T-79.514 Special Course on Cryptology / Privacy-Preserving Data Mining: Comments on “Revealing information while preserving privacy”

Isto Niemi

October 27, 2003

1 Introduction

This review briefly comments the survey [1] of the article “Revealing Information while Preserving Privacy” by Irit Dinur and Kobbi Nissim [2]. The survey examines the tradeoff between privacy and usability in statistical databases.

2 General Comments

The survey gives a good overview of the article by Dinur and Nissim [2]. The structure of the survey is clear and it follows intuitively the structure of the original paper. The amount of details presented is good and it is possible to go through the survey without the original paper. If a proof is skipped it is clearly pointed out.

Some definitions are written down more detailed in the survey [1] than in the original paper [2]. This makes the survey easier to follow for people without special background in cryptography. Though some background is still needed.

By reading the definitions it seems that the output perturbation approach is used but it is not clearly pointed out in the survey. Also the benefits for using the concept of non-privacy are not clear.

3 More Detailed Comments

1 Introduction. In the third paragraph it is said that the random noise is added to the database but in the article [2] it is said that the noise is added to the query results (subset sums). In the same paragraph there is an equation $\epsilon \ll \sqrt{n}$ without n defined.

2 Basic Concepts. The second notation is $dist(c, d) = |i|c_i \neq d_i|$. It is easier to read if a colon is used instead $dist(c, d) = |i : c_i \neq d_i|$.

2.1 Model-Statistical Databases and Statistical Queries. In definition 2 there is a minor notation difference compared to the original one. θ is used instead of σ.

2.2 Privacy Methods for Statistical Databases. The third sentence of the fifth paragraph needs a verb.

2.3 Database Privacy. In the second sentence of the second paragraph the variance of query answers and the estimator variance are mentioned as a privacy measure without cite or better definition.

In the fourth paragraph it is mentioned that Dinur and Nissim reverse the order in which they define the privacy. It might be a good idea to add some motivations why to choose the alternative path.

3 Impossibility Results. No comments.

3.1 Exponential Adversary. No comments.

3.2 Polynomially Bounded Adversary. In the sections “query phase” and “The last inequality holds” there is one missing parenthesis. It should be read $t = n(log^2(n))$.

In the section where the triangle inequality is applied the indexes i and j should be checked and corrected.

3.3 Tightness of the Impossibility Results. In the second sentence of the first paragraph comma is not needed before “that”.

In the second paragraph whitespace is missing. It should be $log\ n$.

4 Feasibility Results. It seems reasonable to leave out the detailed proof of Theorem 10.

5 Conclusions. There is a typo in the second sentence of the first paragraph. It should read “recovered”.

References

[1] E. Oikarinen. Revealing Information while Preserving Privacy. Survey for the Seminar T-79.514: Special Course on Cryptology/Privacy Preserving Data Mining at Helsinki University of Technology, October 2003

[2] I. Dinur and K. Nissim. Revealing Information while Preserving Privacy In Proc. of 22nd ACM SIGMOD-SIGACT-SIGART symposium, pp. 202-210. ACM Press. USA, 2003.