Comments on the survey "Privacy Preserving Data Mining"

Oleg Koshik

October 16, 2005

Introduction

Here I will give my brief comments and opinion on the survey "Privacy Preserving Data Mining". It should be considered that opponent has only basic knowledge in cryptography and has no previous experience in the field of data mining.

General Opinion

I find this survey quite interesting and on the whole well made. It gives a brief but at the same time sufficient overview of the given topic. Differently to some other surveys, this is survey not too technical and contains reasonable amount of motivational and conceptual context. Generally this paper requires no special knowledge to understand it. However, below is paid attention to some moments, which might be cleared or supplemented.

Notions

All main notions regarding to this topic are quite well explained in the text. However, some additional definitions could be given in footnotes, as they may be not clear to inexperienced reader. Such notions are 'clustering' and 'association-rule mining', which are given in the sections 2.1 and 2.2 as examples of data mining functions, as well as 'Zero-knowledge proof' in the end of the second paragraph of the section 3.2.

As I can understand, satisfying the DDH assumption essentially means that the Discrete Log Problem is hard in the given cyclic multiplicative group. If this statement is correct then it could be added to the first footnote on the page 3 in order to make this definition more understandable.

Comments pertaining to the content

In the description and further analysis of the secure-sum it is not quite clear, whether the sum is taken modulo n only one time, when P_1 passes its number to P_2.. Most likely it is so, in such case it should be mentioned in the point 4

of the description, that n must be added to the final sum if the first sum was reduced by n.

To my opinion, the section 3.4 regarding secure function computation is one that needs to be improved. I find the Figure 2 not very succeeded, as it does not make clear the construction and the mechanisms of the combinatorial circuit. The description of the Yao protocol I find not very good as well. I had to read it many times through and afterwards still needed to explore additionally one alternative source to understand it completely. Thus, this protocol should be introduced with some more details.

In the sections considering ID3 and especially privacy-preserving ID3 it might be useful to provide examples of the practical situations, where these algorithms can find application.