannot travel to an RDC, to the students or to the researchers who are awaiting for the access to an RDC but in the meantime can get a better idea about the structure of the data set, test some of their hypotheses or even formulate appropriate research questions. In this talk, I will present some ideas of how to create synthetic SNP data using principles of evolutionary genetics. This is a preliminary and ongoing research. The presentation will be at the conceptual level to encourage discussion, suggestions and comments.

12.30pm – 1.30pm: Working lunch and round table discussion
Organisers Yan-Xia Lin and Chris Mann

Speaker 4 – Ming Ding

Data61 Ming.Ding@data61.csiro.au

Performance Analysis and Optimization in Privacy-Preserving Federated Learning

Abstract: As a popular means of decentralized machine learning, federated learning (FL) has recently drawn a large amount of attention. One of the prominent advantages in FL is its capability of preventing clients' data from being directly exposed to external adversaries. Nevertheless, from a viewpoint of information theory, it is still possible for an attacker to steal private information from eavesdropping/peeking on the shared FL models uploaded by FL clients. In order to address this problem, we develop a novel privacy preserving FL framework based on the concept of differential privacy (DP). To be specific, we first propose a client-oriented DP (CDP) by adding artificial noises to the shared FL models before uploading such models to FL servers. Then, we prove that our proposed CDP satisfies the DP guarantee with adjustable privacy protection levels by varying the variances of the artificial noises. Next, we develop a theoretical convergence upper-bound of the CDP algorithm. Our developed upper-bound reveals that there exists an optimal number of communication rounds to achiever the best convergence performance in terms of loss function values for a given privacy protection level. Furthermore, to obtain such optimal number of communication rounds, which cannot be derived in a closed-form expression, we propose a communication rounds discounting (CRD) method. Compared with the heuristic searching method, our proposed CRD can obtain a better trade-off between complexity and convergence performance. Extensive experiments indicate that our CDP algorithm with an optimization on the number of communication rounds using the proposed CRD can effectively improve the FL training efficiency and FL model quality for a given privacy protection level.

Speaker 5 – James Bailie

Australian Bureau of Statistics

Designing formal privacy mechanisms for the p% rule

Abstract: The p% rule classifies a cell in a frequency table as a disclosure risk if one contributor can use the cell to determine another contributor's value to within p%. This is often possible in economic data when there is a duopoly and therefore the p% rule is an important statistical disclosure control frequently used in national statistical organisations. However, the p% rule is only a method for assessing disclosure risk. While it can say whether a cell is risky or not, it does not provide a mechanism to decrease that risk. To address this limitation, we encode the p% rule into a formal privacy definition using the Pufferfish framework and we develop a perturbation mechanism which is provable private under this framework. This mechanism provides official statisticians with a method for perturbing data which guarantees a Bayesian formulation of the p% rule is satisfied. We motivate this work with an example application.

Speaker 6 – Goran Lesaja

Georgia Southern University goran@georgiasouthen.edu

Controlled Tabular Adjustment Problem: Theory, Models and Computations

Abstract: In this talk we consider a Controlled Tabular Adjustment (CTA) problem for statistical disclosure limitation (control) (SDL) of tabular data. The goal of the CTA model is to find the closest safe table to some original tabular data set that contains sensitive information. The closeness of original and masked table is usually measured using l1 or l2 norm; with each measure having its advantages and disadvantages. We present Second Order Cone reformulation of l1 - CTA and compare it with traditional Linear Programming formulation of – l1 CTA. Computational experiments using interior-point methods show competitiveness of the second order cone model for – l1 CTA when compared to existing models. Furthermore, we explore other measures of closeness between the original and masked tables which are based on table statistics such as chi-square.

3.00pm – 3.15pm: Afternoon Break

Speaker 7 – Anna Oganyan

National Centre for Health Statistics

SDL of high-dimensional data – clustering approach

Abstract: Data sets that are subject to Statistical Disclosure Limitation (SDL) often have many variables of different types that need to be altered for disclosure limitation. To produce a good quality public data set, the data protector needs to account for the relationships between the variables. Hence, ideally SDL methods should not be univariate, that is, treating each variable independently of others, but multivariate, handling several variables at the same time. In practice, multivariate SDL is difficult. Its complexity rapidly increases with the data dimensionality. In this presentation we discuss how different clustering techniques can serve as pre-masking data processing procedure where the subjects being clustered are the variables in the multivariate data sets, and also how data mining techniques such as Association Rule Mining (ARM) can be helpful in developing multivariate SDL methods.

Speaker 8 – Pauline O'Shaughnessy

University of Wollongong bnw722@uowmail.edu.au

A general framework for measuring disclosure

Abstract: Concerns for data privacy motivate the development of privacy protection methods and algorithms. However the current literature lacks a standard to evaluate the performance of these methods. In this talk, we introduce a generalised framework for measuring the disclosure risk of a micro-data released. This framework takes advantage of existing perspectives on disclosure risk measurements to provide a methodology that can be applied to any type of protected data. We demonstrate that this framework is able to properly account for the variation of disclosure risk between various differential privacy protection levels. Then we implement this framework to a real dataset to compare and contrast varying protection mechanisms.

4.15pm: Finish

6.00pm: Informal Dinner

Samaras Restaurant, Corner of Market and Corrimal St, Wollongong

(Please advise on attendance list if you plan to go to the dinner)