Secure Sketch for Multi-Sets

Ee-Chien Chang[*] Vadym Fedyukovych[†] Qiming Li[‡]

March 15, 2006

Abstract

Given the original set X where $|X| = s$, a sketch P is computed from X and made public. From another set Y where $|Y| = s$ and P, we can reconstruct X if $|X \cap Y| \geq |s - t|$, where $t < s$ is some threshold. The sketch P is secure if it does not reveal much information about X. A few constructions have been proposed, but they cannot handle multi-sets, that is, sets that may contain duplicate elements. We observe that the techniques in the set reconciliation protocol proposed by Minsky et al. [3] can be applied and give a secure sketch that supports multi-sets. If X is a subset of an universe with n elements, the running time of the encoding and decoding algorithms will be polynomial w.r.t. s and $\log n$, and the entropy loss due to the sketch is less than $2t(1 + \log n)$.

Keywords: Secure sketch, set difference, multi-set, error-tolerant cryptography.

1 Introduction

For many biometric and multimedia objects, small changes to the data do not affect the authenticity. Hence, traditional cryptographic schemes cannot be directly applied, since they do not allow even the slightest changes. Secure sketch [1] scheme is proposed to recover the original data from their modified versions with the help of a public sketch. With the original recovered, traditional cryptographic schemes can be employed.

In this paper we consider the following problem. Given the *original* X that is an unordered collection of s elements (which may not be distinct) from a universe $\mathcal{U}$ and a threshold t, we wish to compute a sketch P such that (1) X can be recovered given P and Y, where Y is also a collection of s elements in $\mathcal{U}$ such that X and Y has at least $|s - t|$ elements in common, and (2) the amount of information about X revealed by P (i.e., the *entropy loss*) is small.

[*] Email: changec@comp.nus.edu.sg. Department of Computer Science, National University of Singapore.

[†] Email: vf@unity.net.

[‡] Email: qiming.li@ieee.org. Department of Computer Science, National University of Singapore. The author is currently with Department of Computer and Information Science, Polytechnic University.

There are already a few constructions dealing with set difference (e.g., those in [2] and [1]). However, they cannot support multi-sets. Here we propose a secure sketch scheme for multi-sets. Our construction is similar to the set reconciliation protocol in [3], but the problem settings are different.

The proposed scheme gives a sketch of size at most $2t(1 + \log n)$, where n is the size of the universe. In addition, there exists a simple and yet efficient decoding algorithm – we just need to solve a linear system with $2t$ equations and unknowns and find the roots of two degree t polynomials.

2 Related Work

The fuzzy commitment scheme [4] is one of the first formal approaches to achieve robustness against noises, and it makes use of error-correcting codes to recover changes measured by Hamming distance. The set difference metric is first considered by Juels et al. [2], who give a fuzzy vault scheme. The notions of *secure sketch* and *fuzzy extractor* are introduced by Dodis et al. [1], with several constructions for Hamming distance, set difference, and edit distance. In their framework, the secure sketch is used to recover the original from the corrupted data, which is then used to extract a reliable and almost uniform key that can be used with traditional cryptographic schemes. Dodis et al. [1] give three constructions for set difference, with similar entropy loss. The three constructions differ in the sizes of the sketches, efficiency in computation, and also the ease of implementation in practice. One of the constructions has small sketches and achieves “sublinear” (with respect to the size of the universe) decoding by careful reworking of the standard BCH decoding algorithm. Note that these existing schemes for set difference cannot handle multi-sets (i.e., sets that allow duplicate elements).

Minsky et al. [3] proposed a set reconciliation scheme that provides a way for two physically separated parties to compute the union of their data with small amount of communication. Our techniques are similar to theirs, but the problem settings are different.

3 Notations

A multi-set is an unordered collection of elements from a universe $\mathcal{U}$. The elements in a multi-set are not necessarily distinct. To avoid confusion, we use the notations $\subset_m, -_m, \cap_m$ and $\cup_m$ to denote the subset, asymmetric difference, intersection and union respectively on multi-sets. For example, $\{1, 1, 2, 3, 3\} -_m \{1, 2, 2\}$ gives $\{1, 3, 3\}$. We also use $| \cdot |$ to denote the size of the multi-sets. For instance, $|\{1, 3, 3\}| = 3$. Given two multi-sets X and Y, we say that they are t-close if $|X -_m Y| \leq t$.

A secure sketch scheme for multi-set difference on universe $\mathbb{Z}_n$ with threshold t consists of an encoder Enc and a decoder Dec. Given multi-sets X and Y from $\mathcal{U}$, $\mathsf{Dec}(\mathsf{Enc}(X), Y) = X$ if X and Y are t-close. We call $P = \mathsf{Enc}(X)$ the *sketch.*

To measure the security of such a scheme, we follow the definition of *entropy loss* introduced by Dodis et al. [1]. Let $\mathbf{H}_\infty(A)$ be the min-entropy of random variable A, i.e., $\mathbf{H}_\infty(A) = -\log(\max_a \Pr[A = a])$. For two random variables A and B, the *average min-entropy* of A given B is defined as $\widetilde{\mathbf{H}}_\infty(A|B) = -\log(\mathbb{E}_{b\leftarrow B}[2^{-\mathbf{H}_\infty(A|B=b)}])$. By treating X and P as random variables for the original and the sketch, the entropy loss of X given sketch P is defined as $\mathbf{H}_\infty(X) - \widetilde{\mathbf{H}}_\infty(X|P)$. The average min-entropy has the property that, if B is a random variable of ℓ-bit string, we have $\widetilde{\mathbf{H}}_\infty(A|B) \geq \mathbf{H}_\infty(A) - \ell$. Thus, if the sketch is always no more than ℓ bits, then the entropy loss is at most ℓ.

4 Proposed Scheme

We assume that the universe is $\mathbb{Z}_n$, and the original $X \subset_m \mathbb{Z}_n$, where n is a prime. To handle a special case, we firstly assume that X does not contain any element in $\{0, 1, \ldots, 2t-1\}$, and will discuss how to remove this assumption later at the end of this section.

4.1 The encoder Enc.

Given $X = \{x_1, \ldots, x_s\}$, the encoder does the following.

1. Construct a monic polynomial $p(x) = \prod_{i=1}^{s}(x - x_i)$ of degree s.
2. Publish $P = \langle p(0), p(1), \ldots, p(2t-1)\rangle$.

4.2 The decoder Dec.

Given $P = \langle p(0), p(1), \ldots, p(2t-1)\rangle$ and $Y = \{y_1, \ldots, y_s\}$, the decoder follows the steps below.

1. Construct a polynomial $q(x) = \prod_{i=1}^{s}(x - y_i)$ of degree s.
2. Compute $q(0), q(1), \ldots, q(2t-1)$.
3. Let $p'(x) = x^t + \sum_{j=0}^{t-1} a_j x^j$ and $q'(x) = x^t + \sum_{j=0}^{t-1} b_j x^j$ be monic polynomials of degree t. Construct the following system of linear equations with the a_j's and b_j's as unknowns.

$$q(i)p'(i) = p(i)q'(i), \quad \text{for } 0 \leq i \leq 2t-1 \tag{1}$$

4. Find one solution for the above linear system. Since there are $2t$ equations and $2t$ unknowns, such a solution always exists.
5. Solve for the roots of the polynomials $p'(x)$ and $q'(x)$. Let them be X' and Y' respectively.
6. Output $\widetilde{X} = (Y \cup_m X') -_m Y'$.

The correctness of this scheme is straight forward. When there is exactly t replacement errors, we can view $p'(x)$ as the "missed" polynomial whose roots are in $X' = X -_m Y$. Similarly, $q'(x)$ is the "wrong" polynomial, whose roots are in $Y' = Y -_m X$. Since the roots of $p(x)$ and $q(x)$ are in X and Y respectively, we have $q(x)p'(x) = p(x)q'(x)$. This interpretation motivates the equation (1).

When there are less than t replacement errors, there will be many degree t monic polynomials $p'(x)$ and $q'(x)$ that satisfy $q(x)p'(x) = p(x)q'(x)$. For any such $p'(x)$ and $q'(x)$, they share some common roots, which could be some arbitrary multi-set Z. That is, $X' = (X -_m Y) \cup_m Z$, and $Y' = (Y -_m X) \cup_m Z$. In Step 6, this extra Z will be eliminated.

When $X \cap_m \{0, \ldots, 2t-1\} \neq \emptyset$, some equations in (1) would degenerate, which makes the rank of the linear system less than $2t$. In this case, it is not clear how to find the correct polynomial in the solution space. Hence we require that $X \cap_m \{0, \ldots, 2t-1\} = \emptyset$.

Note that in the above we do not require the elements of X and Y to be distinct, so this scheme can handle multi-sets. Furthermore, since the size of each $p(i)$ for $1 \leq i \leq 2t$ is $(\log n)$, the size of P is $2t(\log n)$. Therefore, we have:

THEOREM 1 *When $X \cap_m \{0, \ldots, 2t-1\} = \emptyset$ and n is prime, the entropy loss due to $\mathsf{Enc}_s(X)$ is at most $2t \log n$.*

4.3 Removing the assumption on X and Y.

The assumption that X cannot contain any element from $\{0, \ldots, 2t-1\}$ can be easily relaxed. We can find the smallest prime m such that $m - n \geq 2t$, and then apply the scheme on $\mathbb{Z}_m$. But instead of publishing $p(0), \ldots, p(2t-1)$, we publish $p(m-1), \ldots, p(m-2t)$. In this way, the size of the sketch is $2t \log m$. In practice, this is not a problem since the size of the universe may not be prime, and we will need to choose a larger finite field anyway. For t that is not too large (say, $t \leq n/4$), we can always find at least one prime in $[n+2t, 2n]$. Hence, we have the

COROLLARY 2 *When $t \leq n/4$, the entropy loss due to $\mathsf{Enc}_s(X)$ is at most $2t(1 + \log n)$.*

Note that the decoding amounts to solve a system of $2t$ linear equations, and finding the roots of a polynomial of degree at most t. Hence, the number of arithmetic operations on $\mathbb{Z}_n$ needed during decoding is bounded by a polynomial of s and t.

References

[1] Y. Dodis, L. Reyzin, A. Smith, Fuzzy extractors: How to generate strong keys from biometrics and other noisy data, in: Eurocrypt, Vol. 3027 of LNCS, Springer-Verlag, 2004, pp. 523–540.

[2] A. Juels, M. Sudan, A fuzzy vault scheme, in: IEEE Intl. Symp. on Information Theory, 2002.

[3] Y. Minsky, A. Trachtenberg, R. Zippel, Set reconciliation with nearly optimal communications complexity, in: IEEE Intl. Symp. on Information Theory, 2001.

[4] A. Juels, M. Wattenberg, A fuzzy commitment scheme, in: Proc. ACM Conf. on Computer and Communications Security, 1999, pp. 28–36.