Reviewer's report

Title: De-identifying a Public Use Microdata File from the Canadian National Discharge Abstract Database

Version: 1 Date: 14 February 2011

Reviewer: Aris Gkoulalas-Divanis

Reviewer's report:

This paper aims to analyze the various plausible re-identification attacks to a Canadian DAD PUMF and to design a strategy to de-identify the dataset in a way that data utility is retained to an acceptable extent. After proposing a set of re-identification metrics for evaluating the types of attack, the authors propose an algorithm to anonymize the data. The work is interesting with findings that can be important to researchers in the area but in my opinion needs to undergo revision before accepted for publication. In the following I provide a set of major, minor and discretionary revisions for the paper.

Major Compulsory Revisions

1. The authors should provide some experiments to demonstrate the scalability of the proposed approach by presenting the runtime for the execution of the algorithm under different sample sizes and values of k. Also, how much time does each phase of the algorithm take to execute?

Minor Essential Revisions

1. In the Background (last paragraph - contributions): please correct "we present the results evaluation of".

2. In the Definitions (first paragraph): you mention "we make a distinction among four types of variables". Here you should cite Sweeney's and Samaratis's papers [IJFCS 2002 - "k-anonymity: a model for protecting privacy", TKDE 2001 - "protecting respondents identities in microdata release"] since they were the first to propose this type of categorization.

3. In the Definitions (Sensitive variables) you need to provide examples of sensitive variables as you did for the case of quasi-identifiers (previous section).

4. In the Definitions (Equivalence classes) you mention "For example, all the records in a dataset about 17 year old males admitted on 1st Jan 2008 are an equivalence class". What is the set of quasi-identifiers in this case?

5. In the Definitions (Equivalence classes) you mention "recoded" for the first time. You need to talk about the recoding process before that.

6. In the Definitions (Types of disclosure) you refer to "identity disclosure". You should also cite Sweeney's paper near the definition. Similarly, near the definition of the term "attribute disclosure" you should cite Machannavajjhala's et. al. ICDE 2006 paper on l-diversity ("l-diversity: privacy beyond k-anonymity").
7. In the Definitions (Types of disclosure) you mention "... who live in a particular province had a particular diagnosis". Here you assume that diagnosis is a sensitive attribute but before (in the "Indirectly identifying variables (quasi-identifiers)" section) you mentioned that this is a quasi-identifier. Please explain.

8. You need to restate the definition of attribute disclosure because now it is unclear. In particular, not every inference that could be drawn from the dataset leads to an attribute disclosure. It is the association between an identity and a value in a "sensitive attribute" that occurs with a "sufficiently" high probability.

9. In the Definitions (Types of disclosure) you cite papers [30-40] but not all of these papers correspond to re-identification cases. You should explicitly mention the cases and cite the corresponding papers for each case.

10. In the PUMF Specifications section you mention that "... cases with diagnoses or interventions that indicated abortion of HIV were excluded" from the dataset. Can you provide more information on why this choice was made?

11. In the PUMF Specifications (De-identification methods) section (paragraph one), the authors first mention the term "cell" for a dataset. The authors should explain well what the "cell" is.

12. In section PUMF Specification (De-identification methods) the authors make the following argument: "...suppression affects single records whereas generalization affects all of the records in the dataset". This argument is not true because most of the times suppression does not refer to whole records, i.e. you do not have to suppress a record but you can suppress attributes-values from a record. Also, suppression is a special case of generalization and therefore generalization retains more information so is generally preferred over suppression. Last, generalization does not necessary affect all the records in the dataset (e.g., in the case of local recoding models). I believe that the authors need to restate this argument.

13. In the section "Possible Attacks" (Attack 1) you assume that the attacker does not know whether an individuals's record is contained in the sample. This needs to be made clear from the begining of the section, otherwise the reidentification probability will be different from the one you assume.

14. In the "Possible attacks" section (Attack 1 - last paragraph): please correct "fo" to "for".

15. In the "Possible attacks" section (Attack 3 - first paragraph): please correct "lass" to "class" and "equivakence" to "equivalence".

16. In "Strategies for improving data utility" (disclosing all levels of the hierarchy) you mention that "the commonly used suppression algorithms operate on the whole table rather than on combinations of quasi-identifiers". This is a vague statement and you need to shed light on which are the "commonly used suppression algorithms" as well as to refer to works that do not follow this direction thus do not operate on the whole table.

17. In "Strategies for improving data utility" (disclosing all levels of the hierarchy - paragraph 8) please correct "combinationa".
18. In Results (paragraph 6) please correct the information about Table 11 (it seems that you copy-pasted the legend of the table).

Discretionary Revisions
1. Figure 2 is unnecessary and can be removed.
2. It would be nice if the algorithm is given as part of the main paper and not in the appendix.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.