Secure Sketch: Correcting More Errors Without Entropy Loss

Yen-Lung Lai, Zhe Jin

Monash University Malaysia,
Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor
yenlung.lai@monash.edu, jin.zhe@monash.edu

Abstract. Secure sketch produces public information of its input w without revealing it, yet, allows the exact recovery of w given another value w' that is close to w. Therefore, it can be used to reliably reproduce any error-prone secret sources (i.e., biometric) stored in secret storage. However, some sources have lower entropy compared to the error itself, formally called "more error than entropy", a standard secure sketch cannot show its security promise perfectly to these kinds of sources. Besides, when same input is reused for multiple sketches generation, the complex error process of the input further results to security uncertainty, and offer no security guarantee. Fuller et al., (Asiacrypt 2016) defined the fuzzy min-entropy is necessary to show security for different kind of sources over a family of distributions. This paper focuses on secure sketch. We propose a new technique to generate re-usable secure sketch. We show security to low entropy sources and enable error correction up to Shannon bound via unique decoding. Our security defined information theoretically with Shannon entropy over some worst case random error distribution adding to the input source. In particular, our new technique offers security guarantee for all input distributions with min-entropy at least one bit.

Keywords: Secure Sketch · Error Correction · Fuzzy Extractor · Information Theory

1 Introduction

Traditional cryptography systems rely on uniformly distributed and recoverable random strings for secret. For example, random passwords, tokens, and keys, all are commonly used secrets for deterministic cryptographic applications, i.e., encryption/decryption and password authentication. These secrets must present exactly on every query for a user to be authenticated and get accessed into the system. Besides, it must also consist of high enough entropy, thus making it very long and complicated, further resulted in the difficulty in memorizing it. On the other hand, there existed plentiful non-uniform strings to be utilized for secrets in practice. For instance, biometrics (i.e., human iris, fingerprint) which can be used for human recognition/identification purpose. Similarly, long passphrase

(S. N. Porter, 1982 [1]), answering several questions for secure access (Niklas Frykholm *et al.*, 2001 [2]) or personal entropy system (Ellison *et al.*, 2000 [3]), and list of favorite movies (Juels and Sudan, 2006 [4]), all are non-uniformly distributed random strings that can be utilized for secrets.

As a solution by utilizing non-uniform input for secrets, it raises several security and practicability concerns. Firstly, since it is *not truly random and uniform*, this increased the risk where an adversary may easily be guessed and compromised it, thus reveals the underlying secret. Secondly, most of the available non-uniform strings are *not exactly recoverable*. Therefore, they cannot be used for a typical deterministic cryptographic application. For instance, human biometric data, it is well understood that two biometric readings sourced from the same individual are rarely to be identical. Additionally, precise answer to multiple questions or entering a password through keyboard consistently, from time to time, would be a challenge for human memory although the provided answers are likely to be similar.

Nevertheless, these non-uniform measurements that always selected by human or naturally existing are believed to offer a higher entropy than human-memorable password. Especially, higher security level can be achieved by using longer/more complex human biological measurements, i.e., fingerprint, voice, retina scan, handwriting signature, and others. (N. Frykholm, 2000 [2]), (Jain *et al.*, 2016 [5]). Most importantly, it is memory-free and somewhat difficult to steal, or loss compared to using external key storage, e.g., smart card, token, keys.

The availability of non-uniform information prompted the generation of uniform random string from non-uniform materials. Started by Bennette *et al.*, (1988) [6], identified two major approaches to derive a uniform string from noisy non-uniform sources. The first approach is *information-reconciliation*, by tolerating the errors in the sources without leaking any information. The second approach refers to the *privacy amplification*, which converts high entropy input into a uniform random input. The information-reconciliation process can be classified into interactive (includes multi messages) and non-interactive (only includes single message) versions. For non-interactive line of work, it has been first defined by Dodis *et al.*, (2004) [7] called the fuzzy extractor. Likewise, the fuzzy extractor used two approaches to accomplish the task, which is the secure sketch - for error tolerance, and randomness extractor - for uniform string generation.

In this paper, we only focus on the secure sketch. Secure sketch is more demanding because it allows information-reconciliation, e.g., exact recovery of a noisy secret while offering security assurance to it. Moreover, a secure sketch can be easily extended to fuzzy extractor for uniform string generation by using a randomness extractor. There existing various secure sketch constructions in the literature. Some notable constructions involved the code-offset construction proposed by Juels and Wattenberg (1999) [8] that operates perfectly over hamming matric space. This work generates a sketch through encoding a uniform string with error correction code, then leaving an offset via performing XOR operation with a noisy string. The uniform string can be reproduced by another noisy string by means of error tolerance, provided the error level is lower than a specified threshold. Besides, Juels and Sudan (2006) [4] have also proposed another construction for metric other than hamming called the fuzzy vault. An improved version of the fuzzy vault is proposed by Dodis *et al.*, (2004) [7], and also the Pin-sketch that relies on syndrome encoding/decoding with t-error correcting BCH code $\mathcal{C}$, which works well for non-fixed length input over a universe $\mathcal{U}$.

1.1 Existing Issues in Secure Sketch

We here review some existing issues in the secure sketch.

More error than entropy: The secure sketch must contain some information about the sources to tolerate the errors. More generally, given a point (some value) w, the sketch would allow the acceptance of its nearby point w' within distance t. Therefore, if an adversary can predict an accepting w' with noticeable probability, the sketch must reveal w to the adversary with noticeable probability as well. The tension between the security and error tolerance capability is very strong. Precisely, the security is measured in term of the residual (min-) entropy, which is the starting entropy of w minus the entropy loss. Often, a larger tolerance distance is needed to tolerate more errors. However, exercising larger tolerance distance will offer greater advantages to the adversary in predicting w'. In the end, the residual entropy becomes lower by the increment of t. This consequent to an upper bound of the tolerance distance translated to a lower bound on the entropy loss of the input sources. This event is much worsening for some non-uniform sources with low min-entropy, especially, when the sources consist of *more error than entropy* itself. Since the source entropy rate is lower than the error rate, simply deducting the entropy loss from the sources' min-entropy always output a negative value, hence, show no security. One typical example of a source with more error than entropy refers to the commonly known biometric feature - IrisCode (Daugman, 2006) [9]. The IrisCode is said to provide entropy of 249 bits. Whereas, the IrisCodes generated from the same user of each 2048 bits have shown far more than 249 bits of errors. Therefore, this more error than entropy problem is indeed restricting the usage of a secure sketch from all kind of available sources.

Distribution uncertainty: The predictability of nearby point w' within distance t is not merely entropically connected, but it is also closely tied to the distribution of the sources. A source can be described using a family of distributions $\mathcal{W} = \{W_1, \ldots, W_\gamma\}$. Given a source under a random distribution $W \in \mathcal{W}$ where all points are far apart, the probability for an adversary to predict any nearby point $w' \in W'$ within distance t will be small. The entropy loss of the sketch would be bounded that is proportional to t. In particular, given a source with min-entropy m, a larger distance between the points implies higher min-entropy, thus, the entropy loss due to error tolerance over distance t can be compensated by the high min-entropy. This entropy loss is crude if one has set $t > m$ for error tolerance over distance t (e.g., more error than entropy).

Fuller *et al.*, (2013) [10] showed that under the event when the input distribution is precisely known, the crude entropy loss can be avoided by the measurement of *fuzzy min-entropy*, which defined as the min-entropy with maximized chances for a variable of W within distance t of w':

$$\mathrm{H}^{\text{fuzz}}_{t,\infty}(W) \overset{\text{def}}{=} -\log\left(\max_{w'} \Pr[W \in B_t(w')]\right)$$

where $B_t(w')$ denoted a hamming ball of radius t around w'. Conceivably, the fuzzy min-entropy is equivalent to the residual entropy, which is bounded by the min-entropy $\mathrm{H}_\infty(W) - \log(B_t(w')) \leq \mathrm{H}^{\text{fuzz}}_{t,\infty}(W)$ minus the loss signified by the hamming ball $B_t(w')$ of radius t.

Realistically, it is imprudent to assume the source distribution is precisely known, especially for high entropy sources. The adversary may have higher computation power to model and exam the distribution compared to the designer. Such event always refer to the *distribution uncertainty*, where the fuzzy min-entropy notion is necessary and sufficient only when the security is defined computationally. Viewed this way, one cannot assure information theoretical security without precise knowledge over the input worst-case distribution W (i.e., distance between points is minimum).

Reusability[1] Reusability property is introduced by Boyen (2004) [13]. Given a user comes with a noisy input w (i.e., biometric), the user may enroll w for different applications. Each time the user enrolls using w, he/she must provide slightly different reading w_i due to the error. Therefore, different sketches ss_i and keys R_i can be generated for different applications respectively. The security property of *individual* sketches and keys should hold with all existing sketches $ss_1, ss_2, \ldots, ss_\gamma$. In fact, this property has been well studied for current constructions of secure sketch and fuzzy extractor, but many of them do not satisfied reusability [13] [14] [15] [16].

1.2 Our Contributions

We highlighted our main contributions as follow:

Average fuzzy min-entropy: To correct more errors, larger error tolerance distance is desired. Unfortunately, larger tolerance distance renders higher probability of success in predicting w' within more considerable distance around w. Thus, security diminution cannot be avoided. Our new result has considered the notion of *average fuzzy min-entropy*, which is basically the fuzzy min-entropy with different error tolerance distances.

To be more precise, consider different variable Φ and Φ'. To allow error tolerance within a larger distance $t > t'$, one must maximize the total probability mass of Φ with larger ball $B_t(\phi')$[2] around the string $\phi' \in \Phi'$. Suppose Φ is

[1] The reusability property is different to the unlinkability property [11] [12]. Unlinkability property prevents an adversary from differentiating whether two enrollments correspond to the same physical source, which is not focused in this work.

[2] Sometime, we omit ϕ' or w' to describe the ball B_t or $B_{t'}$, when they are not depend upon their center ϕ' and w' respectively

correlated with some variable W, if the adversary finds out $W \notin B_{t'}(w')$, then the predictability of Φ becomes $\mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \notin B_{t'}(w')\,]\right]$. On average, the average fuzzy min-entropy is:

$$\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t,\infty}(\Phi|W \notin B_{t'}(w')) \overset{\mathrm{def}}{=} -\log\left(\mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \notin B_{t'}(w')\,]\right]\right)$$

Intuitively, average fuzzy min-entropy refers to the fuzzy min-entropy of some variable Φ defined by a larger hamming ball B_t, where only the points outside an existing smaller ball $B_{t'}$ are considered. Substantial fuzzy min-entropy implies more errors can be corrected over larger tolerance distance $t > t'$, hence, higher entropy loss. In this sense, average fuzzy min-entropy $\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t,\infty}(\Phi|W \notin B_{t'}(w'))$ reveals the entropy loss from the fuzzy min-entropy of W over smaller tolerance distance t'. Since the quantity of entropy loss must lower bounded to the entropy of the source to show meaningful security, the average fuzzy min-entropy manifested the minimum entropy loss from a source over the worst case t' (minimum t'). In view of this, the average-fuzzy min-entropy is useful for better monitoring the loss of the min-entropy while providing optimal resilience.

Unique-decoding toward Shannon bound: Our construction uses two error correction codes, one act as 'inner' code $\mathcal{C}^*$ to encode a noisy input string and yield the codeword $c^* \in \mathcal{C}^*$; another one act as 'outer' code to encode the noisy version of c^* and yield another codeword $c \in \mathcal{C}_\xi$. In addition to this, we adopted the principle of *Locality Sensitive Hashing (LSH)* to generate a resilient vector pair (trivially, a pair of longer strings with resilience property) for sketching and recovery. The resilient vector resembles additional random error adding to the codeword c encoded by a the 'outer' code $\mathcal{C}_\xi$. We show that in such particular setting, one is able to uniquely decode the corrupted codewords, and correcting total error rate up to Shannon bound.

Security bound independent to the input length: Info. theoretic secure sketch is always desired. Because it does not introduce additional assumption of computational limits to the attacker, thus offers better security assurance. Most importantly, info. theoretic secure sketch eliminates the distribution uncertainty issue by showing security to all family of input distribution via min-entropy measurement. Notwithstanding its security robustness, the cost imposed by info. theoretic secure sketch to the source entropy requirement is too high, which is at least half of the input length itself [17]. It means that if the entropy is less than half of its input length, it achieves nothing where the underlying secret can be easily revealed due to exhaustive entropy loss caused by error tolerance. We constructed a pair of sketching and recover algorithm. Formal correctness and security proof have been given to show that it satisfies the properties of an info. theoretical secure sketch. In addition to this, the new construction is capable of achieving security bound that merely depend upon the Shannon entropy of some worst case random error distribution rather than its input length.

Reusable secure sketch: Apart from this, the new construction offers extra security property, which is the reusability. We show that the error included implicitly allows reusability. We defined our reusability in information theoretical sense, with a group of computational unbounded adversaries. Our results imply the flexibility of independent re-enrollment of a single source with multiple providers, yet offer security assurance to each of them, as long as the error is kept within specified range. Our reusability emphasizes the case when the providers are not communicating with each other hence it supports security to all of them individually.

1.3 Our Technique

Some notation need to know: This work focus on binary hamming metric where $\mathcal{M}_1 = \{0,1\}^{k^*}$, and $\mathcal{M}_2 = \{0,1\}^n$ denoted two different sizes of metric spaces with $n > k^*$. The distance between different binary string w and w' is the binary hamming distance (e.g., the number of disagree elements) denoted as $\|w \oplus w'\|$ where $\|.\|$ is the hamming weight that count the number of non-zero elements, and $\oplus$ is the addition modulo two operation (XOR). Besides, the error rate in between the input w $w' \in \mathcal{M}_1$ is denoted as $\|w \oplus w'\| \, (k^*)^{-1}$ which is simply their normalized hamming distance. For error correction code notation, since we are more interested in tolerating the errors of a codeword c' instead of its min-distance d, we used t instead of d to explicitly represent an $[n,k,t]_2$ binary code $\mathcal{C}_\xi$ with the tolerance rate denoted as $\xi = t/n$, where $\xi \in (0, 1/4)$. Besides, we also refer to another $[n^*, k^*, t^*]_2$ binary code $\mathcal{C}^*$ with $k^* < n^* \leq k < n$. In particular, we focus on the case when both $\mathcal{C}^*$ and $\mathcal{C}_\xi$ are come with min-distance $d^* \geq 2t^* + 1$ and $d \geq 2t + 1$ respectively. At the same point, we let $t_{(+)} = \lfloor (\xi + \epsilon) k^* \rfloor$ and $t_{(-)} = \lceil (\xi - \epsilon) k^* \rceil$ to describe two different error tolerance distances over the smaller binary metric space $\{0,1\}^{k^*}$, with some error parameter $\epsilon > 0$.

Overview idea: Suppose Alice wishes to conceal a noisy non-uniform string $w \in \{0,1\}^{k^*}$ while allows exact recovery of w from another noisy string $w' \in \{0,1\}^{k^*}$ that is close to w. Then, Alice has to generate a secure sketch which able to tolerate the error in w'. To do so, we invoke the uses of error correction code for conventional secure sketch generation. Our scheme requires two error correction codes, an $[n^*, k^*, t^*]_2$ 'inner' code $\mathcal{C}^*$ is chosen over $\{0,1\}^{k^*}$, and another $[n, k, t]_2$ 'outer' code $\mathcal{C}_\xi$ is chosen over $\{0,1\}^n$, where $k^* < n^* \leq k < n$. Firstly, Alice encodes w using the 'inner' code $\mathcal{C}^*$ to output a codeword c^*. Then, c^* is use to generate a noisy string $v^* \in \{0,1\}^{n^*}$ with w and error e, yielding $v^* = c^* \oplus (0^{n^*-k^*} \| w)$ (with zeros padding at front of w). After that, v^* is being encoded by the 'outer' code $\mathcal{C}_\xi$ to output the final codeword $c \in \mathcal{C}_\xi$. Alice then conceals c by generating a sketch $ss = c \oplus \delta$ which is then made public and leaving the offset δ in the clear. The offset δ is characterized by a pair of resilient vectors $\phi, \phi' \in \{0,1\}^n$, which is generated from a pair of noisy strings $w'_e, w_e \in \{0,1\}^{k^*}$, i.e., $w_e = w \oplus e$ through LSH. The resilient vectors offer resilience for the recovery of w from w' if $\|\delta\| \leq t$ and $\|w \oplus w'\| \leq t^*$.

Likewise the code-offset construction [8], our idea is conceptual simpler but comes with some crucial differences in term of operations. Firstly, the code-offset construction concealing a random and uniform string (called as the witness of w) and involved only single encoding stage; our construction concealing a non-uniform input w that has gone through two different encoding stages with $\mathcal{C}^*$, and $\mathcal{C}_\xi$. Secondly, despite the code-offset construction does not limit to particular types of error correction code (i.e., not necessary to be linear), the sketch size is always bounded by the size of the input w. Comparatively, in our case, Alice is free to choose any error correction code where the sizes of the concealed object and output sketch are not bounded but parametrized by the selected $[n^*, k^*, t^*]_2$ code $\mathcal{C}^*$ and the $[n, k, t]_2$ code $\mathcal{C}_\xi$. Thirdly, of course, our operation comes with additional random errors adding to the input w and w' during sketching and recovery.

For resilient vector generation, we only focus on a particular LSH family called hamming-hash [18]. The hamming hash is considered as one of the easiest ways to construct an LSH family by bit sampling technique. Since it will be a core element in our proposal, it is worth sketching in details on how it works.

Hamming hash strategy. *Let $[k^*] = \{1, \ldots, k^*\}$. For Alice with $w \in \{0,1\}^{k^*}$ and Bob with $w' \in \{0,1\}^{k^*}$. Alice and Bob agreed on this strategy as follow:*

1. *They are told to each other a common random integer $N \in [k^*]$.*
2. *They separately output '0' or '1' depend upon their private string w and w', i.e., Alice output '1' if the N-th bit of w is '1', else output '0'.*
3. *They win if they got the same output, i.e., $w(N) = w'(N)$.*

Based on above strategy, we are interested in the probability for Alice and Bob output the same value which can be described with a similarity function $S(w, w') = P$ with probability $P \in [0, 1]$.

Theorem 1. *Hamming hash strategy is a LSH with similarity function $S(w, w') = 1 - \|w \oplus w'\|(k^*)^{-1}$*

Theorem 1 concluded that Alice and Bob always win with probability described as $P = 1 - \|w \oplus w'\|(k^*)^{-1}$. Observe that, the similarity function for hamming hash correspond to the hamming distance between w and w'.

By repeat step 1 and step 2 of hamming hash strategy n times, with different random integers, Alice and Bob able to output a n bits string $\phi, \phi' \in \{0,1\}^n$ respectively, which we have earlier named as *resilient vectors*.

Theorem 2. *Suppose two resilient vectors $\phi, \phi' \in \{0,1\}^n$ are generated from $w, w' \in \{0,1\}^{k^*}$ respectively by hamming hash strategy with string $N \in [k^*]^n$, then $\mathbb{E}[\|\phi \oplus \phi'\|] = n\,\|w \oplus w'\|\,(k^*)^{-1}$.*

Proof. Let $\|\delta\| = \|\phi \oplus \phi'\|$, base on Theorem 1, we know that, for each time in comparing the hamming hash output (for $i = 1, \ldots, n$), the probability of disagree is describe as:

$$\Pr[\phi(i) \neq \phi'(i)] = \|w \oplus w'\|\,(k^*)^{-1} = 1 - P$$

Therefore, one has i.i.d variable (or Bernoulli variable) for each offset element, $\delta(i)=1$ if $\phi(i)\neq\phi'(i)$ and $\delta(i)=0$ if $\phi(i)=\phi'(i)$. Precisely, $\|\delta\|=\|\phi\oplus\phi'\|=\sum_{i=1}^{n}\delta(i)$, thus, $\|\delta\|\sim\mathrm{Bin}(n,1-P)$ follows binomial distribution of expected distance $\mathbb{E}[\|\delta\|]=n(1-P)$ and s.d. $\sigma=\sqrt{nP(1-P)}$. Hence, $\mathbb{E}[\|\delta\|]=n(1-P)=n\,\|w\oplus w'\|\,(k^*)^{-1}$ and prove the theorem.

Theorem 2 concluded that, any changes in the input hamming distance $\|w\oplus w'\|$ can be described as an Bernoulli variable corresponds to the offset elements $\delta(i)$. Therefore, by introducing additional error $e\in\{0,1\}^{k^*}$ of weight $\|e\|=\lfloor k^*\epsilon\rfloor$ to the inputs, where $\epsilon>0$ (e.g., adding the error simply equivalent to $\|w\oplus w'\oplus e\|$), the probability of disagreeing for each element between the resilient vectors ϕ,ϕ' must increase or decrease by ϵ, which can be described as $1-P\pm\epsilon$.

To make the above argument more precise, we provide the following corollaries to characterize the effect on the offset $\|\delta\|$ with ϵ. In our studies, we always refer to the resilient vectors generated from LSH hamming using the same random integer string $N\in[k^*]^n$. The corollaries are given as follow.

Corollary 1. *Given some random variables $W,W'\in\{0,1\}^{k^*}$, $\Phi,\Phi'\in\{0,1\}^n$ and an error parameter $\epsilon>0$. Let $\xi=t/n$ be the tolerance rate of a $[n,k,t]_2$ code $\mathcal{C}_\xi$. Suppose a resilient vector $\phi'\in\Phi'$ is generated from strings $w'\in W'$. For two hamming ball $B_t(\phi')$ and $B_{t_{(-)}}(w')$ of radius $t_{(-)}=\lceil(\xi-\epsilon)k^*\rceil$ and $t>t_{(-)}$, if $W\in B_{t_{(-)}}(w')$, then, one has the minimum probability to find any variable $\Phi\in B_t(\phi')$ described as $1-\exp(-2n\epsilon^2)$.*

Proof. For $W\in B_{t_{(-)}}(w')$, it means that for all $w\in W$, $\|w\oplus w'\|\leq t_{(-)}$ and so $\|w\oplus w'\|(k^*)^{-1}\leq\xi-\epsilon$ is always true. Multiplying both sides of the inequality with n, then $n\|w\oplus w'\|(k^*)^{-1}\leq n\xi-n\epsilon$ and yield $t\geq\mathbb{E}[\|\delta\|]+n\epsilon$ (by Theorem 2). Therefore, one could have a $t_{\min}$ s.t. $t_{\min}=\mathbb{E}[\|\delta\|]+n\epsilon$. By using *Hoeffding's inequality*, the average probability can be expressed as:

$$\begin{aligned}&\mathbb{E}_{w'\leftarrow W}\left[\min_{\phi'}\Pr\left[\Phi\in B_t(\phi')\;\middle|\;W\in B_{t_{(-)}}(w')\right]\right]\\&\geq\min_{t=t_{\min}}\Pr\left[\|\delta\|\leq t\;\middle|\;\|w\oplus w'\|\leq t_{(-)}\right]=1-\exp(-2n\epsilon^2)\end{aligned}\tag{1}$$

and complete the prove.

Corollary 2. *Given some random variables $W,W'\in\{0,1\}^{k^*}$, $\Phi,\Phi'\in\{0,1\}^n$ and an error parameter $\epsilon>0$. Let $\xi=t/n$ the tolerance rate of a $[n,k,t]_2$ code $\mathcal{C}_\xi$. Suppose a resilient vector $\phi'\in\Phi'$ is generated from strings $w'\in W'$. For two hamming ball $B_t(\phi')$ and $B_{t_{(+)}}(w')$ of radius $t_{(+)}=\lfloor(\xi+\epsilon)k^*\rfloor$ and $t>t_{(+)}$ respectively, if $W\notin B_{t_{(+)}}(w')$, then, one has the maximum probability to find any variable $\Phi\in B_t(\phi')$ described as $\exp(-2n\epsilon^2)$.*

Proof. This proof is instantiated from the proof of Corollary 1. For $W\notin B_{t_{(+)}}(w')$, it means that for all $w\in W$, $\|w\oplus w'\|\geq t_{(+)}$, $\|w\oplus w'\|(k^*)^{-1}\geq\xi+\epsilon$ is always true. Multiplying both sides of the inequality with n, then $n\|w \oplus w'\|(k^*)^{-1} \geq n\xi + n\epsilon$ and yield $t \leq \mathbb{E}[\|\delta\|] - n\epsilon$ (by Theorem 2). Therefore, one could have a $t_{\max}$ s.t. $t_{\max} = \mathbb{E}[\|\delta\|] - n\epsilon$. By using *Hoeffding's inequality*, the average probability can be expressed as (by symmetry):

$$\begin{aligned}&\mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr\left[\Phi \in B_t(\phi') \;\middle|\; W \notin B_{t_{(+)}}(w')\,\right]\right]\\ &\leq \max_{t=t_{\max}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \geq t_{(+)}\,\right] = \exp\left(-2n\epsilon^2\right)\end{aligned} \tag{2}$$

and complete the prove.

The results obtained from Corollary 1 and Corollary 2 imply the following statement: Once the error of parameter $\epsilon > 0$ is introduced to the input, the probability of finding any resilient vector $\phi \in \Phi$ close to $\phi' \in \Phi'$ with in the ball $B_t(\phi')$ will be bounded. These bounds are conditioned on the input W, whether $W \in B_{t_{(-)}}(w')$ or $W \notin B_{t_{(+)}}(w')$, that can be proven in either way by minimizing/maximizing the value of $t = t_{\min}/t_{\max}$ respectively. Accordingly, we have the computed numerical bound for average fuzzy min-entropy described as

$$\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t,\infty}(\Phi|W \notin B_{t_{(+)}}(w')) \geq -\log(\exp(-2n\epsilon^2)) \tag{3}$$

2 Preliminaries

In this section, we briefly highlight and recall some classical notions used in our constructions.

Metric Spaces: A metric space defined $\mathcal{M}$ as finite set along with a distance function $\mathsf{dis} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+ = [0, \infty)$, that takes any non-negative real values and obey symmetric e.g., $= \mathsf{dis}(A, B) = \mathsf{dis}(B, A)$, and triangle inequality, e.g., $\mathsf{dis}(A, C) \leq \mathsf{dis}(A, B) + \mathsf{dis}(B, C)$.

Min-Entropy: For security, one is always interested in the probability for an adversary to predict a random value, i.e., guessing a secret. For a random variable W, $\max_w \Pr[W = w]$ is the adversary's best strategy to guess the most likely value, also known as the predictability of W. The min-entropy thus defined as

$$\mathrm{H}_\infty(W) = -\log\left(\max_w \Pr[W = w]\right)$$

min-entropy also viewed as worst case entropy.

Average min-entropy: Given pair of random variable W, and W' (possible correlated), given an adversary find out the value w' of W', the predictability of W is now become $\max_w \Pr[W = w \mid W' = w'\,]$. The average min-entropy of W given W' is defined as

$$\tilde{\mathrm{H}}_\infty(W|W') = -\log\left(\mathbb{E}_{w' \leftarrow W'}\left[\max_w \Pr[W = w \mid W' = w'\,]\right]\right)$$

Fuzzy min-entropy: Given an adversary try to find w' that is within distance t of w, the *fuzzy min-entropy* is the total maximized probability mass of W within the ball $B_t(w')$ of radius t around w defined as:

$$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) = -\log\left(\max_{w'} \Pr[W \in B_t(w')]\right)$$

high fuzzy min-entropy is a necessary for strong key derivation.

Secure sketch [7] An $(\mathcal{M}, m, \tilde{m}, t)$-secure sketch is a pair of randomized procedures “sketch” (SS) and “Recover” (Rec), with the following properties:

SS: takes input $w \in \mathcal{M}$ returns a secure sketch (e.g., helper string) $ss \in \{0,1\}^*$.

Rec: takes an element $w' \in \mathcal{M}$ and ss. If $\mathsf{dis}(w,w') \leq t$, then $\mathsf{Rec}(w', ss) = w$ with probability $1-\beta$, where β is some negligible quantity. If $\mathsf{dis}(w,w') > t$, then no guarantee is provided about the output of Rec.

The security property of secure sketch guarantees that for any distribution W over $\mathcal{M}$ with min-entropy m, the values of W can be recovered by the adversary who observes ss with probability no greater than $2^{-\tilde{m}}$. That is the residual entropy $\tilde{\mathrm{H}}_\infty(W|W') \geq \tilde{m}$.

Error correction code [19]: Let $q \geq 2$ be an integer, let $[q] = \{1, \ldots, q\}$, we called an $(n,k,d)_q$-ary code $\mathcal{C}$ consist of following properties:

- $\mathcal{C}$ is a subset of $[q]^n$, where n is an integer referring to the *blocklength* of $\mathcal{C}$.
- The *dimension* of code $\mathcal{C}$ can be represented as $|\mathcal{C}| = [q]^k = V$
- The *rate* of code $\mathcal{C}$ to be the normalized quantity $\frac{k}{n}$
- The *min-distance* between different codewords defined as $\min\limits_{c,c^* \in \mathcal{C}} \mathsf{dis}(c, c^*)$

It is convenient to view code $\mathcal{C}$ as a function $\mathcal{C} : [q]^k \to [q]^n$. Under this view, the elements of V can be considered as a message $v \in V$ and the process to generate its associated codeword $\mathcal{C}(v) = c$ is called *encoding*. Viewed this way, encoding a message v of size k, always adding redundancy to produce codeword $c \in [q]^n$ of longer size n. Nevertheless, for any codeword c with at most $t = \lfloor \frac{d-1}{2} \rfloor$ symbols are being modified to form c', it is possible to uniquely recover c from c' by using certain function f s.t. $\mathsf{f}(c') = c$. The procedure to find the unique $c \in \mathcal{C}$ that satisfied $\mathsf{dis}(c, c') \leq t$ by using f is called as *decoding*. A code $\mathcal{C}$ is said to be efficient if there exists a polynomial time algorithm for encoding and decoding.

Linear error correction code [19]: Linear error correction code is a linear subspace of $\mathbb{F}_q^n$. A q-ary linear code of blocklength n, dimension k and minimum distance d is represented as $[n,k,d]_q$ code $\mathcal{C}$. For a linear code, a string with all zeros 0^n is always a codeword. It can be specified into one of two equivalent ways with a generator matrix $G \in \mathbb{F}_q^{n \times k}$ or parity check matrix $H \in \mathbb{F}_q^{(n-k)\times n}$:

- a $[n,k,d]_q$ linear code $\mathcal{C}$ can be specified as the set $\{Gv : v \in \mathbb{F}_q^k\}$ for an $n \times k$ metric which known as the *generator matrix* of $\mathcal{C}$.
- a $[n,k,d]_q$ linear code $\mathcal{C}$ can also be specified as the subspace $\{x : x \in \mathbb{F}_q^n$ and $Hx = 0^{n-k}\}$ for an $(n-k) \times n$ metric which known as the *parity check matrix* of $\mathcal{C}$.

For any linear code, the linear combination of any codewords is also considered as a codeword over $\mathbb{F}_q^n$. Often, the encoding of any message $v \in \mathbb{F}_q^k$ can be done with $O(nk)$ operations (by multiplying it with the generator matrix, i.e., Gv. The distance between two linear codewords refers to the number of disagree elements between them, also known as the *hamming distance*.

Shannon Code [20] Let a binary code $\mathcal{C}$ over $\{0,1\}^n$. We call that $\mathcal{C}$ is an $[t, \varepsilon]$-Shannon code if there exits an encoding and decoding algorithm $\langle \mathsf{Encode}, \mathsf{decode} \rangle$ such that, Encode encode any k bits message to n bits codeword $c \in \mathcal{C}$, and a decode decode any codeword c' for all $t' \leq t$, and $c \in \mathcal{C}$, $\Pr[\mathsf{dis}(c, c') \leq t' \wedge \mathsf{decode}(c' \neq c)] \leq \varepsilon$.

Locality Sensitive Hashing (LSH) [21] Given that $P_2 > P_1$, while $w, w' \in \mathcal{M}$, and $\mathcal{H} = h_i : \mathcal{M} \rightarrow U$, where U refers to the output metric space (after hashing), which comes along with a similarity function S, where i is the number of hash functions h_i. A locality sensitive hashing can be viewed as a probability distribution over a family $\mathcal{H}$ of hash functions follows $P_{h \in \mathcal{H}}[h(w) = h(w')] = S(w, w')$. In particular, the similarity function S described the hashed collision probability in between w and w'.

$$P_{h \in \mathcal{H}}(h_i(w) = h_i(w')) \leq P_1, \quad \text{if } S(w, w') < R_1$$
$$P_{h \in \mathcal{H}}(h_i(w) = h_i(w')) \geq P_2, \quad \text{if } S(w, w') > R_2$$

LSH transforms input w and w' to its output metric space U with property that ensuring similarity inputs render higher probability of collision over U, and vice versal.

3 New Construction-LSH Secure Sketch

We hereby provide the detail of our based design of on a pair of sketching and recover algorithm, that incorporated with LSH, by hamming hash strategy.

3.1 LSH-Hamming hash

We first formulate the hamming-hash algorithm $\Omega^{\mathsf{ham-h}}$ which will be used in our LSH-sketching and recover algorithms description later. Generally, the hamming-hash algorithm $\Omega^{\mathsf{ham-h}} : \mathcal{M}_1 \times [k^*]^n \rightarrow \mathcal{M}_2$ is an iterative process through repeating the hamming hash strategy (steps 1 and 2) up to n times. It serves to sample the input binary string of size k^* into a longer binary string a.k.a resilient vector of size $n > k^*$.

Given input $w \in \{0,1\}^{k^*}$, and $N \in [k^*]^n$, the LSH-hamming hash algorithm describes as follow:

```
$\Omega^{\mathsf{ham-h}}(w, N)$
$\phi \leftarrow \emptyset$
for $i = 1, \ldots, n$ do
    parse $x = w(N(i))$// $x$ is the $N(i)$-th bits of $w$
    $\phi = \phi \| x$
endfor
return $\phi$
```

3.2 LSH-Hamming hash

We denote the LSH-sketching algorithm that employs the hamming-hash algorithm, Ω, an $[n^*, k^*, t^*]_2$ 'inner' code $\mathcal{C}^*$ and an $[n, k, t]_2$ 'outer' code $\mathcal{C}_\xi$ with generator matrix G^* and G respectively[3] as $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}$.

For sketching, one requires to generate a resilient vector ϕ by using the LSH hamming hash algorithm. The size of the resilient vector must same with the output codeword c. Then, the sketch ss can be constructed by simply perform an XOR operation, i.e., $ss = c \oplus \phi$. Besides, to add additional error to the input, we use a random error vector $e \in \mathsf{supp}(\chi)$ over some random distribution χ parametrized by $\epsilon > 0$. Specifically, we have $\|\chi\| = \lfloor k^* \epsilon \rfloor$ where the error vector e is of weight $\|e\| = \|\chi\| = \lfloor k^* \epsilon \rfloor$. The sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}$ with inputs w, N, G^*, G and ϵ is describe as follow:

```
$\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon)$
$\chi \leftarrow_{\$} \{0,1\}^{k^*}$// sample $\chi$ according to the error parameter $\epsilon$
$e \leftarrow_{\$} \mathsf{supp}(\chi)$// sample $e$ from $\chi$ uniformly at random, where $\|e\| = \|\chi\| = \lfloor k^* \epsilon \rfloor$
$c^* = G^* w$;// encode $w$
$v^* = c^* \oplus (0^{n^* - k^*} \| w)$;
$v^* = 0^{k - n^*} \| v^*$;// this padding step is only require if $n^* < k$
$c = G v^*$;// encode $v^*$
$w_e = w \oplus e$;
$\phi \leftarrow \Omega^{\mathsf{ham-h}}(w_e, N)$
$ss = c \oplus \phi$
return $(ss, N, G^*, G)$
```

All steps on $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon)$ can be done in $O(n^2)$. Notably, the size of ss is now depend upon the chosen 'outer' code $\mathcal{C}_\xi$.

[3] Remark that G^* or G can use to generate their corresponding parity check matrices H^* or H respectively

3.3 LSH-Recover

For recovery, suppose one wishes to recover w from another string $w' \in \{0,1\}^{k^*}$. He/she needs to provide another resilient vector ϕ'. This resilient vector can be generated by using the same hamming hash algorithm Ω with error added inputs $w'_e = w' \oplus e$. We want the recovery algorithm to output w for any error rate $\|w \oplus w'\| (k^*)^{-1} \leq \xi \pm \epsilon$ or $\|w \oplus w'\| \leq t_{(+)}$. Therefore, the error included by an error vector $e \in \mathsf{supp}(\chi')$ sampled from another error distribution χ', has shown doubled in amplitude i.e., $\|\chi'\| = \lfloor 2k^*\epsilon \rfloor$ compared to χ in the sketching phase.

We denote the LSH-recover algorithm that employs the hamming-hash algorithm, Ω, an $[n^*, k^*, t^*]_2$ 'inner' code $\mathcal{C}^*$ with generator matrix G^*, an $[n, k, t]_2$ 'outer' code $\mathcal{C}_\xi$ with generator matrix G, and a decoding algorithm f as $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$. The recover algorithm $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ with inputs ss, w', N, G^*, G and ϵ to recover w is describe as follow:

```
$\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, G^*, G, \epsilon)$

$\chi' \leftarrow_\$ \{0,1\}^{k^*}$ // sample $\chi'$ with error parameter $\epsilon$ i.e., $\|\chi'\| = \lfloor 2k^*\epsilon \rfloor$
for $i = 1, \ldots, |\mathsf{supp}(\chi')|$
  $e_i \leftarrow_\$ \mathsf{supp}(\chi')$// sample $e_i$ uniformly at random, where $\|e_i\| = \|\chi'\| = \lfloor 2k^*\epsilon \rfloor$
  $w'_{e_i} = w' \oplus e_i$
  $\phi'_i \leftarrow \Omega^{\mathsf{ham-h}}(w'_{e_i}, N)$
  $c'_i = ss \oplus \phi'_i$// also $ss \oplus \phi'_i = c \oplus (\phi \oplus \phi'_i)$
// try to decode the codeword:
  $c \leftarrow \mathsf{f}(c'_i)$// first decoding
  if first decoding is succeeded
  $v^* \leftarrow G^{-1}c$
  $c'^* = v^* \oplus (0^{n^*-k^*} \| w')$
  $c^* \leftarrow \mathsf{f}(c'^*)$// second decoding
  if second decoding is succeeded
    $w \leftarrow G^{*-1}c^*$
    return $w$
  endif
  endif
endfor
```

A brief description of the recovery mechanism is given as follow. Suppose Bob has intercepted with a sketch $ss = c \oplus \phi$. Firstly, he has to double the error parameter from ϵ to 2ϵ and generate a resilient vector $\phi'_i \leftarrow \Omega^{\mathsf{ham-h}}(w'_{e_i}, N)$. The hamming weight of the offset can be conveniently represented as $\|\delta_i\| = \|\phi \oplus \phi'_i\|$. By means of the similarity preservation property of LSH, the offset, δ_i is expected to be low as well if w and w' are close to each other. Bob then performs $ss \oplus \phi'_i$ to output the nearest codeword c'_i. The errors over c'^* and c'_i can be tolerated by means of error correction with codes $\mathcal{C}^*$, $\mathcal{C}_\xi$ respectively and a decoding function f. Such decoding process is repeat for $i = 1, \ldots, |\mathsf{supp}(\chi')|$ iterations to try with all possible input error patterns over $|\mathsf{supp}(\chi')|$.

4 Resilience

We now consider the resilience of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$. Generally, the resilience measures on how probable the offset $\|\delta\|$ can be tolerated in facilitating the recovery of v^* and so w from the sketch. High resilience implies high probability to tolerate the offset or correcting the errors.

Clearly, the resilience of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is bounded by the resilience of the selected codes pair $\mathcal{C}^*$ and $\mathcal{C}_\xi$. This also mean that the value of ϵ should be bounded by the maximum achievable error tolerence rate, determined among the chosen codes pair $\mathcal{C}^*$ and $\mathcal{C}_\xi$, i.e, $0 < \epsilon \leq \max\{t^*/k^*, \xi\}$. Since $t^* < t$, certainly $0 < \epsilon \leq \xi$. Choosing a 'good' code with a high value of ξ is non-trivial, this is because different code is subjected to different set of parameters (n, k, t) and there is no straightforward way to determine which the most efficient one is. The design of such code under different set of parameters (n, k, t) is another broad research topic. We direct the interested user refer to the works of Macwilliams, (1977) [22], and Peterson and Weldo, (1972) [23]. Nevertheless, by padding zeros on the input v^*, the selection of different codes pair $\mathcal{C}^*$ and $\mathcal{C}_\xi$ are highly relieved, since once can easily find a pair of code $\mathcal{C}^*$ and $\mathcal{C}_\xi$ with $k^* < n^* \leq k < n$.

In this section, we are more interested in the probability to recover the original input w. We will leave the discussion of the topic regarding resilience bound to the following Section 4.1.

Further simplification is done by describing the term *overwhelming* if the value of $1-\beta$ is close to one (e.g., negligible β). As we shall see, negligible β means substantial average fuzzy min-entropy.

We hereby formalize the *completeness* of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ for our resilience claim. It captures the scenario when the players are honest, which is define under the following definition.

Definition 1. *For some random variables $W, W' \in \mathcal{M}_1$, and $SS \in \mathcal{M}_2$, where $\mathcal{M}_1 = \{0,1\}^{k^*}$ and $\mathcal{M}_2 = \{0,1\}^n$. Given $N \in [k^*]^n$, an $[n^*, k^*, t^*]_2$ linear code $\mathcal{C}^*$ and an $[n, k, t]_2$ linear code $\mathcal{C}_\xi$ with generator matrix $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$ respectively, where $t^* < t$ and $k^* < n^* \leq k < n$. For $\epsilon \in (0, \xi]$, let $t_{(-)} = \lceil (\xi - \epsilon)k^* \rceil$, and $\xi = t/n$. For a sketch ss generated through $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon) = ss$, where $ss \in SS$. For all $w \in W$, $w' \in W'$, then one has $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, G^*, G, \epsilon) = w$ can be achieved with overwhelming probability at least $1-\beta$ if $\|w \oplus w'\| \leq t_{(-)}$. We said $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is complete in (β, ϵ)-fuzziness if above statement holds.*

We hereby provide a proposition with proof for the completeness claim on $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$.

Proposition 1. *For sufficiently large n, $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is complete in (β,ϵ)-fuzziness if $\|w \oplus w'\| \leq t_{(-)} \leq t^*$, where $\beta = \exp(-2n\epsilon^2)$.*

Proof. Clearly, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, G^*, G, \epsilon) = w$ can only be achieved if both decoding processes $\mathsf{f}(c')$[4] and $\mathsf{f}(c'^*)$ can be done successfully. To be specific, for the first decoding stage, it follows that $\mathsf{f}(c') = \mathsf{f}(c \oplus \delta) = \mathsf{f}((c \oplus \phi) \oplus \phi') = \mathsf{f}(c \oplus (\phi \oplus \phi'))$. If $\|\phi \oplus \phi'\| = \|\delta\| \leq t$, this decoding will success and return c and so v^*. For the second decoding stage. It follows $\mathsf{f}(c'^*) = \mathsf{f}(v^* \oplus (0^{k^*} \| w')) = \mathsf{f}(c'^* \oplus (w \oplus w'))$. If $\|w \oplus w'\| \leq t_{(-)} \leq t^*$, the second decoding stage will success as well by returning c^* and so w. Therefore, both scenarios $\|\delta\| \leq t$ and $\|w \oplus w'\| \leq t^*$ must hold for successful recovery of w. For any value of $t_{(-)} = \lceil (\xi - \epsilon) k^* \rceil \geq 0$, the second decoding stage will always success with probability one if $\|w \oplus w'\| \leq t_{(-)} \leq t^*$. Therefore, we only need to consider the probability for the scenario $\|\delta\| \leq t$ to hold (for first decoding stage to success). Follow Corollary 1. When $\|w \oplus w'\| \leq t_{(-)}$, this probability is therefore expresses as:

$$\min_{t = t_{\min}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(-)} \right] = 1 - \beta$$

It follows $1 - \beta$ is overwhelming with negligible quantity $\beta = \exp\left(-2n\epsilon^2\right)$ when n is sufficiently large. Hence, the proposition is prove.

Proposition 1 concluded that given two linear error correction codes $\mathcal{C}^*$ and $\mathcal{C}_\xi$, for $\epsilon \in (0, \xi]$, when $\|w \oplus w'\| \leq t_{(-)}$, then $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, G^*, G, \epsilon) = w$ can be achieved with overwhelming probability at least $1 - \beta$ with decoding function f if one has the value of n is sufficiently large and $t^* \geq t_{(-)}$.

4.1 Correcting More Errors in Polynomial Time

This section provided the details explanation for error correction of error rate up to $\xi + \epsilon$ or $t_{(+)}$.

Focusing on the second decoding stage, it only success when $\|w \oplus w'\| \leq t_{(-)} \leq t^*$. This result demonstrating a limited amount of the error rate $\|w \oplus w'\| \leq \xi - \epsilon$ can be corrected by the 'inner' code $\mathcal{C}^*$ with probability one. On the other hand, one can actually correct more errors with additional 'outer' code $\mathcal{C}_\xi$ imposes on top of $\mathcal{C}^*$. In particular, when $\|w \oplus w'\|(k^*)^{-1} \leq \xi + \epsilon$ or $\|w \oplus w'\| \leq t_{(+)}$. By introducing additional random error of rate -2ϵ during recovery phase, the error rate is possible to be reduced down to $\|w \oplus w'\|(k^*)^{-1} \leq \xi - \epsilon$, hence the remaining errors can be corrected with probability one when $t_{(-)} \leq t^*$.

Based on the above argument, we hereby provide the *soundness* of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ to characterize the resilience of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ for higher error rate, e.g., at most $\|w \oplus w'\|(k^*)^{-1} \leq \xi + \epsilon$. Formally, this soundness of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ covers the scenario where any adversary is capable of sampling any w' s.t. $\|w \oplus w'\| \leq t_{(+)}$ holds.

[4] we here omitted the i-th notation for neater presentation

Often, for efficient recovery, we always hope that $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ can be done successfully in a few iterations, polynomial in the sketch size (n). The definition below defines the soundness of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ in correcting the errors probability at least $1-\beta$ and efficiently under the event when $\|w \oplus w'\| \leq t_{(+)}$.

Definition 2. *For some random variables $W, W' \in \mathcal{M}_1$, and $SS \in \mathcal{M}_2$, where $\mathcal{M}_1 = \{0,1\}^{k^*}$ and $\mathcal{M}_2 = \{0,1\}^n$. Given $N \in [k^*]^n$, an $[n^*, k^*, t^*]_2$ linear code $\mathcal{C}^*$ and an $[n,k,t]_2$ linear code $\mathcal{C}_\xi$ with generator matrix $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$ respectively, where $k^* < n^* \leq k < n$. For $\epsilon \in (0,\xi]$, let $t_{(+)} = \lfloor(\xi+\epsilon)k^*\rfloor$, $t_{(-)} = \lceil(\xi-\epsilon)k^*\rceil$, and $\xi = t/n$. For all $w \in W, w' \in W'$, given a sketch ss generated by $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w,N,G^*,G,\epsilon) = ss \in SS$. We said $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ is efficient if $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}(ss,w',N,G^*,G,\epsilon) = w$ can be done in time $\mathsf{poly}(n)$*

For the 'inner' code $\mathcal{C}^*$, we propose to use a type of efficient code named BCH code [23] with efficient decoding algorithm f via algebric method, i.e., syndrome decoding [24]. To claim the soundness of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$, we have to show the existence of such 'outer' code $\mathcal{C}_\xi$ for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ with efficient decoding algorithm f as well.

Let denote $h_2(\epsilon) = -(\epsilon)\log(\epsilon) - (1-\epsilon)\log(1-\epsilon)$ be the binary entropy function of error rate ϵ, we provide a theorem with proof of the existence of the 'outer' code $\mathcal{C}_\xi$, and the efficiency of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ itself.

Theorem 3. *Suppose code $\mathcal{C}^*$ is a $[n^*, k^*, t^*]$ BCH code. For $\epsilon \in [(k^*)^{-1}, \xi]$, and $\exp(-2n\epsilon) < 0.125$, there exists another $[n,k,t]$ BCH code $\mathcal{C}_\xi$ with syndrome decoding algorithm f for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ to correct the error rate of $\xi+\epsilon$ with probability at least $1-\exp(-2n\epsilon)$ in time $\mathsf{poly}(n)$ when $\lceil kh_2(\epsilon)\rceil \leq \log(n+1)$, $\lfloor 2k^*\epsilon\rfloor \leq \|w \oplus w'\| \leq t_{(+)}$, and $t^* \geq t_{(-)}$.*

Proof. We first provide the existance proof for code $\mathcal{C}_\xi$. We knew that, for some positive integer $m' > 3$ and $t < 2^{m'-1}$, there exits an BCH code (computation in Galois field $GF(2^{m'})$) with parameters $n = 2^{m'} - 1$, $n-k \leq m't$ and minimum distance $d \geq 2t-1$. The total number of codewords in such BCH code must be bounded by $2^{m'} = n+1$. For $\exp(-2n\epsilon^2) < 0.125$, there are at least $1/\exp(-2n\epsilon^2) = 2^{\log(1/\exp(-2n\epsilon^2))} > 2^3$ number of codewords which are consider as 'similar' codewords (e.g., $c'_i \in \mathcal{C}_\xi \wedge \|\delta\| \leq t$) for the first decoding stage to success with probability at least $1-\exp(-2n\epsilon^2)$ (see Proposition 1). It follows that $2^{\lceil\log(1/\exp(-2n\epsilon^2))\rceil} \geq 2^{\log(1/\exp(-2n\epsilon^2))}$, and we could have $\lceil\log(1/\exp(-2n\epsilon^2))\rceil = m' = \log(n+1) > 3$ to claim the existence of such 'outer' BCH code $\mathcal{C}_\xi$ with efficient syndrome decoding algorithm f.

We now revert to our main goal of correcting the error rate $\xi+\epsilon$ or $t_{(+)}$. Noting that doubling the error during decoding would cause the input error rate to increase (i.e., $\|w \oplus w'\|(k^*)^{-1} \leq \xi+\epsilon+2\epsilon$) or decrease (i.e., $\|w \oplus w'\|(k^*)^{-1} \leq \xi+\epsilon-2\epsilon$). Clearly, when the error rate decreases, it means the errors $\|w \oplus w'\| \leq t_{(-)}$, then, the second decoding stage will success with probability one if $t^* \geq t_{(-)}$ by the chosen 'inner' code $\mathcal{C}^*$. We have to show the probability of the errors to decrease from $\|w \oplus w'\| \leq t_{(+)}$ to $\|w \oplus w'\| \leq t_{(-)}$. Given error vector

$\|e_i\| = \lfloor 2k^*\epsilon \rfloor$, we have $|\mathsf{supp}(\chi')| = \binom{k^*}{\lfloor 2k^*\epsilon \rfloor}$ possible ways to describe all different combinations of the error vector $e_i \in \mathsf{supp}(\chi')$. It follows $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ maximally run in $\binom{k^*}{\lfloor 2k^*\epsilon \rfloor}$ iterations. This probability can be expressed as:

$$\Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(+)}\right] = \frac{\binom{\|w \oplus w'\|}{\lfloor 2k^*\epsilon \rfloor}}{\binom{k^*}{\lfloor 2k^*\epsilon \rfloor}}$$

For efficiency claim, to keep $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ run in low number of iterations, we need $\binom{k^*}{\lfloor 2k^*\epsilon \rfloor}/\binom{\|w \oplus w'\|}{\lfloor 2k^*\epsilon \rfloor} \leq 2^{m'}$ to ensure $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ maximally run in $2^{m'} = n+1$ iterations. With $\lfloor 2k^*\epsilon \rfloor \leq \|w \oplus w'\| \leq t_{(+)}$, by *Stirling's approximation*, $\binom{k^*}{\lfloor 2k^*\epsilon \rfloor} \leq \binom{k^*}{2k^*\epsilon} \leq 2^{k^* h_2(\epsilon)} < 2^{n^* h_2(\epsilon)} \leq 2^{kh_2(\epsilon)}$ holds when $\epsilon \in [(k^*)^{-1}, 1/4]$, hence $\frac{\binom{\|w \oplus w'\|}{\lfloor 2k^*\epsilon \rfloor}}{\binom{k^*}{\lfloor 2k^*\epsilon \rfloor}} \geq 2^{-kh_2(\epsilon)}$. Recall $\epsilon \in (0, \xi]$ must hold (see Proposition 1), therefore $\epsilon \in [(k^*)^{-1}, \xi]$. For $\lceil kh_2(\epsilon) \rceil \leq \log(n+1) = \lceil \log(1/\exp(-2n\epsilon^2)) \rceil$, the solution is then follow:

$$\begin{aligned}
&\Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(+)}\right] \geq 2^{-kh_2(\epsilon)} \\
&\geq 2^{-\lceil kh_2(\epsilon) \rceil} \geq 2^{-\lceil \log(1/\exp(-2n\epsilon^2)) \rceil} = 2^{-m'} = \frac{1}{2^{O(\log(n+1))}} = \frac{1}{\mathsf{poly}(n)} \qquad (4)
\end{aligned}$$

After $\mathsf{poly}(n)$ iterations, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ would success and output w with probability at least $1 - \exp(-2n\epsilon^2) > 0.875$ and complete the prove.

In summary, for $\epsilon \in [(k^*)^{-1}, \xi]$, given $\exp(-2n\epsilon^2) < 0.125$, $\lfloor 2k^*\epsilon \rfloor \leq \|w \oplus w'\| \leq t_{(+)}$ and $t^* \geq t_{(-)}$, one needs $\lceil kh_2(\epsilon) \rceil \leq \log(n+1)$ to ensure efficient decoding with f in $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ to correct the error rate of $\xi + \epsilon$.

We would provide more details discussion regarding the error correction bound in the next subsection.

4.2 Error Correction up to Shannon Bound

In the previous section, we have demonstrated the resilience of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$, in term of the probability in correcting the errors at most $t_{(+)}$. Although, high probability in correcting the errors does not always mean high number of errors can be corrected. Therefore, this section will provide the discussion on how much errors can be corrected by using $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$. Formally, we call this as the resilience bound of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$.

Generally, to study the resilience bound, the error model of the system must be conceived. It is mean to say that, without any knowledge on the error process of the input, it is difficult to precisely model and determine the resilience bound of a given error correcting construction. It is also heedless for one to believe that people have a complete understanding of the complex error pattern, or the distribution that is overtaking by the noisy non-uniform sources, i.e., biometric.

To study the resilience bound without the knowledge of the input error process, one can always use the *perfect correctness* model. Recall that, high resilience means the errors can be corrected with overwhelming probability $1-\beta$. Ideally, it is natural to let $\beta = 0$, which will easily lead to the perfect correctness model, so, the errors can be corrected with probability one. This means there will be only one unique solution for every w' within distance t. Hence, the decoding process always returns the original value w precisely (e.g., unique decoding). In this model, the fuzzy min-entropy notion may not necessary, since one can easily show infinite fuzzy min-entropy without any dissension for security. Therefore, this model is useful and suitable for who try to avoid certain assumption about the exact properties of the stochastic error process, or the computational power of an adversary to carry out decoding successfully.

However, inevitably, under the perfect correctness model, one always tied to a very strong bound in term of the resilience. Typically, one can only uniquely decode the codeword by using an error correction code with min-distance $d \geq 2t+1$. Saying so, the Plotkin bound (see [25]) has revealed the limited maximum number of codewords in a code of blocklength n and minimum distance d. More formally, there can be only at most $2n$ codewords with $d > n/2$, which means given the residual entropy larger or equal to $\log(n)$, there has no error correction code can correct $n/4$ errors with probability one and so for a secure sketch.

Despite of this, for sufficiently large n, the code $\mathcal{C}_\xi$ would contain large distance in between the codewords itself (i.e., $d \geq 2t+1$) with overwhelming probability ([26], Theorem 8). In such an event, one has a slightly relaxed notion of correctness called *probabilistic correctness model*. Notably, our construction naturally categorized under this relaxed model, where the first decoding stage $\mathsf{f}(c'_i)$ will not succeed with probability one, rather $1-\beta$, with some probability to fail. The failure in decoding is subjected to the condition of either $W \in B_{t_{(+)}}(w')$ or $W \notin B_{t_{(+)}}(w')$ for a given sketch ss. Therefore, a higher distance between the codewords implicitly reduces the failure in decoding. This relaxed notion of correctness is essential for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ to free from the Plotkin bound and allow it to correct more errors by increment of the error parameter ϵ. In contrary, we have our perfect correctness guarantee only when $W \in B_{t_{(-)}}(w')$ via second decoding $\mathsf{f}(c'^*)$, with number of correctable errors up to $t_{(-)} \leq t^* < t$.

We now show that the probabilistic correctness model has allowed us to correct more errors, arbitrarily close to $n/2$. Credited by the LSH-hamming hash, the errors in a pair of resilient vectors can be described by using the Bernoulli process. More formally, our works following the random error model which was famously considered by Shannon [20]. Shannon provided the noisy channel coding theorem saying that, for any discrete memoryless channel, the error tolerance rate is characterized by the maximum mutual information between the input and outputs. Precisely, in a binary symmetric channel, like our case, there exists a code encoding k bits into n bits which able to tolerate the error of probability p for every single bit, if and only if:

$$k < \lfloor (1-h_2(p))n \rfloor$$

Since $h_2(p)$ is maximally one when $p = 1/2$, conversely, this theorem indicates the existence of a secure sketch even for high error rate as long as p is smaller than $1/2$.

Our efficiency claim for maximum error correction using $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ is as follow:

Proposition 2. *If both $\mathcal{C}^*$ and $\mathcal{C}_\xi$ are BCH codes with syndrome decoding algorithm f, where $\xi = t/n \in (0, 1/4)$. For $\epsilon \in [(k^*)^{-1}, \xi]$ and $\exp(-2n\epsilon^2) < 0.125$, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ is efficient when $\lceil kh_2(\epsilon)\rceil \leq \log(n+1)$, $\lfloor 2k^*\epsilon\rfloor \leq \|w \oplus w'\| \leq t_{(+)}$, and $t^* \geq t_{(-)}$ where the maximum correctable error rate is $2\epsilon \leq 2\xi < 1/2$.*

Proof. By Theorem 3, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ can correct the error rate of $\|w \oplus w'\|\,(k^*)^{-1} \leq \xi + \epsilon$ efficiently when $\lceil kh_2(\epsilon)\rceil \leq \log(n+1)$, $\lfloor 2k^*\epsilon\rfloor \leq \|w \oplus w'\| \leq t_{(+)}$, and $t^* \geq t_{(-)}$. clearly, with maximum $\epsilon = \xi$, the maximum correctable error rate is $\|w \oplus w'\|\,(k^*)^{-1} \leq \xi + \epsilon \leq 2\epsilon \leq 2\xi < 1/2$

Finally, we give a corollary whom proof is instantiated by the proof of Theorem 3 and Proposition 2, to formalize $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ as an $[2t, \beta]$ Shannon code.

Corollary 3. *For any pair of BCH codes $\mathcal{C}^*$ and $\mathcal{C}_\xi$ with efficient decoding algorithm f used in $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$, $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is an efficient $[2t, \beta]$-Shannon code with $\beta = \exp(-2n\epsilon^2)$*

It is useful to have an example to show how $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ works according to our claim follows Corollary 3 with two BCH codes.

Example 1. Suppose one wishes to correct some errors over an input $w \in \{0,1\}^{k^*}$ of length $k^* = 11$. Given an $[1023, 56, 191]_2$ BCH code is chosen for 'outer' code $\mathcal{C}_\xi$. By efficiency argument, he/she needs $\lceil kh_2(\epsilon)\rceil \leq \log(n+1)$, thus, $h_2(\epsilon) \leq 0.1786$, so $\epsilon \leq 0.0269$. In such a case, one has the correctable error rate is bounded at most $\xi + \epsilon = 0.2136$ (i.e., $t_{(+)} = \lfloor(0.2136)k^*\rfloor$, and he/she needs an 'inner' code $\mathcal{C}^*$ that could correct at least $t_{(-)} = \lceil(0.01599)k^*\rceil = 1$ error, i.e., $[n^*, k^*, t^*]$ code $\mathcal{C}^*$ with $n^* = 31, k^* = 11, t^* = 5$. On the other hand, let say the input is of length $k^* = 4$. For an $[1023, 11, 255]_2$ 'outer' code $\mathcal{C}_\xi$ where $\xi = 0.2493$. One can easily compute $h_2(\epsilon) \leq 0.9091$. Then, he/she is capable of choosing maximum $\epsilon = \xi = 0.2493$ for maximum error correction capacity, which is at most of rate $\xi + \epsilon = 0.4986$, which is equivalent to $2t = 510$ over $\{0,1\}^n$. The errors can be corrected with overwhelming probability at least $1 - 5.95 \times 10^{-56}$. Meanwhile, he/she would need to choose an 'inner code' $\mathcal{C}^*$ with $n^* = 7, k^* = 4$ and $t^* = 1$. since $t_{(-)} = 1 \leq t^*$, the second decoding stage would success with probability one by syndrome decoding algorithm f.

Apart from this, computationally efficient code achieve Shannon bound is also found by Forney in 1965, named as *concatenated code* [27]. Our construction relies on similar approach as the concatenated code for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ with an 'outer' code $\mathcal{C}_\xi$ build on top of an 'inner' code $\mathcal{C}^*$, where both are linear codes as well. However, obviously, our construction operating differently by taking into consideration over the input error's distributions parametrized by ϵ, and the resilient vector.

5 Security

Recall that for all $w \in W$ in some random distribution W over $\mathcal{M}_1$, the soundness of $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}$ captured the scenario where any adversary $\mathcal{A}$ is capable of sampling any $w' \in W'$ satisfy $\|w \oplus w'\| \, (k^*)^{-1} \leq \xi + \epsilon$ or $\|w \oplus w'\| \leq t_{(+)}$. The security of our proposal depends upon the hardness in searching a variable W satisfy $W \in B_{t_{(+)}}(w')$. Since error correction implies entropy loss, this loss must be indicated by the number of codewords in $\mathcal{C}_\xi$ which is consider as 'similar' (e.g., $c'_i \in \mathcal{C}_\xi \wedge \|\delta\| \leq t$) for the first decoding stage to success. Therefore, any 'similar' codewords would contribute to additional information for an attacker to differentiate whether $W \in B_{t_{(+)}}(w')$. In the absence of these additional information, W is hidden in information theoretically fashion with security proportional to the entropy contained in the distribution W over $\mathcal{M}_1$.

We now formalize the security of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$. We assume an original input w is randomly sampled from some random distribution W (not mandatory uniform) over a metric space $\mathcal{M}_1 = \{0,1\}^{k^*}$. Besides, we restrain another sample $w' \in W'$ that show at least error rate of $\|w \oplus w'\| \, (k^*)^{-1} \geq \xi + \epsilon$ (holds for $\|w \oplus w'\| \geq t_{(+)}$) with the original sample w. Recall the number of correctable error is bounded proportional to the introduced error of parameter ϵ (i.e., 2ϵ), we hereby giving $\mathcal{A}$ full power in choosing any other error parameter $\epsilon' \leq \epsilon$ in recovering w for maximum error tolerance.

We sake to characterize the security of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ by using an adversary $\mathcal{A}$ comes with unlimited computation power and probable of sampling any w' satisfies $\|w \oplus w'\| \geq t_{(+)}$. The security is formalize by using an attack running together with $\mathcal{A}$. Formally, $\mathcal{A} : \mathcal{M}_1^2 \times \mathcal{M}_2 \times \mathbb{F}_2^{n^* \times k^*} \times \mathbb{F}_2^{n \times k} \times [k^*]^n \rightarrow \mathcal{M}_1$ is just an algorithm that is computationally unbounded, aim to recover w from a sketch $ss \in \mathcal{M}_2$, with the generator matrices $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$, an integer string $N \in [k^*]^n$ and $w' \in \mathcal{M}_1$ and error parameter $\epsilon' \leq \epsilon$. The attack is denote as $\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A})$ with LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}$, and inputs N, G^*, G, ϵ, and $\mathcal{A}$ as follow:

```
Attack($\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}$)
1 : $w \leftarrow W$// sample according to some distribution $W \in \mathcal{M}_1$
2 : $w' \leftarrow W'$
3 : if $\|w \oplus w'\| \leq t_{(+)}$, repeat step 2 until $\|w \oplus w'\| \geq t_{(+)}$
4 :     if $\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon), w', N, G^*, G, \epsilon') = w$
5 :         Output true
6 :     else
7 :         Output false
8 :     endif
9 : endif
```

We then have the following definition for our security.

Definition 3. *Let $\mu > 0$ and $\epsilon \in [(k^*)^{-1}, \xi]$. Given some random variables $W, W' \in \mathcal{M}_1$, and $SS \in \mathcal{M}_2$, where $\mathcal{M}_1 = \{0,1\}^{k^*}$ and $\mathcal{M}_2 = \{0,1\}^n$. Given $N \in [k^*]^n$, an $[n^*, k^*, t^*]_2$ linear code $\mathcal{C}^*$ and an $[n, k, t]_2$ linear code $\mathcal{C}_\xi$ with generator matrices $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$ respectively, where $\xi = t/n$ and $k^* < n^* \leq k < n$. For all $w \in W$, $w' \in W'$, and $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon) = ss \in SS$, $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is $(\mu, \xi + \epsilon)$-information theoretically secure if $\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}) = \mathbf{true}\right] \leq \mu$ for any computationally unbounded adversary $\mathcal{A}$.*

According to Definition 3, we give a theorem with proof to generally characterize the information theoretical security of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$.

Theorem 4. *Let a positive integer $m \geq 1$, if both $\mathcal{C}^*$ and $\mathcal{C}_\xi$ are BCH codes with syndrome decoding algorithm f, then, the algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is $(\mu, \xi+\epsilon)$-information theoretically secure with $\mu = \left(\frac{2^{-m-1}}{1-2^{-m}}\right)$, when $\lceil k h_2(\epsilon) \rceil > \lceil \log(1/\exp(-2n\epsilon^2)) \rceil$, and $\|w \oplus w'\| \geq t_{(+)}$.*

Proof. : We first analyse the amount of effort for any adversary $(\mathcal{A})$ in recovering w from a sketch when $\|w \oplus w'\| \geq t_{(+)}$. We claim that this amount of effort is large and bounded by the number of codewords which is consider as 'similar' for the first decoding stage to success.

Specifically, given $\|e'\| = \lfloor k^* \epsilon' \rfloor \leq \lfloor k^* \epsilon \rfloor$, one could easily note that the results of Corollary 1 and 2 are indeed offer the tighter bound for the probability of getting a similar pair of resilient vectors with offset $\|\delta\| \leq t$. In particular, let $t'_{(+)} = \lfloor (\xi + \epsilon') k^* \rfloor$, and $t'_{(-)} = \lceil (\xi - \epsilon') k^* \rceil$, where $\beta = \exp(-2n(\epsilon)^2)$:

$$\min_{t=t_{\min}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t'_{(-)}\right] = 1 - \exp(-2n(\epsilon')^2) \leq 1 - \beta$$

$$\max_{t=t_{\max}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \geq t'_{(+)}\right] = \exp(-2n(\epsilon')^2) \geq \beta \tag{5}$$

Therefore, the number of codewords which are consider as 'similar' is at most $1/\exp(-2n(\epsilon')^2) \leq 1/\beta$ by Eq. 5. It follows this number can be further bounded as $\log(1/\beta) = 2^{\log(1/\beta)} \leq 2^{\lceil \log(1/\beta) \rceil}$, hence:

$$\Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w, N, G^*, G, \epsilon), w', N, G^*, G, \epsilon') = w\right] \leq 2^{-\lceil \log(1/\beta) \rceil} \tag{6}$$

We now revert to our main problem where w is not sampled randomly instead of selected by the algorithm itself according to some distribution W (could be non-uniformly) over $\mathcal{M}_1$. We will show that in this case, the probability to sample any $W \in B_{t_{(+)}}(w')$ is at most $\left(\frac{2^{-m-1}}{1-2^{-m}}\right)$

It should be described as follow:

$$\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}) = \textbf{true}\right] = \Pr\left[\|w \oplus w'\| \leq t_{(+)} \;\middle|\; \|\delta\| \leq t\right]$$

To continue, we denote two events $\{\mathsf{Event}_a, \mathsf{Event}_b\}$ where $a, b \in \{0,1\}$ as follow:

$$\mathsf{Event}_a = \begin{cases} \|\delta\| \leq t, & a = 0 \\ \|\delta\| > t, & a = 1 \end{cases}$$

$$\mathsf{Event}_b = \begin{cases} \|w \oplus w'\| \leq t_{(+)}, & b = 0 \\ \|w \oplus w'\| \geq t_{(+)}, & b = 1 \end{cases}$$

By using *Bayes' law*:

$$\begin{aligned}
&\Pr\left[\|w \oplus w'\| \leq t_{(+)} \;\middle|\; \|\delta\| \leq t\right] = \frac{\Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(+)}\right] \Pr\left[\|w \oplus w'\| \leq t_{(+)}\right]}{\Pr\left[\|\delta\| \leq t\right]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0}\right]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right] + \Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=1}\right] \Pr\left[\mathsf{Event}_{b=1}\right]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right] + \Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=1}\right] (1 - \Pr\left[\mathsf{Event}_{b=0}\right])} \\
&= \frac{1}{1 + \frac{\Pr[\mathsf{Event}_{a=0} \,|\, \mathsf{Event}_{b=1}](1 - \Pr[\mathsf{Event}_{b=0}])}{\Pr[\mathsf{Event}_{a=0} \,|\, \mathsf{Event}_{b=0}]\Pr[\mathsf{Event}_{b=0}]}} \\
&< \frac{1}{\frac{\Pr[\mathsf{Event}_{a=0} \,|\, \mathsf{Event}_{b=1}](1 - \Pr[\mathsf{Event}_{b=0}])}{\Pr[\mathsf{Event}_{a=0} \,|\, \mathsf{Event}_{b=0}]\Pr[\mathsf{Event}_{b=0}]}} = \frac{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=1}\right] (1 - \Pr\left[\mathsf{Event}_{b=0}\right])} \\
&= \left(\frac{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=1}\right]}\right) \left(\frac{\Pr\left[\mathsf{Event}_{b=0}\right]}{1 - \Pr\left[\mathsf{Event}_{b=0}\right]}\right) \qquad (7)
\end{aligned}$$

By Eq. 5 and 6, we have $\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=1}\right] \leq 2^{-\lceil \log(1/\beta) \rceil}$.

By the result from Theorem 3 (Eq. 4), we have (for $\epsilon \in [(k^*)^{-1}, \xi]$, $\lfloor 2k^*\epsilon \rfloor \leq \|w \oplus w'\| \leq t_{(+)}$):

$$\Pr\left[\mathsf{Event}_{a=0} \;\middle|\; \mathsf{Event}_{b=0}\right] = \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(+)}\right] \geq 2^{-kh_2(\epsilon)} \geq 2^{-\lceil kh_2(\epsilon) \rceil}$$

To look for the solution of $\Pr\left[\mathsf{Event}_{b=0}\right]$, we need to use the fuzzy min-entropy notion, s.t. $\max_{w'} \Pr\left[W \in B_{t_{(+)}}(w')\right] = 2^{-\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t_{(+)},\infty}(W)}$. Therefore:

$$\begin{aligned}
&\Pr\left[\mathsf{Event}_{b=0}\right] = \Pr\left[\|w \oplus w'\| \leq t_{(+)}\right] \\
&\leq \max_{w'} \Pr\left[W \in B_{t(+)}(w')\right] = 2^{-\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t_{(+)},\infty}(W)}
\end{aligned}$$

Taking into consideration over all possible distributions for W of fuzzy min-entropy at least one bit, $\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t_{(+)},\infty}(W) \geq 1$, it follows $\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t_{(+)},\infty}(W) \geq \mathrm{H}_\infty(W) = m \geq 1$. Doing so implies the worst case distribution W comes with min-entropy $\mathrm{H}_\infty(W) = m$ at least one bit. Eq. (7) can be interpreted as:

$$\left(\frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\right]}\right)\left(\frac{\Pr\left[\mathsf{Event}_{b=0}\right]}{1-\Pr\left[\mathsf{Event}_{b=0}\right]}\right) = \left(\frac{2^{-\lceil kh_2(\epsilon)\rceil}}{2^{-\lceil \log(1/\beta)\rceil}}\right)\left(\frac{2^{-m}}{1-2^{-m}}\right)$$

For $\lceil kh_2(\epsilon)\rceil > \lceil \log(1/\exp(-2n\epsilon^2))\rceil$, $\lceil kh_2(\epsilon)\rceil - \lceil \log(1/\exp(-2n\epsilon^2))\rceil \geq 1$ must hold by minimum. The maximum achievable probability is thus:

$$\begin{aligned}&\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}) = \mathbf{true}\right] < \left(\frac{2^{-\lceil kh_2(\epsilon)\rceil}}{2^{-\lceil \log(1/\beta)\rceil}}\right)\left(\frac{2^{-m}}{1-2^{-m}}\right)\\ &\leq \left(\frac{2^{-m-1}}{1-2^{-m}}\right) \qquad (8)\end{aligned}$$

hence complete the prove.

Eventually, we give the below proposition to formalize $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ as an information theoretically secure sketch.

Proposition 3. *$\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is an efficient $(\mathcal{M}_2, m, m, t_{(+)})$-secure sketch*

Proof. We start from the proof of **correctness**. Obviously, the correctness simply follows the proof of Theorem 3, where we claimed when $\lfloor 2k^*\epsilon \rfloor \leq \|w \oplus w'\| \leq t_{(+)}$ and $\lceil kh_2(\epsilon)\rceil \leq \lceil \log(1/\exp(-2n\epsilon^2))\rceil = \log(n+1)$, the errors can be corrected with probability at least $1-\exp(-2n\epsilon^2) > 0.875$ efficiently.

For **security** proof, by Theorem 4 (Eq. 8). For $\epsilon \in [(k^*)^{-1}, \xi]$ and $\|w \oplus w'\| \geq t_{(+)}$, the residual entropy required for a computationally unbounded attacker to differentiate whether $W \in B_{t_{(+)}}(w')$ from a sketch $ss \in SS$ under some distribution SS over $\mathcal{M}_2$, two generator matrices $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$, an integer string $N \in [k^*]^n$ and $w' \in W'$ over $\mathcal{M}_1$ can be expressed as:

$$\begin{aligned}&\tilde{\mathrm{H}}_\infty(W|SS, G^*, G, N, W', \epsilon) = -\log\left(\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}) = \mathbf{true}\right]\right)\\ &\geq m + 1 + \log(1-2^{-m}) \qquad (9)\end{aligned}$$

For $m \geq 1$, the term $0 \geq \log(1-2^{-m}) \geq -1$ maximally contribute to entropy loss of one bit, hence, the average minimum entropy is describe as:

$$\tilde{\mathrm{H}}_\infty(W|SS, G^*, G, N, W', \epsilon) \geq m$$

and complete the prove.

Since $\lceil kh_2(\epsilon)\rceil > \lceil \log(1/\exp(-2n\epsilon^2))\rceil$ is a necessary condition to show security of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ (information theoretically). Proposition 3 therefore suggested the following statement:

"*For any input with error parameter $0 < \epsilon \leq \xi$, high Shannon entropy of error rate ϵ is a necessary and sufficient condition to show security of a secure sketch with inputs min-entropy $m \geq 1$ and correcting a total number of error at most $2\epsilon \leq 2\xi$*"

5.1 Security Bound on Secure Sketch

In this section, we consider the security bound on the secure sketch. Formally, this security bound also refer to the best possible security can offer by a secure sketch construction. Particularly, we are interested in the best possible security by using the new sketching and recover algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$.

If a secure sketch allows recovery of the input from some errors with high probability, it must consist of enough information to describe the error pattern. According to Dodis *et al.* [7], in a random error model, under the relaxed correctness notion, describing the outcome of n independent coin flips with probability of error, p requires $nh_2(p)$ bits of entropy. Therefore, the sketch must loss $nh_2(p)$ bits of entropy. They used the Shannon entropy to describe the security bound in this model and assumed the input is random and uniformly distributed. Since $nh_2(p)$ bits of entropy is loss from the sketch, the upper bound residual entropy is thus reduced to $n(1-h_2(p)-o(1))$. larger value of $p \in (0, 1/2)$ results to lower residual entropy.

In our model, the entropy loss can be described by $\lceil \log(1/\exp(-2n\epsilon^2))\rceil + 1$ (see Eq. 9). Observably, $\lceil \log(1/\exp(-2n\epsilon^2))\rceil + 1$ will show higher value (i.e., $\lceil \log(1/\exp(-2n\epsilon^2))\rceil + 1 > nh_2(p)$) with larger ϵ or n. This result suggested a better achievable lower bound to describe the error pattern in the resilient vectors of size n by using $\lceil \log(1/\exp(-2n\epsilon^2))\rceil + 1$ rather than $nh_2(p)$.

In fact, we have shown that, the upper bound residual entropy in our construction is $m + \lceil kh_2(\epsilon)\rceil - \lceil \log(1/\exp(-2n\epsilon^2))\rceil - 1$. Apparently, this residual entropy is always bounded by $m + \lceil kh_2(\epsilon)\rceil$. Given $\lceil kh_2(\epsilon)\rceil \leq \lceil \log(1/\exp(-2n\epsilon^2))\rceil$, entropy loss cannot be avoided, therefore, high min-entropy became a necessary condition to show security for any sources under a family of distributions $\{W_1, \ldots\} \in \mathcal{W}$ over $\mathcal{M}_1$. In viewed of this, meaningful security (e.g., at least one bit) can only be showed over any distribution $W \in \mathcal{W}$ with entropy (i.e., fuzzy min-entropy) larger than the total entropy loss. On the other hand, if $\lceil kh_2(\epsilon)\rceil > \lceil \log(1/\exp(-2n\epsilon^2))\rceil$, our results (see Theorem 4 and Proposition 3) replied that one can always show meaningful security for any random distributions $W \in \mathcal{M}_1$, including the worst case distribution with min-entropy $\mathrm{H}_\infty(W) = m \geq 1$. In other words, we could have $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ to accept any sources with any distribution of min-entropy at least one bit, yet shown not entropy loss with the published sketch.

Table 1 tabulated the security bound for various β-correct probabilistic secure sketch in correcting the error in probability $1-\beta$. To differentiate our proposal from other existing scheme, we stress here we only refer $\beta = \exp(-2n\epsilon^2)$ in our proposal (LSH sketch).

Security Bound for β-Correct Secure Sketch		
Computational	Best possible security	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(1-\beta)$
Computational	FRS sketch(universal hash functions) [28]	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(1/\beta) - \log\log(\mathsf{supp}(W)) - 1$
Computational	Layer hiding hash (strong universal hash function)[29]	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(1/\beta) - 1$
Info. theoretic	**LSH sketch**	$\mathrm{H}_\infty(W) = m \geq 1$ (If $\lceil kh_2(\epsilon)\rceil > \lceil \log(1/\beta)\rceil$)

Table 1: Summary of security bound of β-correct secure sketch in term of fuzzy-min entropy.

6 Reusability

We focus on the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ in this section. First stated by Boyen, 2004 [13], any information theoretical secure sketch or fuzzy extractor must leak certain amount of fresh information about the input for each time it reuses/re-enrolls. The reusability property allows the reuse/re-enrollment of the noisy data with multiple providers. Trivially, if $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}}\rangle$ can show reusability property, it also suggested a reusable fuzzy extractor for uniform random strings generation.

In the context of showing reusability, $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}$ may run in multiple times for enrollment of correlating samples $w_1, w_2, \ldots, w_\gamma$. Each enrollment should return a sketch ss_i which possesses individual security that holds even under the existence of other sketches for $i \in \{1, \ldots, \gamma\}$. Boyens works on assuming a single adversary should be able to perform some perturbation on the original input w^* to yield a list of correlating samples $w_1, w_2, \ldots, w_\gamma$, further gains advantages in recovering w_i from its corresponding sketch ss_i. The works of Boyen on reusability has focused on a particular class of perturbation which is the transitive and isometric permutation applied to w^*. This constraint applied to the perturbation is unlikely in a real and practical scenario. However, his work has encouraged the needs of showing reusability for a secure sketch to offer stronger security guarantee.

Apart from Boyen works, Fuller *et al.*, (2016) [28] provided a modified definition of reusability that covered a more realistic scenario. In their works, they split the adversary into a group of adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$. This group of adversaries implicitly defined different distributions over the published sketch $\{ss_1, \ldots, ss_\gamma\}$. Each sketch is subjected to a particular adversary in the group to show security individually. The act of showing security for a group of adversaries manifested the reusability for independent re-enrollment of the original input with multiple providers that may not trust each other. They utilized set of functions $f_1, \ldots, f_\gamma$

to sample $w', \ldots, w_\gamma$ s.t. $w_i = f_i(w^*, ss_1, \ldots, ss_i)$. These set of functions come with the main property, is to offer fresh min-entropy to the new sample w_i over a particular distribution W_i. The security is defined computationally with fuzzy min-entropy and holds for a large class of family of distributions $\{W_1, \ldots, W_\gamma\}$ over $\mathcal{M}$.

Our intuition of showing reusability for a group of adversary follows the works proposed by Fuller *et at.,* [28]. The goal is to show security to the original sample w^* for different independent re-enrollments come with certain degree of perturbations. It considered a stronger notion of reusability compare to the previous case studied by Boyen and Fuller *et al.,*. It means to show security for any perturbation applied to the input as long as the perturbation is kept within some limited strength, i.e., the maximum number of altered bits is bounded. This notion is more applicable to real case scenario since it does not introduce any assumption on the type of perturbation applied to the input but only provides a bound on it. To do so, we have introduced additional random error $\{e_1, \ldots, e_\gamma\}$, s.t. $\|e_i\| = \lfloor k^*\epsilon_i \rfloor \leq \lfloor k^*\epsilon \rfloor$, i.e., $\epsilon_i \leq \epsilon$ acting as perturbation to the input w^* to sample a list of correlating reading $\{w_1, \ldots, w_\gamma\}$. The usage of random error is better fit to real case scenario, since any perturbation occurs during re-enrollment must cause certain amount of bits flip to the original sample w^*.

To formalize the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ with perturbation parameter $\epsilon_i \leq \epsilon$, we assume an original input w^* is randomly sampled from a metric space $\mathcal{M}_1 = \{0,1\}^{k^*}$, then we apply perturbation on w^* to generate a list of correlated samples $\{w_1, \ldots, w_\gamma\}$ over some random distribution $\{W_1, \ldots, W_\gamma\} \in \mathcal{M}_1$ (parametrized by $\epsilon_i \leq \epsilon$) respectively. Such realization of perturbation in fact can be done straightforwardly with $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w^*, N, G^*, G, \epsilon_i) \rightarrow ss_i$ to output a random sketch $ss_i \in SS_i$ over $\mathcal{M}_2 = \{0,1\}^n$. In such a case, we could have another sample $w' \in W'$ over $\mathcal{M}_1$ that show at least error rate of $\|w_i \oplus w'\| (k^*)^{-1} \geq \xi + \epsilon \geq \xi + \epsilon_i$, where $\|w_i \oplus w'\| \geq t_{(+)}$ is always true by $\epsilon_i \leq \epsilon$ (for $i = 1, \ldots, \gamma$). We aim to characterize the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ by using a group of adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$ comes with unlimited computation power. Formally, each adversary $\mathcal{A}_i : \mathcal{M}_1 \times \mathcal{M}_2 \times \mathbb{F}_2^{n^* \times k^*} \times \mathbb{F}_2^{n \times k} \times [k^*]^n \rightarrow \mathcal{M}_1$ is simply an algorithm that is computationally unbounded to output $w_i \in \mathcal{M}_1$ from a public sketch $ss_i \in \mathcal{M}_2$, with input $w' \in \mathcal{M}_1$, generator matrices $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$ and an integer string $N \in [k^*]^n$. Our formalization uses a second attack running with $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$. Likewise the formulation of Attack, each adversary $\mathcal{A}_i$ is giving full power of choosing any other error parameter $\epsilon'_i \leq \epsilon_i$ in recovering w_i for maximum error tolerance. The second attack is depict on $\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\})$ with input LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}$, N, G^*, G, ϵ and $\mathcal{A}_i$, which is consider as succeeded if at least one adversary $\mathcal{A}_i \in \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$ has successfully recover w_i from the sketch ss_i

```
$\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\})$

 1 : $w^* \leftarrow W$// sample according to some distribution $W \in \mathcal{M}_1$
 2 : for $i = 1 : \gamma$
 3 : $e_i \leftarrow_\$ \{0,1\}^{k^*}$// the weight $\|e_i\| = \lfloor k^* \epsilon_i \rfloor \leq \lfloor k^* \epsilon \rfloor$
 4 : $w_i = w^* \oplus e_i$// $w_i \in W_i$
 5 : $w' \leftarrow W'$
 6 :     if $\|w_i \oplus w'\| \leq t_{(+)}$, repeat step 5 until $\|w_i \oplus w'\| \geq t_{(+)}$
 7 :         if $\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w^*, N, G^*, G, \epsilon_i), w', N, G^*, G, \epsilon'_i) = w_i$
 8 :             Output true
 9 :         else
10 :             Output false
11 :         endif
12 :     endif
13 : endfor
```

The reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ can be generally characterized by the definition below.

Definition 4. *Let $\mu' > 0$, $\epsilon \in [(k^*)^{-1}, \xi]$ and $\epsilon_i \leq \epsilon$ and $i = 1, \ldots, \gamma$. Let $W_i, W', W^* \in \mathcal{M}_1$ and $SS \in \mathcal{M}_2$ be some random variable over $\mathcal{M}_1 = \{0,1\}^{k^*}$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively. Given $N \in [k^*]^n$, an $[n^*, k^*, t^*]_2$ linear code $\mathcal{C}^*$ and an $[n, k, t]_2$ linear code $\mathcal{C}_\xi$ with generator matrices $G^* \in \mathbb{F}_2^{n^* \times k^*}$ and $G \in \mathbb{F}_2^{n \times k}$ respectively, where $\xi = t/n$ and $k^* < n^* \leq k < n$. For all $w' \in W'$, $w_i \in W_i$, $w^* \in W$ and $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w^*, N, G^*, G, \epsilon_i) = ss_i \in SS_i$, we said algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is (ϵ, μ', γ)-reusable if ons has the probability*
$$\Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}) = \mathbf{true}\right] \leq \mu'$$

Theorem 5. *Let a positive integer $m \geq 1$, if both $\mathcal{C}^*$ and $\mathcal{C}_\xi$ are BCH codes with syndrome decoding algorithm f, then, the algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is (ϵ, μ', ∞)-reusable with $\mu' = \left(\frac{2^{-m-1}}{1-2^{-m}}\right)$, when $\lceil k h_2(\epsilon) \rceil > \lceil \log(1/\exp(-2n\epsilon^2)) \rceil$ and $\|w_i \oplus w'\| \geq t_{(+)}$.*

Proof. Clearly, Attack_2 will output a true result if at least one of the adversary $\mathcal{A}_i$ has successfully recover w_i from the sketch ss_i.

Since, $\epsilon'_i \leq \epsilon_i \leq \epsilon$, one obtains the same bound by Eq. 5 and Eq. 6:

$$\Pr\left[\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w^*, N, G^*, G, \epsilon_i), w', N, G^*, G, \epsilon'_i) = w_i\right] \leq 2^{-\lceil \log(1/\beta) \rceil}$$

Therefore, the reusability attack Attack_2 is at least as hard as Attack over single adversary setting. It follows:

$$\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \mathcal{A}) = \mathbf{true}\right]$$
$$\leq \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}, N, G^*, G, \epsilon, \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}) = \mathbf{true}\right] \leq \left(\frac{2^{-m-1}}{1-2^{-m}}\right)$$

with $\mu' \geq \mu = \left(\frac{2^{-m-1}}{1-2^{-m}}\right)$ and complete the prove.

Theorem 5 concluded that the original input $w^* \in W$ can be reused for γ times to generate γ number sketches $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}^*,\mathcal{C}_\xi}(w^*, N, G^*, G, \epsilon_i) = ss_i \in SS_i$ with perturbation parameter $\epsilon_i \leq \epsilon$. Viewed this way, adding error of parameter ϵ larger than the input perturbation $\epsilon_i \leq \epsilon$ while sketching implicitly allows reusability (under the case where all providers are not communicating to each other).

References

1. S. N. Porter, “A password extension for improved human factors,” *Computers & Security*, vol. 1, no. 1, pp. 54–56, 1982.
2. N. Frykholm and A. Juels, “Error-tolerant password recovery,” in *Proceedings of the 8th ACM conference on Computer and Communications Security*. ACM, 2001, pp. 1–9.
3. C. Ellison, C. Hall, R. Milbert, and B. Schneier, “Protecting secret keys with personal entropy,” *Future Generation Computer Systems*, vol. 16, no. 4, pp. 311–318, 2000.
4. A. Juels and M. Sudan, “A fuzzy vault scheme,” *Designs, Codes and Cryptography*, vol. 38, no. 2, pp. 237–257, 2006.
5. A. K. Jain, K. Nandakumar, and A. Ross, “50 years of biometric research: Accomplishments, challenges, and opportunities,” *Pattern Recognition Letters*, vol. 79, pp. 80–105, 2016.
6. C. H. Bennett, G. Brassard, and J.-M. Robert, “Privacy amplification by public discussion,” *SIAM journal on Computing*, vol. 17, no. 2, pp. 210–229, 1988.
7. Y. Dodis, L. Reyzin, and A. Smith, “Fuzzy extractors: How to generate strong keys from biometrics and other noisy data,” in *International conference on the theory and applications of cryptographic techniques*. Springer, 2004, pp. 523–540.
8. A. Juels and M. Wattenberg, “A fuzzy commitment scheme,” in *Proceedings of the 6th ACM conference on Computer and communications security*. ACM, 1999, pp. 28–36.
9. J. Daugman, “Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons,” *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1927–1935, 2006.
10. B. Fuller, X. Meng, and L. Reyzin, “Computational fuzzy extractors,” in *International Conference on the Theory and Application of Cryptology and Information Security*. Springer, 2013, pp. 174–193.
11. E. J. Kelkboom, J. Breebaart, T. A. Kevenaar, I. Buhan, and R. N. Veldhuis, “Preventing the decodability attack based cross-matching in a fuzzy commitment scheme,” *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 107–121, 2011.

12. M. Gomez-Barrero, J. Galbally, C. Rathgeb, and C. Busch, "General framework to evaluate unlinkability in biometric template protection systems," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1406–1420, 2018.
13. X. Boyen, "Reusable cryptographic fuzzy extractors," in *Proceedings of the 11th ACM conference on Computer and communications security*. ACM, 2004, pp. 82–91.
14. M. Blanton and M. Aliasgari, "Analysis of reusability of secure sketches and fuzzy extractors," *IEEE transactions on information forensics and security*, vol. 8, no. 9, pp. 1433–1445, 2013.
15. ——, "On the (non-) reusability of fuzzy sketches and extractors and security in the computational setting," in *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*. IEEE, 2011, pp. 68–77.
16. K. Simoens, P. Tuyls, and B. Preneel, "Privacy weaknesses in biometric sketches," in *Security and Privacy, 2009 30th IEEE Symposium on*. IEEE, 2009, pp. 188–203.
17. Y. Dodis and D. Wichs, "Non-malleable extractors and symmetric key cryptography from weak secrets," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 2009, pp. 601–610.
18. A. Gionis, P. Indyk, R. Motwani *et al.*, "Similarity search in high dimensions via hashing," in *Vldb*, vol. 99, no. 6, 1999, pp. 518–529.
19. V. Guruswami, *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*. Springer Science & Business Media, 2004, vol. 3282.
20. C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
21. M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 380–388.
22. F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*. Elsevier, 1977.
23. E. R. Berlekamp, *Algebraic coding theory*. World Scientific Publishing Co, 2015.
24. W. W. Peterson and E. J. Weldon, *Error-correcting codes*. MIT press, 1972.
25. M. Sudan, "Lecture notes for an algorithmic introduction to coding theory," *Course taught at MIT*, 2001.
26. V. Guruswami, "Introduction to coding theory, lecture 2: Gilbert-varshamov bound," *University Lecture*, 2010.
27. G. D. Forney, "Concatenated codes." *Phd Thesis*, 1965.
28. B. Fuller, L. Reyzin, and A. Smith, "When are fuzzy extractors possible?" in *Advances in Cryptology–ASIACRYPT 2016: 22nd International Conference on the Theory and Application of Cryptology and Information Security, Hanoi, Vietnam, December 4-8, 2016, Proceedings, Part I 22*. Springer, 2016, pp. 277–306.
29. J. Woodage, R. Chatterjee, Y. Dodis, A. Juels, and T. Ristenpart, "A new distribution-sensitive secure sketch and popularity-proportional hashing," in *Annual International Cryptology Conference*. Springer, 2017, pp. 682–710.