Review comments on "Secure Approximate Matching" survey, T-79.514 Special Course on Cryptology

Sami Vaarala

November 18, 2003

1 Introduction

We review the survey on secure approximate matching given in [1]. The survey is, for the most part, based on the paper "Protocols for Secure Remote Database Access with Approximate Matching" by Du and Atallah [2]. Review comments are given by sections of the survey. Finally, typos and other minor nits are given in the last section.

2 General comments

The paper describes a well selected subset, the SSO/SSCO problem, of the 25 page long paper. Some short description of the general idea behind the methods not described in detail would be good. Overall, the survey is easy to read.

3 Comments on Introduction

When using fingerprints, digitization errors occur both when the original fingerprint is scanned, and when a potential fingerprint is scanned. The text now is "... in the process of storing (digitizing)", should probably simply say "in the process of digitizing".

Example about DNA involving Alice and Bob. The example itself is good and motivates the problem well. One detail, though: knowledge of the query (= Alice's DNA) implies knowledge of the query result when we consider Bob (who also knows that database and can thus easily repeat the query. The text is a bit ambiguous now, and could be understood such that Alice's DNA and the match result are "independent" pieces of information.

SDA problem definition should probably state that all strings (q, $t_i \in T$) belong to some arbitrary set of strings (S for instance). Also, the result of the SDA procedure (for Alice) is a bit unclear. Does Alice get the matching string, or just knowledge whether such a string exists? The text now is: "Alice wants to know whether there exists a string $t_i \in T$ that matches q."

A *match* is not defined precisely, which is OK in the case of exact matching using some criteria. But suppose we have some comparison function $f(a, b) \rightarrow \{\text{true}, \text{false}\}$. If f returns true for multiple objects $t \in S$, what is the result of SDA? If the query is repeated, should the same result be received each time (i.e. deterministic behavior) or is the result a random selection from the set of strings which match? The $Match$ function in Section 3 probably wraps this decision inside the function, so both behaviors are OK depending on the $Match$ function definition. This ambiguity can be easily resolved by one sentence in the problem definition discussion, saying that the handling of this case depends on the definition of a match.

4 Comments on Metrics of Interest

Components of strings a, b are not described explicitly.

5 Comments on Considered Models

The subsection begins with "These models differ in ... , (ii) who is the owner of and who possesses T, ...". This is contradicted at end of the paragraph: "In each model, Bob possesses T".

The second to last paragraph does not make sense as is. If Bob does not have access to T, surely Bob cannot know any of its contents. If the text assumes that T is "encrypted" as part of some retrieval method, it is a problem with the method if something is revealed to Bob? If the intent of the paragraph is simply to illustrate why it is not a good idea to simply trust Bob to carry out queries properly and forget the results, and hand over the database in unencrypted form to Bob, the point should be made in more conrete terms.

It might be good to point out why the SSO/SSCO results are better than PIM/PIMPD results. I believe the difference is that in SSO/SSCO it is not a problem if Alice gets to know some information from Bob's calculations, which allows some optimizations.

In this section it seems to be assumed that the alphabet of strings has a finite number of elements. It seems to me that not all of the results depend on that. The alphabet should in any case be introduced in the Introduction where the SDA problem is introduced.

6 Comments on Overview of Results

The section is easy to understand; the table of results increases readability.

7 Comments on In-Depth Look at the Protocols

In Section 5.1 (description of SSO), the description of the construction using a random $(n+3) \times (n+3)$ invertible matric would be improved by giving a toy example with e.g. $n = 2$. Otherwise the text is readable and clear.

Section 5.2 could use some of the clarifications (writing down computation steps) of Section 5.1.

8 Comments on Conclusion

None.

9 Typos and other nits

Typos in Introduction

- "group of object" should be "group of objects"
- italics in "that matches q"
- "qnor" should be "q nor" in SDA definition

Typos in Considered Models

- "examplify" should be "exemplify"

Typos in Overview of results

- "In addition, for the SSO and SSCO models protocols dealing with the $\sum_{i=1}^{n}(a_i - b_i)^2$ is given" is missing "metric" or something similar.

Typos in In-Depth Look at the Protocols

- "..., he interested reader ..." should be "..., the interested reader ...".

References

[1] Matti Järvisalo. Secure Approximate Matching. T-79.514 Special Course on Cryptology / Privacy Preserving Data Mining, 2003.

[2] W. Du and M. Atallah. Protocols for Secure Remote Database Access with Approximate Matching. 7th ACM Conference on Computer and Communications Security (ACM-CCS 2000).