Secure Sketch for Multi-Set Difference

Ee-Chien Chang* Vadym Fedyukovych† Qiming Li‡

Abstract

Recently there has been active research on secure sketch schemes. The current constructions with the set difference as the distance metric cannot be extended to multi-sets. In this paper, we give an efficient secure sketch scheme for multi-sets, with entropy loss and sketch size similar to previous constructions.

keywords: Secure sketch, set difference, multi-set.

1 Introduction

In some applications, small changes to the data do not affect their authenticity. This cannot be done with traditional cryptographic schemes since they do not allow even the slightest changes. The fuzzy commitment scheme [3] 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 security level. 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. None of these existing schemes for set difference can be extended to handle multi-sets (i.e., sets that allow duplicated elements) with minor modifications.

In this paper, we consider secure sketch for multi-sets, where the distance metric is the difference between multi-sets, which we will define formally later.

*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 and Information Science, Polytechnic University. This work was done when the author was in Department of Computer Science, National University of Singapore.

Our construction is similar to the set reconciliation protocol in [4], 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, and t is the number of errors we want to tolerate. 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 Notations

Let $\mathcal{U}$ be the universe. In the rest of this paper, we assume that $\mathcal{U} = \{0, \ldots, n-1\}$ to be a set of n distinct integers. We assume that the original data X is an enssemble of s elements, and we write $X = \{x_1, \ldots, x_s\}$, where $x_i \in \mathcal{U}$ for all $1 \leq i \leq s$. Note that the elements in X are not necessarily distinct. We call such X a multi-set, and write $X \subseteq_m \mathcal{U}$. For two multi-sets of the same size, we define their distance as below.

Definition 1 *For any $X = \{x_1, \ldots, x_s\}$ and $Y = \{y_1, \ldots, y_s\}$ such that $X, Y \subseteq_m \mathcal{U}$, the multi-set difference between X and Y is*

$$D(X, Y) = s - \max_f |\{i \mid x_i = y_{f(i)}\}| \tag{1}$$

for all one-to-one corresponse f on $\{1, \ldots, s\}$.

In other words, we find the maximum match between X and Y, and the difference is the number of elements that do not match. We then further define the "closeness" between X and Y as below.

Definition 2 *For two multi-sets X and Y, we say that they are close if $D(X, Y)$ is defined, and $D(X, Y) \leq t$ for some threshold t.*

A secure sketch scheme with universe $\mathcal{U}$ and threshold t consists of an encoder Enc and a decoder Dec, such that given multi-sets X and Y from $\mathcal{U}$, $\mathsf{Dec}(\mathsf{Enc}(X), Y) = X$ if X and Y are 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 $\widetilde{\mathbf{H}}_\infty(A|B) = -\log(\mathbb{E}_{b \leftarrow B}[2^{-\mathbf{H}_\infty(A|B=b)}])$. This definition is useful in the analysis, since for any ℓ-bit string B, we have $\widetilde{\mathbf{H}}_\infty(A|B) \geq \mathbf{H}_\infty(A) - \ell$.

We say that the sketch scheme is m-secure if for all random variable X from $\mathcal{U}$, the entropy loss of P is at most m. That is, $\mathbf{H}_\infty(X) - \widetilde{\mathbf{H}}_\infty(X \mid \mathsf{Enc}(X)) \leq m$.

3 Proposed Scheme

To handle a special case, we 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.

3.1 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 $P = \langle p(0), p(1), \ldots, p(2t-1)\rangle$.

3.2 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.

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

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 $\widetilde{X} = (Y \cup X') \setminus Y'$.

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 $X' = X \setminus Y$. Similarly, $q'(x)$ is the "wrong" polynomial, whose roots are in $Y' = Y \setminus X$. 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 (2).

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, $X' = (X \setminus Y) \cup Z$, and $Y' = (Y \setminus X) \cup Z$. In Step 6, this extra Z will be eliminated.

When $X \cap \{0, \ldots, 2t-1\} \neq \emptyset$, some equations in (2) 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 $X \cap \{0, \ldots, 2t-1\} = \emptyset$.

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 the

THEOREM 3 *When $X \cap \{0, \ldots, 2t-1\} = \emptyset$, the entropy loss due to $\mathsf{Enc}_s(X)$ is at most $2t \log n$.*

3.3 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 4 *When $t \leq n/4$, the entropy loss due to $\mathsf{Enc}_s(X)$ is at most $2t(1+\log n)$.*

References

[1] Yevgeniy Dodis, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In *Eurocrypt'04*, volume 3027 of *LNCS*, pages 523–540. Springer-Verlag, 2004.

[2] Ari Juels and Madhu Sudan. A fuzzy vault scheme. In *IEEE Intl. Symp. on Information Theory*, 2002.

[3] Ari Juels and Martin Wattenberg. A fuzzy commitment scheme. In *Proc. ACM Conf. on Computer and Communications Security*, pages 28–36, 1999.

[4] Yaron Minsky, Ari Trachtenberg, and Richard Zippel. Set reconciliation with nearly optimal communications complexity. In *ISIT*, 2001.