How to Correct More Errors in a Secure Sketch

Lai Yen Lung

Monash University Malaysia,
Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor
yenlung.lai@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 biometric data stored in a database, without jeopardizing the user privacy. In addition to this, secure sketch enables fuzzy extractor, by using a randomness extractor to convert the noisy reading w' of its original value w into the same uniform key R. Standard secure sketch should work on all type of available input sources. 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 different 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. Our security defined information theoretically with min-entropy under distribution uncertain setting. In particular, our new technique offers security guarantee for all family of input distributions, as long as the sources possessing "meaningful amount" of min-entropy that is equivalent to the min-entropy of some random distributions over a larger metric space, parametrized by a chosen error correction code.

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 [Por82]), answering several questions for secure access (Niklas Frykholm *et al.*, 2001 [FJ01]) or personal entropy system (Ellison *et al.*, 2000 [EHMS00]), and list of favorite movies (Juels and Sudan, 2006 [JS06]), all are non-uniformly distributed random strings that can be utilized for secrets.

As a solution by utilizing non-uniform input for secrets, it raised 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 [FJ01]), (Jain *et al.*, 2016 [JNR16]). 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) [BBR88], 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) [DRS04] called the fuzzy extractor. Likewise, the fuzzy extractor used two approaches to accomplish the task, which are 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) [JW99] 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 noise level is lower than a specified threshold. Besides, Juels and Sudan (2006) [JS06] have also proposed another construction for metric other than hamming called the fuzzy vault. A fuzzy vault is a vault over a field $\mathbb{F} \times \mathbb{F}$ that protecting an unorder sets, usually, represented as different genuine points. The genuine points reveal a secret which is encoded by using an error correction code. Protection of the genuine points can be done by adding extra chaff points into the vault to conceal the genuine points. Given another set of query points matched with the genuine points in at least some reasonable number, the secret can be reproduced through error tolerance. An improved version of the fuzzy vault is proposed by Dodis *et al.*, (2004) [DRS04], 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 $\mathcal{U}$ [DRS04].

1.1 Existing Issues in Secure Sketch

We here review some existing issues in a secure sketch. As a highlight, these issues are mainly due to the trade-off between security and error tolerance, and they have not considered by the constructions we have mentioned previously. Alternative approaches have introduced to solve these issues recently by showing security computationally [CFP$^+$16] [FMR13], yet diverged from the original definition of a 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. Some other useful discussion on how the low entropy sources must be taken into consideration when constructing a fuzzy extractor can be found in [CFP+16], by Canetti *et al.*, (2016). One typical example of a source with more error than entropy refers to the commonly known biometric feature - IrisCode (Daugman, 2006) [Dau06]. The IrisCode is said to provide entropy of 249 bits. Whereas, the IrisCodes generated from the same user of each 2048 bits have showed 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 $\mathcal{W} = \{W_1, \dots, W_\gamma\}$. 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. In some condition, the source may only consist of distributions where all points are far apart with min-distance d. Then, this source must possess a certain quantity of min-entropy m over d, which identifies the predictability of $w' \in W$. Viewed this way, the distributions of the source determine its maximum tolerance distance $t \leq d$ with min-entropy m over the family of distributions $\mathcal{W}$. In particular, given a source with min-entropy m, larger min-distance between the points allows larger tolerance distance, which also means, more entropy loss can be compensated by higher min-entropy. However, in some worst scenario, the points may be distributed very close to each other. Therefore, the value of t is preferably to be small as well. For any point $w' \in W$ over this 'worse case' distribution W, and one has set $t > d$, the sketch must lose entropy by means of the number of similar points within distance t for error tolerance. The entropy loss of the sketch would be bounded that is proportional to this value. Under the case when $t > d$, the sketch is said to loss all min-entropy and show no security (e.g., more error than entropy).

Fuller *et al.*, (2013) [FMR13] have showed that under the event when the input distribution is precisely known and the security is defined computationally, 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':

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

where $B_t(w')$ denoted a hamming ball of radius t around w'. Conceivably, the fuzzy min-entropy is equivalent to the residual entropy, it can be 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, due to error tolerance.

However, 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. This leads to another problem called *distribution uncertainty*.

The distribution uncertainty problem potentially to be resolved by showing security to a family of distributions rather than a single distribution, which can be easily achieved by using the traditional way of measurement with min-entropy, e.g., min-entropy minus the loss. Most importantly, the notion of min-entropy has considered all possible distributions, included the worst case distribution over error tolerance distance t, which also known as the worst case entropy. In this regard, measuring the entropy loss with min-entropy certainly captured more relevance security property for a secure sketch. Nonetheless, doing so will reduce to the precedent more error than entropy problem which is intended to be resolved by using fuzzy min-entropy measurement.

Reusability[1] e-usability property is introduced by Boyen (2004) [Boy04]. 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

[1] The reusability property is different to the unlinkability property [KBK+11] [CS08] [GBGRB18]. Unlinkability property prevents an adversary from differentiating whether two enrollments correspond to the same physical source, which is not focus in this work.

using w, he/she must provide slightly different reading w_i due to the noise. 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 $ss_1, ss_2, \dots, ss_\gamma$. 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 [Boy04] [BA13] [BA11] [STP09].

1.2 Our Contributions

We highlighted our main contributions as follow:

Correcting more errors with 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. For this reason, we noticed merely relying on fuzzy min-entropy of single tolerance distance t' is insufficient; additional property is required to correct more errors in a source.

Consider another variable 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 ϕ'. Suppose Φ is 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, we meant to look for the fuzzy min-entropy of a variable Φ that is defined by a larger hamming ball B_t, but it comes with an additional property: only the points outside the smaller ball $B_{t'}$ are considered. In brief, if one can show substantive fuzzy min-entropy for every point outside the ball $B_{t'}$, it implies more errors can be corrected over larger tolerance distance $t > t'$. Otherwise, the average fuzzy min-entropy must offer security according to the maximized probability for a variable $\Phi \in B_t(\phi')$ within distance t that is outside the ball $B_{t'}$, by fuzzy min-entropy definition.

Undoubtedly, correcting more errors means higher entropy loss. Therefore, in some 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'. This can be argued by knowing some values outside the ball $B_{t'}(w')$ must add advantages of predicting a value inside the ball $B_{t'}(w')$. In addition to this, since the min-entropy should be lower bounded to the fuzzy min-entropy, the minimum security can offer by average fuzzy min-entropy over t implies the minimum entropy loss of the min-entropy over the worst case distribution. In light of this, the average-fuzzy min-entropy is useful for better monitoring the loss of the min-entropy measured under smaller tolerance distance while providing optimal resilience. We obtain its definition by merely combined the average min-entropy and fuzzy min-entropy notions.

Showing info. theoretic security with min-entropy over larger metric space: 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. It also shows security to all family of input distribution, without putting extra stringent distribution requirement to the sources, i.e., by 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 length itself [DW09]. 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 that offers info. theoretical security, which frees from the stiff constraint where the source entropy must be at least half of its input length. The new construction is capable of achieving security bound that merely depends upon the input entropy rather than its input length. Notably, it shows the best possible

[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

security which is at most half of the input entropy could offer (i.e., $m/2$), regardless of its input length. Our construction relies on *Local Sensitive Hashing (LSH)* to generate a resilient vector pair (trivially, a pair of longer strings with resilience property) for sketching and recover instead of using the original input string. Doing so would allow us to apply the average fuzzy min-entropy notion and correct more errors over a larger metric space. In fact, in our exposition, we show that the min-entropy of the resilient vectors is bounded by the min-entropy of the sources regardless to any chosen error correction code (not parametrized by the code length or the input length). Our works supported a statement: high min-entropy is necessary and sufficient for a source to show meaningful security. On the other hand, high input length is necessary for better error tolerance.

Reusable secure sketch: Apart from this, the new construction offers extra security property, which is the reusability. In the beginning, our design is meant to provide better security bound to the secure sketch, through the insertion of additional random error during the sketching phase. Eventually, we find out 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 $\mathcal{M}_1 = \{0,1\}^l$, and $\mathcal{M}_2 = \{0,1\}^n$ denoted two different sizes of metric spaces with $n > l$. 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 counts the number of non-zero elements, and $\oplus$ is the addition modulo two operation (XOR). Besides, the error rate of w and w' is denoted as $\|w \oplus w'\| \,/\, |w|$ which is simply the normalized hamming distance, given their size (length) $|w| = |w'|$. For error correction code notation, since we are more interested in tolerating the errors of a codeword c', 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 $\xi = tn^{-1}$ over larger binary metric space $\{0,1\}^n$. At the same point, we let $t_{(+)} = \lfloor (\xi + \epsilon) l \rfloor$ and $t_{(-)} = \lfloor (\xi - \epsilon) l \rfloor$ to describe the error tolerance distances over the smaller binary metric space $\{0,1\}^l$, with some error parameter $\epsilon \in (0, 1/2 - \xi)$.

Main idea: Suppose Alice wishes to conceal a noisy non-uniform string $w \in \{0,1\}^l$ while allows exact recovery of w from another noisy string $w' \in \{0,1\}^l$ 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 use of error correction code for conventional secure sketch generation, but comes with additional random error adding to the noisy input w' during recovery phase. The random error can be added by simply an XOR operation in between w' with some random error vector $e \in \{0,1\}^l$ to form $w_e = w' \oplus e$. Given a $[n,k,t]_2$ code $\mathcal{C}_\xi$ is chosen over $\{0,1\}^n$, in contrary to direct encoding w with $\mathcal{C}_\xi$, Alice encodes a longer string $v \in \{0,1\}^k$ by padding w with additional random bits string $r \in \{0,1\}^{k-l}$ drawn uniformly at random, i.e., $v = w \| r$. The output of the encoding process is a codeword $c \in \mathcal{C}_\xi$. After this, she 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 $w'_e, w \in \{0,1\}^l$ through LSH. The resilient vectors offer resilience for the recovery of w from w' if $\|\delta\| \leq t$.

Likewise the code-offset construction [JW99], 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); our construction concealing a non-uniform input padded with additional random bits. Therefore the concealed object is not entirely random and uniform in our case. 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 as she like, but with new liberty, i.e., the sizes of the concealed object and output sketch have not bounded but parametrized by the selected $[n,k,t]_2$ code $\mathcal{C}_\xi$. Thirdly, of course, our operation comes with additional random error added to the input w' during recovery.

In our work, for resilient vector generation, we only focus on a particular LSH family called hamming-hash [GIM$^+$99]. 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 $[l] = \{1, \ldots, l\}$. For Alice with $w \in \{0,1\}^l$ and Bob with $w' \in \{0,1\}^l$. Alice and Bob agreed on this strategy as follow:*

1. *They are told to each other a common random integer $N \in [l]$.*
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. *A hamming hash strategy is a LSH with similarity function $S(w, w') = 1 - \|w \oplus w'\| l^{-1}$.*

Theorem 1 concluded that Alice and Bob always win with probability described as $P = 1 - \|w \oplus w'\| l^{-1}$. 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 $w, w' \in \{0,1\}^l$ respectively by hamming hash strategy with a random integer string $N \in [l]^n$, the expected hamming distance is $\mathbb{E}[\|\phi \oplus \phi'\|] = n \|w \oplus w'\| l^{-1}$.*

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 described as:

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

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, $\|\delta\| \sim$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, $\mathbb{E}[\|\delta\|] = n(1-P) = n \|w \oplus w'\| \, l^{-1}$ 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 $e \in \{0,1\}^l$ of weight $\|e\| = l\epsilon$ to the inputs, where $\epsilon \in (0, 1/2 - \xi)$ (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 shifted 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. To avoid notation clutter, we always refer to the resilient vectors generated from LSH hamming using the same random integer string $N \in [l]^n$. The corollaries are given as follow.

Corollary 1. *Let W and Φ be some random variable over $\{0,1\}^l$ and $\{0,1\}^n$ respectively, let $\xi \in (0, 1/2)$ be the tolerance rate of a $[n, k, t]_2$ code $\mathcal{C}_\xi$ and $\epsilon \in (0, 1/2 - \xi)$ be the error parameter. 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 $t_{(-)} = \lfloor (\xi - \epsilon) l \rfloor$ and $t > t_{(-)}$, given a variable $W \in B_{t_{(-)}}(w')$, then, one has the average minimum probability to find any variable $\Phi \in B_t(\phi')$ described as $\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr\left[\Phi \in B_t(\phi') \,\middle|\, W \in B_{t_{(-)}}(w')\right]\right] \geq 1 - \exp(-2n\epsilon^2)$.*

Proof. For $W \in B_{t_{(-)}}(w')$, it means that any string $w \in W$ must show an error rate of $\|w \oplus w'\| l^{-1} \le \xi - \epsilon$. Based on Theorem 2, w can be used to produce its corresponding resilient vector $\phi \in \varPhi$ that shows an expected offset with ϕ' described as $\mathbb{E}[\|\phi \oplus \phi'\|] = \mathbb{E}[\|\delta\|]$ s.t. $\mathbb{E}[\|\delta\|] \le t - n\epsilon$ (by multiplying both sides of the inequality with n). It follows, there will be a minimum value of $t_{\min}$ s.t. $t_{\min} = \mathbb{E}[\|\delta\|] + n\epsilon$. Therefore, By using *Hoeffding's inequality*, one able to calculate the average minimum probability:

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

and complete the prove. □

Corollary 2. *Let W and $\varPhi$ be some random variable over $\{0,1\}^l$ and $\{0,1\}^n$ respectively, let $\xi \in (0, 1/2)$ be the tolerance rate of a $[n,k,t]_2$ code $\mathcal{C}_\xi$ and $\epsilon \in (0, 1/2 - \xi)$ be the error parameter. Suppose a resilient vector $\phi' \in \varPhi$ is generated from strings $w' \in W$. For two hamming ball $B_t(\phi')$ and $B_{t_{(+)}}(w')$ of radius $t_{(+)} = \lfloor(\xi + \epsilon)l\rfloor$ and $t > t_{(+)}$, given a variable $W \notin B_{t_{(+)}}(w')$, then, one has the average maximum probability to find any variable $\varPhi \in B_t(\phi')$ described as $\mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr\left[\varPhi \in B_t(\phi') \;\middle|\; W \notin B_{t_{(+)}}(w')\right]\right] \le \exp\left(-2n\epsilon^2\right)$.*

Proof. This proof is instantiated from the proof of Corollary 1. For $W \notin B_{t_{(+)}}(w')$, it means that any string $w \in W$ must show error rate of $\|w \oplus w'\| l^{-1} \ge \xi + \epsilon$. More precisely, $\|w \oplus w'\| \ge \lfloor(\xi + \epsilon)l\rfloor > t_{(+)}$. According to Theorem 2, w is capable to produce its corresponding resilient vector $\phi \in \varPhi$ that will show an expected offset with ϕ' described as $\mathbb{E}[\|\delta\|] \ge t + n\epsilon$. Thus, there will be a maximum value of $t_{\max}$ s.t. $t_{\max} = \mathbb{E}[\|\delta\|] - n\epsilon$. Therefore, By using *Hoeffding's inequality*, one able to calculate the average maximum probability, by symmetry:

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

and complete the prove. □

The results obtained from Corollary 1 and Corollary 2 imply the following statement: Once the error is introduced into the input, the probability to find any resilient vector $\phi' \in \varPhi$ close to its original reading ϕ within the ball $B_t(\phi')$ will be bounded due to the error effect. 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 average fuzzy min-entropy described as

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

by definition.

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 $\mathsf{dis} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+ = [0, \infty)$. The distance function can take any non-negative real values and obey symmetric e.g., $= \mathsf{dis}(A,B) = \mathsf{dis}(B,A)$, and triangle inequality, e.g., $\mathsf{dis}(A,C) \le \mathsf{dis}(A,B) + \mathsf{dis}(B,C)$.

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, $\max\limits_{w} \Pr[W = w]$ 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

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

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 w' of W, the predictability of W is now conditioned as $\max\limits_{w} \Pr[W = w \mid W' = w']$. 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 [DRS04] A $(\mathcal{M}, m, \tilde{m}, t)$-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 high probability $1 - \beta$. 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 [Gur04]: 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 [Gur04]: 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 $Hx = 0^n\}$ 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*.

Local Sensitive Hashing (LSH) [Cha02] Given that $P_2 > P_1$, while $w, w' \in \mathcal{M}$, and $\mathcal{H} = h_i : \mathcal{M} \to U$, where U[3] is the hashed metric space depends to similarity function defined by S and i refers to the number of hash functions h_i. A local sensitive hashing is a probability distribution on a family $\mathcal{H}$ of hash functions such that $P_{h\in\mathcal{H}}[h(w) = h(w')] = S(w, w')$. With a similarity function S define on the collection of 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 is the hashing of object collection w and w' by means of multiple hash functions h_i. The use of h_i enables decent approximation of the pair-wise distance of w and w' in terms of collision probability. LSH ensures that w and w' with high similarity render higher probability of collision in the hashed domain; on the contrary, the data points far apart each other result in a lower probability of hash collision.

3 Our 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 described later. Generally, the hamming-hash algorithm $\Omega^{\mathsf{ham-h}} : \mathcal{M}_1 \times [l]^n \to \mathcal{M}_2$ is an iterative process through repeating the hamming hash strategy (steps 1 and 2) up to $n > 1$ times. It serves to sample the input binary string of size l into a longer binary string a.k.a resilient vector of size $n > l$.

Given input $w \in \{0,1\}^l$, and $N \leftarrow_{\$} [l]^n$, the LSH-hamming hash algorithm described 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] The notation used here is different with our exposition. In our exposition, $\mathcal{M} = \mathcal{M}_1$ and $U = \mathcal{M}_2$, where $|\mathcal{M}_1| < |\mathcal{M}_2|$. In traditional LSH, $|U|$ is usually smaller than $\mathcal{M}$ for different objectives, i.e., fast similarity search.

3.2 LSH-Sketching

We denote the LSH-sketching algorithm that employed the hamming-hash algorithm, Ω and a $[n,k,t]_2$ code $\mathcal{C}_\xi$ with parity check matrix H[4] as $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$.

For sketching, one is required to generate a resilient vector ϕ by using the LSH hamming hash algorithm. The size of the resilient vector must same as the sampled codeword c. Then, the sketch ss can be constructed by simply perform an XOR operation, i.e., $ss = c \oplus \phi$. The sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$ used input w, N, H described as follow:

```
$\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w,N,H)$
$r \leftarrow_{\$} \{0,1\}^{k-l}$ // sample $r$ uniformly at random
$v = w\|r$;
$c = Hv$;
$\phi \leftarrow \Omega^{\mathsf{ham-h}}(w,N)$
$ss = c \oplus \phi$
return $(ss,N,H)$
```

All steps on $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w,N,H)$ can be done in $\mathcal{O}\left(n^2\right)$. Notably, the size of v and ss are now depend upon the chosen code $\mathcal{C}_\xi$ (parametrized by k and n respectively). Often, the XOR operation $c \oplus \phi$ works perfectly under the case when the size of the codeword and the resilient vector are equal, i.e., $|c| = |\phi| = n$.

Assuming in a scenario that is without any random bits padding, direct encoding w must add $n-l$ number of redundant symbols for $|c| = |\phi| = n$ to hold, which will lead to exhaustive entropy loss when the sketch is published. As a solution to this, we padded the input to form a longer string v before encoding takes place, hence reduced the number of redundancy. Doing so can minimize the entropy loss from the sketch during encoding phase.

In fact, the idea of random bits padding for secure sketch has been earlier proposed by Woodage *et al.* [WCD+17] for password typo correction. Their works padded random bits on shorter sketches that protecting the same password. The effort required to recover the password from all sketches of the same size is increased, so, it reduced the entropy loss.

Noting that for any random bit padded input $v \in V$ over some random distribution V, our strategy should introduce a changes over the input metric space form $\mathcal{M}_1$ to $\mathcal{M}_2$ for $W \in \mathcal{M}_1$ and $V \in \mathcal{M}_2$ respectively. In fact any secure sketch construction technique that allows changing in between metric space can be viewed as *biometric embedding*, first identified by Dodis *et al.,* (2004) [DRS04]. Generally, biometric embedding used a transformation function f_b to transform the input $w, w' \in \mathcal{M}_a$ over a metric space $\mathcal{M}_a$ to another metric space $\mathcal{M}_b$, i.e., $f_b(w), f_b(w') \in \mathcal{M}_b$. The transformation function itself must come with some useful properties for secure sketch construction (see Section 4.3 in [DRS04] for more details). Like wise, our construction can be considered as an realization of biometric embedding with resilient vector where the achievable security is bounded by the input min-entropy over the original metric space before transformation ([DRS04], Lemma 4.7).

3.3 LSH-Recover

For recovery, given any random error $e \in \mathsf{supp}(\chi)$ over some random distribution χ parametrized by $\epsilon \in (0, 1/2 - \xi)$, i.e., $\|\chi\| = l\epsilon$. Suppose one wishes to recover w from another string $w' \in \{0,1\}^l$. He/she needs to provide another resilient vector ϕ'. This resilient vector can be generated by using the same hamming hash algorithm Ω with inputs $w'_e = w' \oplus e$ (after adding the error e) and the public known integer string N. The offset is manifested by the way of measuring the hamming distance on the resilient vectors pair, $\delta = \phi \oplus \phi'$. Often, we allow the recovery algorithm to run iteratively for all $e_i \in \mathsf{supp}(\chi)$ (for $i = 1, \ldots, |\mathsf{supp}(\chi)|$).

[4] Sometimes, we replace H with G if generator matrix is desired for code $\mathcal{C}_\xi$

We denote the LSH-recover algorithm that employed the hamming-hash algorithm, Ω, and a $[n,k,t]_2$ code $\mathcal{C}_\xi$ with parity check matrix H and a decoding algorithm f as $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$. The recover algorithm $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ used input string ss, w', N, ϵ and H to recover w is described as follow:

```
$\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, H, \epsilon)$
Set $\chi = \{0,1\}^l$// $\|\chi\| = \epsilon l$
for $i = 1, \ldots, |\mathsf{supp}(\chi)|$
  $\mathcal{L} \leftarrow \emptyset$
  $e_i \leftarrow_\$ \mathsf{supp}(\chi)$
  $w'_{e_i} = w' \oplus e_i$
  $\phi'_i \leftarrow \Omega^{\mathsf{ham-h}}(w'_{e_i}, N)$
  $c'_i = ss \oplus \phi'_i$// also $ss \oplus \phi'_i = c \oplus (\phi \oplus \phi'_i)$
  $c_i \leftarrow \mathsf{f}(c'_i, H)$
  $v_i \leftarrow H^{-1} c_i$
  $w_i \leftarrow v_i$// look for $w_i$ from first $k$ bits of $v_i$
  $\mathcal{L} \bigcup w_i$
endfor
return $\mathcal{L}$
```

If the final decoding process $\mathsf{f}(c, H)$ is successful, the algorithm returns a list of outputs $\mathcal{L}$ where $w \in \mathcal{L}$. Else, it will output all wrong results and $w \notin \mathcal{L}$.

A brief description of the recovery mechanism is given as follow. Saying that, the recovery is allowed to success when the inputs' error rate $\|w \oplus w'\|\, l^{-1} \leq \xi$ by some error correction code $\mathcal{C}_\xi$. Suppose Bob has intercepted with a sketch $ss = c \oplus \phi$. Firstly, he has to generate a resilient vector $\phi' \leftarrow \Omega^{\mathsf{ham-h}}(w'_e, N)$. The hamming weight of the offset can be conveniently represented as $\|\delta\|$. By means of the similarity preservation property over LSH, the offset, δ is expected to be low as well if w and w' are close to each other. Expressly, if w and w' are close enough, one would have $\|\delta\| \leq t$, with distance t specified by the error correction code $\mathcal{C}_\xi$, where $\xi = t/n$. Eventually, Bob can perform $ss \oplus \phi'$ to output the nearest codeword c'. The errors over c' can be tolerated by means of error correction with code $\mathcal{C}_\xi$ with decoding function f.

When comes into decoding, it follows that $\mathsf{f}(c', H) = \mathsf{f}(c \oplus \delta, H) = \mathsf{f}((c \oplus \phi) \oplus \phi', H) = \mathsf{f}(c \oplus (\phi \oplus \phi'), H)$. If $\|\phi \oplus \phi'\| = \|\delta\| \leq t$, the decoding will success and its efficiency follows the decoding algorithm f itself. Thereafter, v can be recovered successfully and so w by looking at the first l symbols of v. Above process is repeated for $i = 1, \ldots, \mathsf{supp}(\chi)$ iterations to list all possible solutions for w over a list $\mathcal{L}$.

4 Resilience

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

Obviously, the resilience of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is bounded below the resilience of the selected code $\mathcal{C}_\xi$. Choosing a 'good' code with a high value of ξ is non-trivial, this is because different code $\mathcal{C}_\xi$ 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) [MS77], and Peterson and Weldo, (1972) [Ber15]. 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.

For the seek of simplicity, we combined the results of Eq. (1) and Eq. (2). Formally, we let $\beta = \exp\left(-2n\epsilon^2\right)$. Thus, $\mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr\left[\Phi \in B_t(\phi') \;\middle|\; W \notin B_{t_{(+)}}(w')\right]\right] \leq \beta$. accordingly. On the other hand, the average minimum probability to find $\Phi \in B_t(\phi')$ can be represented as $\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr\left[\Phi \in B_t(\phi') \;\middle|\; W \in B_{t_{(-)}}(w')\right]\right] \geq 1 - \beta$.

Further simplification is done by describing the term *overwhelming* given the value of $1-\beta$ comes with some negligible quantity β. As we shall see, negligible value of β means substantial average fuzzy min entropy, since $\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t,\infty}(\Phi | W \notin B_{t_{(+)}}(w')) \geq -\log(\beta)$. In view of this, apart from the security it could offer with, the average fuzzy min entropy is promoting higher resilience.

Our explication of resilience evinced by the *completeness* of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. It captured the scenario when the players are honest, which is defined under the following definition

Definition 1. *Let W and Φ be some random variable over a metric space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $w, w' \in W$, $N \in [l]^n, \epsilon \in (0, 1/2 - \xi)$, an $[n,k,t]_2$ linear code $\mathcal{C}_\xi$ with $\xi = tn^{-1} \in (0, 1/2)$ and parity check matrix $H \in \mathbb{F}^{(n-k)\times n}$. For a sketch ss generated through $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N) = ss$, then the probability for $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, H, \epsilon) = w$ is overwhelming if the error rate $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$. We said $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is complete in (ξ, ϵ)-fuzziness if above statement holds.*

We hereby provide a proposition with proof to characterize the resilience property of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. For efficiency purpose, we will focus on particular $[n,k,t]$ code $\mathcal{C}_\xi$ named BCH code with efficient decoding algorithm f via algebric method, i.e., syndrome decoding [PW72].

Proposition 1. *If syndrome decoding algorithm f is used for an $[n,k,t]$ BCH code $\mathcal{C}_\xi$, an LSH-sketching and recover algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is complete in (ξ, ϵ)-fuzziness if n is sufficiently large.*

Proof. Recall that any offset as $\delta \in \{0,1\}^n$ with $\|\delta\| \leq t$ is required for a successful decoding. For an error correction threshold $t > 0$, the usage of syndrome decoding, f can decode the corrupted codeword, c' if $\|\delta\| \leq t$, described as $\mathsf{f}(c', H) = \mathsf{f}(c \oplus (\phi \oplus \phi'), H) = c$. Eventually, one has $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N, H, \epsilon) = w$.

To claim our completeness, we utilize the result in Corollary 1. Suppose $\|w \oplus w'\| l^{-1} \leq \xi$, with additional random error e added to the input w', such as $\|w \oplus w' \oplus e\| l^{-1}$. The error included will lead to the changes in the final error rate between w and w', to either $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$ or $\|w \oplus w'\| l^{-1} \leq \xi + \epsilon$. Focusing on the case when $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$, one has:

$$\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr\left[\Phi \in B_t(\phi') \;\middle|\; W \in B_{t_{(-)}}(w')\right]\right] \geq 1 - \beta$$

Observe that $1 - \beta$ 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 a $[n,k,t]_2$ BCH code $\mathcal{C}_\xi$, under the scenario where $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$, or formally, it also equivalent to the case when $\|w \oplus w'\| \leq t_{(-)}$, the offset can be tolerated with overwhelming probability if one has the value of n is sufficiently large for any decoding algorithm f. The secure sketch itself is considered as efficient by the efficiency of syndrome decoding f itself.

It is useful to have an example to show how our results can be applied practically with a particular classes of error correction code- BCH code, which is an efficient one.

Example 1. Let $w, w' \in \{0,1\}^l$, $l = 100$. Suppose one wishes to correct some errors say 10 bits. It means $10/l = 0.1 = \|w \oplus w'\| l^{-1}$. Suppose a $[1023, 101, 175]_2$ BCH code with $\xi = 175/1023 = 0.1711$ is used for encoding/decoding. To show resilience (correctness), recall the completeness hold if $\|w \oplus w'\| \leq \lfloor (\xi - \epsilon) l \rfloor$. It follows that one can calculate the weight of random error $\|e\| = l\epsilon$ following $10 \leq \lfloor (0.1711 - \epsilon)100 \rfloor$, thus $\epsilon \leq 0.0711$ and $\|e\| \leq \lfloor (0.0711)100 \rfloor = 7$ bits. The error can be corrected with overwhelming probability $1 - \exp\left(-2n\epsilon^2\right) = 1 - 3.22 \times 10^{-5}$.

In fact the syndrome decoding algorithm f itself will always success without error (success with probability one, i.e., $\beta = 0$) if $\|\delta\| \leq t$ [Ber15]. However, adding random error eventually boils down this perfect correctness notion into probabilistic correctness notion. Precisely, the error added into the input w would affect the distance of the resilient vectors pair $\|\phi \oplus \phi'\|$ described by their collision probability. Therefore the distance over the resilient vector will be probabilistic as well.

4.1 Correcting More Error via List-Decoding

Recall that, once the error added to the input w', the final error rate would be $\|w \oplus w'\| l^{-1} \leq \xi \pm \epsilon$. To correct the error with overwhelming probability, the completeness statement (hold only if $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$) implies the added random error must come with error parameter $\epsilon \leq \xi$ for a constant value of error tolerance rate $\xi \in (0, 1/2)$. In fact, above result demonstrating a limited amount of inputs' error rate $\|w \oplus w'\|$ could be corrected by the code $\mathcal{C}_\xi$. For instance, since $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$ must hold by argument of completeness, then, the total number of error rate needs to be corrected will be $\|w \oplus w'\|\, l^{-1} + \epsilon \leq \xi$, thus lesser inputs' errors can be corrected in this scenario.

Conversely, one can actually correct more error with any code $\mathcal{C}_\xi$ when $\|w \oplus w'\| l^{-1} \leq \xi + \epsilon$. To be specific, given some inputs' error rate of $\|w \oplus w'\|$, it is possible to have $\|w \oplus w'\| l^{-1} \leq \xi + \epsilon$ after adding the random error. Therefore, the inputs' error rate can be reduced to $\|w \oplus w'\| l^{-1} - \epsilon \leq \xi$. Saying so, we have the soundness of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ to correct a total amount of error rate $\leq \xi$ by code $\mathcal{C}_\xi$.

The definition below captured the soundness of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ in errors with overwhelming probability and efficiently under the event when $\|w \oplus w'\| l^{-1} \leq \xi + \epsilon$. As all steps on $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, H)$ can be done efficiently in $\mathcal{O}\left(n^2\right)$, our focus will mostly on the recovery algorithm itself $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$.

Definition 2. *Let W and $\varPhi$ be some random variable over a metric space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $w, w' \in W$, $N \in [l]^n, \epsilon \in (0, 1/2 - \xi)$, an $[n,k,t]_2$ linear code $\mathcal{C}_\xi$ with $\xi = tn^{-1} \in (0, 1/2)$ and parity check matrix $H \in \mathbb{F}^{(n-k)\times n}$. For a sketch ss generated through $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N) = ss$, We said $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ is efficient if it can run in polynomial time and correct the error rate of $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ with overwhelming probability.*

We provide a proposition with proof to show that $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ can be done in an efficient manner follows Definition 2.

Proposition 2. *For any polynomial time decoding algorithm f used for an $[n,k,t]$ code $\mathcal{C}_\xi$, an LSH-recover algorithm $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ is efficient for sufficiently large n.*

Proof. We first argue on the statement where the error can be corrected with overwhelming probability. In particular, given the final error rate (after error added) holds as $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$, same interpretation can be used by Eq. 1 and replacing $\epsilon = -\epsilon$, which yields:

$$\begin{aligned} &\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr\left[\varPhi \in B_t(\phi') \,\middle|\, W \in B_{t_{(+)}}(w')\,\right]\right] = \min_{t = t_{\min}} \Pr\left[\|\delta\| \leq t \,\middle|\, \|w \oplus w'\| \leq t_{(+)}\,\right] \\ &\geq 1 - \exp\left(-2n\epsilon^2\right) \end{aligned} \tag{4}$$

Eq. 4 concluded that $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ can correct the errors with overwhelming probability under the case $\|w \oplus w'\|\, l^{-1} - \epsilon \leq \xi$ and $\epsilon = -\epsilon$ if n is sufficiently large. In other words, the error can be corrected with overwhelming probability only if the introduced error parameters $\epsilon = -\epsilon$ is reducing the original input errors rate $\|w \oplus w'\|\, l^{-1}$.

For efficiency claim, given any error parameter ϵ, we have a total number of $|\mathsf{supp}(\chi)| = \binom{l}{l\epsilon}$ possible ways to describe all different combination of the random error $e_i \in \mathsf{supp}(\chi)$. Noting that all these possible ways of description should include both scenarios when $\epsilon = +\epsilon$ (inputs' errors rate increasing) or $\epsilon =$

$-\epsilon$ (inputs' errors rate decreasing). Therefore $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ maximally run in $\binom{l}{l\epsilon}$ iterations. It follows that $\Pr\left[\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon\right] = \frac{\binom{\|w\oplus w'\|}{l\epsilon}}{\binom{l}{l\epsilon}}$. Let $y = l - \|w \oplus w'\|$, then

$$\frac{\binom{\|w\oplus w'\|}{l\epsilon}}{\binom{l}{l\epsilon}} = \frac{\binom{l-y}{l\epsilon}}{\binom{l}{l\epsilon}} > (1 - \frac{y}{l - l\epsilon})^{l\epsilon} \approx \exp(-\frac{y\epsilon}{1-\epsilon}) = \frac{1}{\mathsf{poly}(\epsilon)} \tag{5}$$

Briefly, after $\mathsf{poly}(\epsilon)$ iteration, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ would success and output w with overwhelming probability $1-\beta$. The efficiency of the decoding algorithm f itself follows. □

Proposition 2 concluded that given a $[n,k,t]_2$ BCH code $\mathcal{C}_\xi$, under the scenario where $\|w \oplus w'\| l^{-1} \leq \xi + \epsilon$, or formally, it also equivalent to the case when $\|w \oplus w'\| \leq t_{(+)}$, the offset can be tolerated with overwhelming probability after $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ has run in $\mathsf{poly}(\epsilon)$ iterations and n is large enough. The secure sketch itself is considered as efficient if f is efficient.

Remark that $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ could return a list of solutions $\mathcal{L}$ after running $\mathsf{poly}(\epsilon)$ iterations, we then called it as list-decoding for recovery of w.

We give another example to showcase how $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ works according to our claim follows Proposition 2.

Example 2. Let $w, w' \in \{0,1\}^l$, $l = 100$. Suppose one wishes to correct some errors say 30 bits. It means $30/l = 0.3 = \|w \oplus w'\| l^{-1}$. Suppose a $[1023, 101, 175]_2$ BCH code with $\xi = 175/1023 = 0.1711$ is used for encoding/decoding. To show resilience (correctness), under the case when $\|w \oplus w'\| \leq \lfloor(\xi + \epsilon)l\rfloor$. It follows that one can calculate the weight of random error $\|e\| = l\epsilon$ following $30 \leq \lfloor(0.1711 + \epsilon)100\rfloor$, thus $\|e\| \geq \lceil l(0.3 - 0.1711)\rceil = 13$ bits. By Eq. 5, after $-\log \frac{\binom{100}{13}}{\binom{30}{13}} = 2^{26}$ number of iterations, the error can be corrected with overwhelming probability $1 - \exp\left(-2n\epsilon^2\right) = 1 - 1.72 \times 10^{-15}$.

Intuitively, we introduce additional 13 bits of information while recovery phase by adding extra random error to facilitate the decoding process. To be specific, since 30 bits of error contributed to 0.3 error rate (equivalent to ≈ 307 bits of errors over $\mathcal{M}_2$) which is required to be corrected by the chosen BCH code $\mathcal{C}_{175/1023}$. However, since $\mathcal{C}_{175/1023}$ is only capable to correct $t = 175$ bits of error, adding extra random error of 13 bits on the inputs is possible to reduce the original error by 0.13 (equivalent to ≈ 134 bits of errors over $\mathcal{M}_2$). Therefore, the total errors that need to be corrected over $\mathcal{M}_2$ are now become $307 - 134 = 173$ bits, which is suffice and easy to be handled by using $\mathcal{C}_{175/1023}$.

4.2 Error Correction up to Shannon Bound

In the previous section, we have demonstrated the resilience of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$, in term of the probability in correcting the errors. 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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$. Formally, we called this as the resilience bound of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$.

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.

Principally, 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 return 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. Formally, once the error pattern of the input sources is precisely modelled and known, one can easily determine the min-distance d between the codewords so that the decoding process must succeed without any error. On the other hand, computational hardness assumption must be applied to show meaningful security with fuzzy min-entropy in case of it is not infinite.

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 [Sud01]) 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 $4/n$ 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 ([Gur10], 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 decoding process 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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ to free from the Plotkin bound and allows it to correct more errors by increment of n.

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 [Sha01]. 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:

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

where $h_2(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function of error rate p. 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 $1/2$. Therefore, we obtain

Proposition 3. *With sufficiently large n, there exists a $[n,k,t]_2$ code $\mathcal{C}_\xi$ for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ to correct the errors with overwhelming probability as long as the total error rate satisfy $\xi + \epsilon < 1/2$.*

Proof. This proof is straightforward by using the Shannon noisy channel coding theorem over binary symmetric channel. Formally, the total error rate including the introduced random error of parameter e can be described as:

$$\|w \oplus w'\|\, l^{-1} = \xi \pm \epsilon$$

By completeness itself, the error can be corrected with overwhelming probability when $\|w \oplus w'\|\, l^{-1} = \xi - \epsilon$ for sufficiently large n. On the other hand, by the the efficiency of the algorithm $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ itself, the error can be corrected overwhelming probability when $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ with sufficiently large n. For both cases, we have $p = \xi \pm \epsilon < 1/2$ with $\xi \in (0, 1/2)$ and $\epsilon \in (0, 1/2 - \xi)$, hence complete the prove. □

The result of Proposition 3 indeed can be demonstrated by any $[n,k,t]$ BCH code $\mathcal{C}_\xi$ with syndrome decoding function f. Clearly, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ can correct the errors rate of $\xi + \epsilon$ with overwhelming probability after $\mathsf{poly}(\epsilon)$ iterations (see Proposition 2 and Example 2).

Apart from this, computationally efficient code achieve Shannon bound is also found by Forney in 1965, named as *concatenated code* [For65]. This outcome suggested one can choose an appropriate concatenated code to apply on $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ since the code can be linear as well.

5 Security

We now formalize the security of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. We assume an original input w is randomly sampled from a metric space $\mathcal{M}_1 = \{0,1\}^l$, over some random distribution $W \in \mathcal{M}_1$ (not mandatory uniform). Besides, we restrain another sample $w' \in W$ that show at least error rate of $\|w \oplus w'\|\, l^{-1} \geq \xi$ with the original sample w. This step is orthodox to show error tolerance up to distance t with code $\mathcal{C}_\xi$, where $\xi = t/n$. We seek to characterize the security of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ by using an adversary $\mathcal{A}$ comes with unlimited computation power. The security is formalized by using an attack running together with $\mathcal{A}$. Formally, $\mathcal{A} : \mathcal{M}_1^2 \times \mathcal{M}_2 \times \mathbb{F}_2^{(n-k)\times n} \times [l]^n \to \mathcal{M}_1$[5] is just an algorithm that is computationally unbounded, aim to recover w from a sketch $ss \in \mathcal{M}_2$, with the parity check matrix $H \in \mathbb{F}_2^{(n-k)\times n}$, an integer string $N \in [l]^n$ and $w' \in \mathcal{M}_1$ and error of distribution χ , where $\|\chi\| = l\epsilon$, and $\mathcal{M}_2 = \{0,1\}^n$. Meanwhile, we imposed an additional requirement for $\mathcal{A}$ in running the attack. To be specific, once $\mathcal{A}$ has successfully outputted the original string w, the attack is consider succeeded only if the error rate $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$. The attack is denoted as $\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, H, \epsilon, \mathcal{A})$ with input LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$, N, H, ϵ, and $\mathcal{A}$ as follow:

```
Attack($\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, H, \epsilon, \mathcal{A}$)
1 : $w \leftarrow_\$ \{0,1\}^l,\ w' \leftarrow_\$ \{0,1\}^l,$
2 : if $\|w \oplus w'\|\, l^{-1} \leq \xi$, repeat step 1 until $\|w \oplus w'\|\, l^{-1} \geq \xi$
3 : if $\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, G), w', N, H, \epsilon) = w$ & $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$
4 :     Output true
5 : else
6 :     Output false
```

The additional requirement we have imposed is meant to provide a more complete security evaluation on the input W. More explicitly, once the error e has added to the input w' during the recovery phase, it should introduce some uncertainty to the total error rate between w and w'. This error effect must be taken into account for more complete security study. For instance, given $\|w \oplus w'\| \geq \xi$, after additional error e of weight $\|e\| = \|\chi\| = l\epsilon$ is included, it may lead to either $\|w \oplus w'\| \geq \xi + \epsilon$ or $\|w \oplus w'\| \geq \xi - \epsilon$. Since the correctness result can be applied to the case when $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ by Proposition 2, focusing on both cases when $\|w \oplus w'\| \geq \xi + \epsilon$ and $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ should complete our security evaluation. We then have the following definition for our security.

Definition 3. *Let β and β' be some negligible quantity. Let W and Φ be some random variable over a metric space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $N \in [l]^n$, $H \in \mathbb{F}_2^{(n-k)\times n}$, and $\epsilon \in (0, 1/2-\xi)$, where $\xi \in (0, 1/2)$, the algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is a $(\mathcal{M}_2, m, \min\{-\log(\beta), -\log(\beta')\}, t)$ secure sketch if $\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \mathcal{A}) = \mathbf{true}\right] \leq \beta'$ and $\Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N,\ H, \epsilon), w', N, H) = w\right] \leq \beta$ for any computationally unbounded adversary $\mathcal{A}$.*

Finally, we provide a general characterization of the information theoretical security of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$, and show that it is a $(\mathcal{M}_2, m, \min\{-\log(\beta), -\log(\beta')\}, t)$ secure sketch. This proposition comes with a proof according to Definition 3

Theorem 3. *The algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is a $(\mathcal{M}_2, m, \min\{-\log(\beta), -\log(\beta')\}, t)$ secure sketch with $\beta' = 2^{-m}/\beta$ and $\beta = \exp(-2n\epsilon^2)$ if n is sufficiently large.*

Proof. (*sketch*): We here provide a brief overview of the main proof. More complete and detail proof can be found in the appendix. The **correctness** is clear, simply follow (Proposition 1 and Proposition 2). Formally,

[5] Note that we here omitted the step of recovering the padded input $v \in \{0,1\}^k$, rather, we directly refer the recovered object to be $w \in \mathcal{M}_1$. This is because once v is recovered successfully, it is trivial to look for w by the first l bits of v.

given any pair of string $w, w' \in \mathcal{M}_1$, under the case when $\|w \oplus w'\|\, l^{-1} \leq \xi \pm \epsilon$, the offset can be tolerated with overwhelming probability at least $1 - \beta = 1 - \exp(-2n\epsilon^2)$ for negligible β, or after $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ run in $\mathsf{poly}(\epsilon)$ iterations if n is sufficiently large.

For **security**, due to the introduced error effect, the error rate in the resilient vectors can simply described into two different cases, which are $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$ and $\|w \oplus w'\| l^{-1} \geq \xi - \epsilon$. Based on this, our security proof only need to focus on two different parts: (1) when $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$, and (2) when $\|w \oplus w'\|\, l^{-1} \geq \xi - \epsilon$.

Proof for Part (1): Given any pair $w, w' \in W$ with $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$, it follows that:

$$\begin{aligned}&\Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, G, \epsilon), w', N, G) = w\right] = \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| l^{-1} \geq \xi + \epsilon\right] \\ &\leq \max_{t = t_{\max}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| l^{-1} \geq \xi + \epsilon\right] = \exp\left(-2n\epsilon^2\right)\end{aligned}$$

Thus, we found $\beta = \exp(-2n\epsilon^2)$ and claim our security for this part.

However, since the error added is random during the sketching, the condition $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$ must not hold every time. We then proceed to the proof for the remaining Part (2).

Proof for Part (2): When $\|w \oplus w'\|\, l^{-1} \geq \xi - \epsilon$, since the correctness result can be applied to the case when $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ by Proposition 2, focusing on both cases when $\|w \oplus w'\|\, l^{-1} \geq \xi + \epsilon$ and $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ should complete our security evaluation. Therefore, the proof for this part follows the terminology in Attack. This attack will output true if the adversary $\mathcal{A}$ succeeded in recover w and able to show the sampled pair (w, w') comes with $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$. It should be described as follow:

$$\begin{aligned}&\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \mathcal{A}) = \mathbf{true}\right] = \Pr\left[\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon \;\middle|\; \|\delta\| \leq t\right] \\ &= \frac{\Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon\right] \Pr\left[\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon\right]}{\Pr\left[\|\delta\| \leq t\right]} \\ &= \frac{(1-\beta)\alpha}{(1-\beta)\alpha + (1-\alpha)\beta} = \frac{\alpha}{\beta}\left(\frac{1-\beta}{1-\alpha}\right) \leq \frac{\alpha}{\beta} = \beta'\end{aligned}$$

The second line result obtained by using *Bayes' law*. For the third line result, it follows: given $t_{(+)} = \lfloor(\xi + \epsilon)l\rfloor$, and let $\alpha = \Pr\left[\|w \oplus w'\| \leq t_{(+)}\right] \leq \max_{w'} \Pr\left[W \in B_{t_{(+)}}(w')\right]$. Then, by combining the results from Corollary 1 and Proposition 2 (Eq. 5), one has $\Pr[\|\delta \leq t\|] = (1-\beta)\alpha + \beta(1-\alpha)$. Since the source must come with certain quantity of fuzzy min-entropy over $t_{(+)}$, and the fuzzy min-entropy must be upper bound to the min-entropy, thus, $-\log(\alpha) = \mathrm{H}^{\mathrm{fuzz}}_{t_{(+)},\infty}(W) \geq \mathrm{H}_\infty(W) = m$, and $\alpha \leq 2^{-m}$ by minimum entropy of m. Therefore β can be any value $\geq \alpha$ to show security.

In the end, the maximum probability of recovering w for both Part (1) and Part (2) described as $\max\{2^{-m}/\beta, \beta\}$. □

Remark: The events when $\|w \oplus w'\|\, l^{-1} \geq \xi + \epsilon$ and $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ can also be represented as the cases when $\|w \oplus w'\| > t_{(+)}$ and $\|w \oplus w'\| \leq t_{(+)}$ respectively. In our exposition, we usually refer to the former representation to show more meaningful details with ξ and ϵ. This show $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ offers correctness when $\|w \oplus w'\| \leq t_{(+)}$ but no guarantee for the case when $\|w \oplus w'\| > t_{(+)}$ further supported the definition for a standard secure sketch.

The proof of Theorem 3 demonstrated one can construct an info. theoretic secure sketch with security does not depend upon the length itself provided the min-entropy of the source in $\mathcal{M}_1$ is high enough. Therefore, to show meaningful security, the sources must at least come with sufficient amount of min-entropy, where the entropy loss can be characterized by average fuzzy min-entropy, which parametrized by a chosen $[n, k, t]$ code $\mathcal{C}_\xi$.

We here provide discussion on how the average fuzzy min-entropy ($-\log\beta$) help in monitoring the entropy loss. Recall the fuzzy min-entropy definition. The residual entropy of Φ is equivalent to its fuzzy min-entropy over tolerance distance t, which can be bounded as $\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(\Phi) \geq \tilde{\mathrm{H}}_\infty(\Phi | W, ss) = \mathrm{H}_\infty(\Phi) - \lambda$, where λ refers to the entropy loss. Since we have used the min-entropy as lower bound for $-\log \alpha$, the computed fuzzy min-entropy must be minimum as well, which is:

$$\begin{aligned} \mathrm{H}^{\text{fuzz}}_{t,\infty}(\varPhi) &= \mathrm{H}_\infty(\varPhi) - \lambda = \min\{-\log(\beta'), -\log(\beta)\} \\ &= \min\{-\log(\alpha/\beta), -\log(\beta)\} \\ &= \min\{\mathrm{H}_\infty(W) + \log(\beta), -\log(\beta)\} \end{aligned} \tag{6}$$

Given the case when $m/2 \leq -\log(\beta)$, through direct comparison, we have $\mathrm{H}_\infty(\varPhi) - \lambda = \mathrm{H}_\infty(W) + \log(\beta)$. Since $\mathrm{H}_\infty(W) \geq m$, it follows $\mathrm{H}_\infty(\varPhi) \geq m$, so the entropy loss would be bounded as $\lambda \leq -\log(\beta)$. On the other hand, if $m/2 > -\log(\beta)$, one has $\mathrm{H}_\infty(\varPhi) - \lambda = -\log(\beta)$, hence larger entropy loss can be seen described as $\lambda \leq m - (-\log(\beta))$.

An alternative to always ensure meaningful security can be provided is to have a precise knowledge setting on the input distribution during the sketching phase. This setting can be achieved by using the universal hashing to disambiguate the points as proposed by Fuller *et al.,* [FRS16].

Given the source in some distribution W over $\mathcal{M}_1$, which has no min-entropy to support meaningful security, showing security on it seems to be an extra move. Nevertheless, there have a plethora of sources with "reasonable" amount of min-entropy. Our construction offers the advantage to show security with min-entropy measurement that can be view as the min-entropy over larger metric space by the resilient vector. In light of this, one can always show security to all family of distribution with our construction, but not always all of them are meaningful ones.

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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$.

If a secure sketch allows recovery of the input from some errors with high probability, it must consist enough information to describe the error pattern. According to Dodis *et al.* [DRS04], 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 described the security bound in this model by assuming W is drawn from uniform. 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.

Under the same model, the bound with $nh_2(p)$ bits entropy loss is possible to be applied in our case as well, by letting $p = \|w \oplus w'\| \, l^{-1} + \epsilon$. However, through comparing the mathematical description of the average fuzzy min-entropy $-\log(\beta)$ and $nh_2(p)$, it shows that there is no compiling need to consider the error rate of the input $\|w \oplus w'\| \, l^{-1}$ to outline the entropy loss. Clearly, $-\log(\beta) = -\log(\exp(-2n\epsilon^2))$ will show lower value with smaller ϵ without the knowledge of the input error rate $\|w \oplus w'\| \, l^{-1}$. Recall this entropy loss can simply described by $-\log(\beta)$ under the case when $m/2 \geq -\log(\beta)$ (see Eq. 6). This result suggested a better achievable lower bound to describe the error pattern in the resilient vectors of size n by using $-\log(\beta)$ rather than $nh_2(p)$. Additionally, it is well-understood that W is not uniform in our case, therefore, the lower bound residual entropy described by $n(1 - h_2(p) - o(1))$ may not directly applicable to us. In fact, we have shown that, the upper bound residual entropy in our construction is $\min\{\mathrm{H}_\infty(W) + \log(\beta), -\log(\beta)\}$. Apparently, this residual entropy is always bounded by the min-entropy of the source instead of the blocklength of the code n.

Perceivably, min-entropy has shown to offer more meaningful results relatively to Shannon entropy, especially for the case when the inputs are not uniform. These results have motivated the usage of min-entropy instead of Shannon entropy to avoid overestimation on the residual entropy, which is critical while designing any cryptographic application such as randomness extractor or key derivation. In spite of that, for any discussion related to resilience, the Shannon bound is always a good reference point to exam the existence of such a code for error correction.

We have the following proposition to describe the best possible security for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$, whose proof is straightforward.

Proposition 4. *The best information theoretical security with algorithm pair* $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ *is* $m/2$

Proof. Since we have a $(\mathcal{M}_2, m, \min\{-\log(\beta), -\log(\beta')\}, t)$ secure sketch with $\beta' = 2^{-m}/\beta$ and $\beta = \exp(-2n\epsilon^2)$ (Theorem 3). Therefore, the best possible security balances both sites which is:

$$\begin{aligned} m + \log(\beta) &= -\log(\beta) \\ m/2 &= -\log(\beta) \end{aligned}$$

□

Given a source with min-entropy m, one can always choose a desired security level by average fuzzy min-entropy via computing $-\log\beta$. This security holds for computationally unbounded adversary $\mathcal{A}$ with conditioned on $m/2 \geq -\log\beta$ by Proposition 4.

Table 1 tabulated the security bound for various β-correct secure sketch.

Security Bound for β-Correct Secure Sketch		
Computational	Best possible security	$\mathrm{H}^{\text{fuzz}}_{t,\infty}(W) - \log(1-\beta)$
Computational	FRS sketch(universal hash functions) [FRS16]	$\mathrm{H}^{\text{fuzz}}_{t,\infty}(W) - \log(\frac{1}{\beta}) - \log\log(\mathsf{supp(W)}) - 1$
Computational	Layer hiding hash (strong universal hash function)[WCD+17]	$\mathrm{H}^{\text{fuzz}}_{t,\infty}(W) - \log(\frac{1}{\beta}) - 1$
Info. theoretic	**LSH sketch**	$\min\{\mathrm{H}_\infty(W) + \log(\beta), -\log(\beta)\}$[6]

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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ in this section. First stated by Boyen, 2004 [Boy04], 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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ can show reusability property, it also suggested a reusable fuzzy extractor for uniform random strings generation.

In the context of showing reusability, $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$ 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) [FRS16] 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

[6] We used t' instead of t to remark the LSH sketch emphasis on different tolerance distances explicitly

$f_1, \ldots, f_\gamma$ 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 $\mathcal{M}$.

We now formalize the reusability of algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. Basically, it follows the previous security setting, but only comes with slight extension (from single adversary to multi adversaries setting). We assume an original input w^* is randomly sampled from a metric space $\mathcal{M}_1 = \{0,1\}^l$, over some random distribution $W \in \mathcal{M}_1$ (not mandatory uniform). Again, we restrain another sample $w' \in W$ that show at least error rate of $\|w^* \oplus w'\| \, l^{-1} \geq \xi$. We aim to characterize the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ by using a group of adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$ comes with unlimited computation power. To do so, we have introduced additional random error $\{e'_1, \ldots, e'_\gamma\}$ with $\|e'_i\| \leq l\epsilon' < l\epsilon$ parametrized by another error parameter $\epsilon' \in (0, 1/2 - \xi - \epsilon)$. Formally, e'_i 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^*.

Briefly, we seek to show reusability defined in information-theoretical sense as well. Our work is considered as 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.

The security is formalized by using an attack running together with $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$. Formally, each adversary $\mathcal{A}_i : \mathcal{M}_1^2 \times \mathcal{M}_2 \times \mathbb{F}_2^{(n-k)\times n} \times [l]^n \to \mathcal{M}_1$ is simply an algorithm that is computationally unbounded to output w_i from a public sketch $ss_i \in \mathcal{M}_2$, with input $w' \in \mathcal{M}_1$, a parity check matrix $H \times \mathbb{F}_2^{(n-k)\times n}$, an error parameter $\epsilon \in (0, 1/2 - \xi)$ and an integer string $N \in [l]^n$. Follow previous security setting, similar requirement is imposed on $\mathcal{A}_i$ in running the attack. That is, once $\mathcal{A}_i$ has successfully outputted the string w_i, the attack is only considered succeeded if the error rate $\|w_i \oplus w'\| \leq \xi + \epsilon + \epsilon'$. The attack is denoted as $\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\})$ with input LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$, N, H, and $\mathcal{A}_i$ as follow:

```
$\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\})$
1 :  $w^* \leftarrow_\$ \{0,1\}^l, \quad w' \leftarrow_\$ \{0,1\}^l$
2 :  **if** $\|w^* \oplus w'\| \, l^{-1} \leq \xi$, **repeat step 1 until** $\|w^* \oplus w'\| \, l^{-1} \geq \xi$
3 :  **for** $i = 1 : \gamma$
4 :  $e'_i \leftarrow_\$ \{0,1\}^l$ // the weight $\|e'_i\| = l\epsilon'_i \leq l\epsilon' < l\epsilon$
5 :  $w_i = w^* \oplus e'_i$
6 :      **if** $\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w_i, N,\ H), w', N, H, \epsilon) = w_i$ & $\|w_i \oplus w'\| \, l^{-1} \leq \xi + \epsilon + \epsilon'$
7 :      Output **true**
8 :      **else**
9 :      Output **false**
10 : **endfor**
```

Our intuition of showing reusability for a group of adversary follows the works proposed by Fuller *et at.,* [FRS16]. The goal is to show security to the original sample w^* for different independent re-enrollment, with some perturbations. Reusability can only be claimed if the security holds for all adversaries corresponds to individual re-enrollment of w^* respectively. Since each re-enrollment is subjected to different providers, and the providers may not communicating and trusted to each other, therefore showing security individually to each adversary $\mathcal{A}_i$ is necessary to support our claim. We give the definition below to characterized the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$.

Definition 4. *Let β_2 and β_2' be some negligible quantities. Let W and Φ be some random variable over a metric space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $N \in [l]^n$, $\epsilon \in (0, 1/2 - \xi)$ and $\epsilon' \in (0, 1/2 - \xi - \epsilon)$, where $\xi \in (0, 1/2)$ and $H \in \mathbb{F}_2^{(n-k)\times n}$. Let $\epsilon^*_{(-)} = \epsilon - \epsilon'$, the algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is $(\max\{\beta_2, \beta_2'\}, \epsilon^*_{(-)}, \gamma)$-reusable if one has $\max_\gamma \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, H, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}) = \textbf{true}\right] \leq \beta_2'$ and $\max_i \Pr\left[\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w_i, N,\ H), w', N, H, \epsilon) = w_i\right] \leq \beta_2$ for a group of computational unbounded adversary $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$.*

Recall we have initially introduced an error e of weight $\|e\| = l\epsilon$ during sketching. Given another error $\|e_i'\| = l\epsilon_i \leq l\epsilon'$, where $\epsilon' < \epsilon$, it means $\|e_i' \oplus e\|\, l^{-1}$ must within $\epsilon \pm \epsilon'$. Suppose the error rate between w^* and w' satisfies $\|w^* \oplus w'\| \geq \xi$, the total error effect (for sketching and perturbation) will cause the changes of the final error rate to either $\|w^* \oplus w'\|\, l^{-1} \geq \xi + (\epsilon \pm \epsilon')$ or $\|w^* \oplus w'\|\, l^{-1} \geq \xi - (\epsilon \pm \epsilon')$. Manifestly, further simplification can be done by letting $\epsilon^* = \epsilon \pm \epsilon'$, which therefore allows one to describe the final error rate as $\|w^* \oplus w'\|\, l^{-1} \geq \xi + \epsilon^*$ and $\|w^* \oplus w'\|\, l^{-1} \geq \xi - \epsilon^*$ respectively. Consequently, doing so can easily lead us to the security reduction from multi-adversaries setting to single adversary setting which has been covered by the prove of Theorem 3.

Based on the reasoning above, adding error while sketching implicitly allows reusability. Hence, the proof of reusability is trivial in our case. Nevertheless, it is worth to detail the reduction of the security property from multiple adversaries setting to single adversary setting over $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$.

The following lemma is given to characterize the reusability of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$. The proof demonstrated the security reduction of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ from multi-adversaries setting to single adversary setting.

Lemma 1. *The algorithm pair $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}\rangle$ is $(\max\{\beta_2, \beta_2'\}, \epsilon^*_{(-)}, \infty)$-reusable, with $\beta_2' = 2^{-m}/\beta_2$ and $\beta_2 = \exp(-2n(\epsilon^*_{(-)})^2)$ given some integer $m > 0$.*

Proof. (Sketch) We reiterate the proof of this lemma is similar to the proof of the security in Theorem 3. We briefly highlighted the main prove as follows.

There consisted only two errors we have introduced, e of weight $\|e\| = l\epsilon$ during sketching, and e_i' of weight $\|e_i\| = l\epsilon_i' \leq l\epsilon' < l\epsilon$ to show reusability. Given the error rate for w^* and w' satisfies $\|w^* \oplus w'\| \geq \xi$, the error (for reusability) will lead to either $\|w_i \oplus w'\|\, l^{-1} \geq \xi + \epsilon_i'$ or $\|w_i \oplus w'\|\, l^{-1} \geq \xi - \epsilon_i'$. Together with the error added during our sketching phase, the final error rate can be described as $\|w_i \oplus w'\|\, l^{-1} \geq \xi + \epsilon \pm \epsilon_i'$ or $\|w_i \oplus w'\|\, l^{-1} \geq \xi - \epsilon \pm \epsilon_i'$. Likewise the single adversary setting, the second inequality can be replaced by $\|w_i \oplus w'\|\, l^{-1} \leq \xi + \epsilon \pm \epsilon_i'$ because of the correctness should hold when $\|w_i \oplus w'\|\, l^{-1} \leq \xi + \epsilon \pm \epsilon_i'$ as well.

Therefore, the error rate in the resilient vectors can simply analysed under these two parts, which are Part(1): when $\|w_i \oplus w'\|\, l^{-1} \geq \xi + \epsilon \pm \epsilon_i'$, and Part(2): when $\|w_i \oplus w'\|\, l^{-1} \leq \xi + \epsilon \pm \epsilon_i'$.

We will let $\epsilon^*_{(-)} = \epsilon - \epsilon'$ and $\epsilon^*_{(+)} = \epsilon + \epsilon'$ throughout the whole prove to show the reduction of our result from multiple adversaries to single adversary setting. Since we have multiple adversaries needed to consider, our aim of Part (1) proof is to find the maximum probability to correct the offset among all of them. Let $\beta_{2,i} = \exp(-2n(\epsilon \pm \epsilon_i')^2)$, the maximum probability described as follow:

$$\begin{aligned}
&\max_i \Pr\left[\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w_i, N,\ H), w', N, H, \epsilon) = w_i\right] \\
&= \max_{w_i} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w_i \oplus w'\| l^{-1} \geq \xi + \epsilon \pm \epsilon_i'\right] = \beta_{2,i} \leq \exp\left(-2n(\epsilon^*_{(-)})^2\right)
\end{aligned}$$

The last line result follows by taking the maximum value for $\beta_{2,i}$, clearly, the maximum value of $\beta_{2,i} = \exp(-2n(\epsilon \pm \epsilon_i')^2)$ refer to the case when $\epsilon \pm \epsilon_i'$ is minimum, which is $\epsilon - \epsilon'$, since $\epsilon_i' \leq \epsilon'$. Let $\beta_2 = \exp\left(-2n(\epsilon - \epsilon')^2\right) = \exp\left(-2n(\epsilon^*_{(-)})^2\right)$, the security for Part (1) is claimed.

On the other hand, the main prove for Part (2) is to show security hold for all adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$ in running Attack_2. Formally, it can be described as:

$$\begin{aligned}
&\max_\gamma \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}) = \textbf{true}\right] \\
&= \max_{w_i} \Pr\left[\|w_i \oplus w'\|\, l^{-1} \leq \xi + \epsilon \pm \epsilon_i' \;\middle|\; \|\delta\| \leq t\right]
\end{aligned}$$

We first look for $\Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi + \epsilon \pm \epsilon_i' \;\middle|\; \|\delta\| \leq t\,\right]$, then follow by its maximum value. By the results of Corollary 1 and Proposition 2 and Bayes' law (follows the proof of Theorem 3):

$$\Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi + \epsilon \pm \epsilon_i' \;\middle|\; \|\delta\| \leq t\,\right] = \frac{1-\beta_{2,i}}{\beta_{2,i}} \left(\frac{\alpha_{2,i}}{1-\alpha_{2,i}}\right)$$

Above result depends upon the error parameter ϵ_i for each $\mathcal{A}_i$. In particular, we let $t'_{2,i} = \lfloor(\xi + \epsilon \pm \epsilon'_i)l\rfloor$, and so $\alpha_{2,i} = \Pr\left[\|w_i \oplus w'\| \leq t'_{2,i}\right] \leq \max_{w'} \Pr\left[W \in B_{t'_2}\right]$, where $t'_2 = \lfloor(\xi + \epsilon + \epsilon'_i)l\rfloor$ (since $(\xi + \epsilon^*_{(+)}) \geq (\xi + \epsilon^*_{(-)})$ by maximum). Therefore we have $\alpha_2 \leq \max_{w'} \Pr\left[W \in B_{t'_2}\right]$ and one can bound $-\log \alpha_{2,i} \geq -\log \alpha_2 \geq \mathrm{H}^{\mathrm{fuzz}}_{t'_2,\infty}(W) \geq \mathrm{H}_\infty(W) = m$. It follows:

$$\begin{aligned} &\max_{\gamma} \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}) = \mathbf{true}\right] \\ &= \max_{w_i} \Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi + \epsilon \pm \epsilon_i' \;\middle|\; \|\delta\| \leq t\,\right] \leq \frac{\alpha_2}{\beta_2} = \frac{2^{-m}}{\beta_2} \end{aligned}$$

by maximum $\beta_{2,i} \leq \beta_2$ (prove in Part (1)) and maximum $\alpha_{2,i} \leq 2^{-m}$.

Consequently, one has the new security results in a multiple adversaries setting (a group of adversary). The maximum probability to decode the codeword successfully is $\max\{\beta'_2, \beta_2\}$ with $\beta'_2 = 2^{-m}/\beta_2$ and $\beta_2 = \exp(-2n(\epsilon^*_{(-)})^2)$. This result holds for all the adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_\gamma\}$. The prove holds with $\gamma = \infty$. □

With the proof of Lemma 1, we concluded that $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ allows the re-enrollment of the input w^* for $\gamma = \infty$ number of times as long as the error (perturbation) e' has bounded weight $\|e'_i\| \leq l\epsilon'$ for $i = \{1, \ldots, \gamma\}$. The security holds for all adversaries is $\min\{-\log(\beta'_2), -\log(\beta_2)\}$. Noticeably, the security over multi-adversaries setting is similar to single adversary setting, with the only changed error parameter from ϵ (single adversary) to $\epsilon^*_{(-)}$ (multi-adversaries). We therefore obtain the following proposition

Proposition 5. *If a pair of LSH-sketching and recover algorithm* $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ *is* $(\max\{\beta_2, \beta'_2\}, \epsilon^*_{(-)}, \infty)$*-reusable, it is also a* $(\mathcal{M}_2, m, \min\{-\log(\beta_2), -\log(\beta'_2)\}, t)$ *secure sketch for sufficiently large* n.

We omitted the proof of Proposition 5 since it is straightforward. Precisely, its **correctness** claims by Proposition 1 (after reduction from multiple adversaries setting to single adversary setting) for sufficient large n and its **security** claims by Lemma 1 itself.

7 A Toy Example

In this section, a toy example is given to demonstrate how $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ can be practically applies to real cases. This example focuses on one of the common-known noisy sources which consist of more error than entropy-*IrisCode*. The IrisCode is a binary representation extracted from the human iris, and it is being used to perform biometric authentication. It has been viewed as the strongest biometric [PPJ03] due to its uniqueness and resistant against false matching. We adopted the IrisCode of vector $w \in \{0,1\}^l$ with $l = 2048$, which is first considered by Daugman in 2006. [Dau06]. Based on the *degree of freedom* argument, this IrisCode is believed to come with entropy around 249 bits. Additionally, it is commonly conceived that depends on different transformation, from the original eye images to IrisCode generation, the error content in different IrisCode of the same user lye in between $10\% - 35\%$ [FSS17]. Therefore, a traditional secure sketch should loss all information and show no security for any error correction on error rate $\geq 249/2048 = 0.1215$.

We here show how $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is able to correct an error rate $249/2048 = 0.1215$ yet offer security guarantee. To do so, an $[65535, 2061, 10967]_2$ BCH code $\mathcal{C}_{10967/65535}$ is chosen, with $\xi = 0.1674$. To correct an error rate of $0.15 > 0.1215$, one can compute the value ϵ required, which is $307 \leq \lfloor(0.1674 - \epsilon)2048\rfloor$, thus $\epsilon \leq 0.0174$ and $\|e\| \leq \lfloor(0.0174)2048\rfloor = 35$ bits. In such a case, one could enjoy information theoretical security of $-\log(\exp(-2n\epsilon^2)) = 58$ bits, where the error can be corrected with overwhelming probability at least $1 - 2^{-58}$.

To correct more error, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ also offered the list-decoding strategy which run in polynomial time (see Section 4.1). Suppose now the error rate increased to 0.1750. Then, $\mathcal{C}_{10967/65535}$ has shown insufficient error correction capacity for such error rate, hence cannot correct the errors with overwhelming probability as promised. In such event, adding the random error e would help to achieve $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ and so $\|w \oplus w'\|\, l^{-1} - \epsilon \leq \xi$ can hold. Consequently, one can easily compute that $0.1750 \leq (0.1674 + \epsilon)$, and so $\epsilon \geq 0.0076$. Thus $\|e\| \geq \lceil l(0.1750 - 0.1674) \rceil = 16$ bits are required. Precisely, by Eq. 5, after $-\log \frac{\binom{2048}{16}}{\binom{356}{16}} = 2^{41}$ number of iterations, the error can be corrected with overwhelming probability $1 - \exp\left(-2n\epsilon^2\right) = 1 - 1.94 \times 10^{-3}$, with information theoretical security $-\log(\exp(-2n\epsilon^2)) = 11$ bits. To show re-usability in this case, we refer perturbation e' with $\|e'\| \leq l\epsilon'$, which can be described using perturbation parameter $\epsilon' < \epsilon$. For instance, given any perturbation e' that maximally caused 3 bits flipped on the input, i.e. $\|e'\| = 3$, so $\epsilon' = 3/2048 < \epsilon$. by Lemma 1 results, it follows $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is $(\max\{\beta_2, \beta'_2\}, \epsilon^*_{(-)}, \infty)$-reusable, where one can compute that $\beta_2 = 2^{-m}/\beta_2$, where $\beta'_2 = \exp(-2n(\epsilon^*_{(-)})^2)) = 0.0076$ and $\epsilon^*_{(-)} = \epsilon - \epsilon' = 0.0061$, with information theoretical security reduced to $-\log(\beta'_2) = 8$ bits.

7.1 The Cost of Information Theoretical Security

As we shall see, based on our toy example given in Section 7, one can show considerable high information theoretical security (58 bits) for any inputs' error rate $\|w \oplus w'_G\|\, l^{-1} + \epsilon \leq \xi$ but lower security (16 bits) for any inputs' error rate $\|w \oplus w'_G\|\, l^{-1} - \epsilon \leq \xi$ with code $\mathcal{C}_\xi$. Apparently, correcting more error introduced lower security level (information theoretically).

Noting that for error correction over error rates $\|w \oplus w'_G\|\, l^{-1} - \epsilon \leq \xi$, it only holds if $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ has run in time after $\mathsf{poly}(\epsilon)$ iterations. This result suggested that the computation complexity required for error correction indeed goes higher. To visualize this, suppose we have two different cases: a *genuine case* with recovery input w'_G and an *imposter case* with recovery input w'_I. Obviously, for $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ to be useful and applicable, its correctness must hold when $\|w \oplus w'_G\|\, l^{-1} \leq \xi + \epsilon$ under the genuine case. Saying so, as long as $\|w \oplus w'_G\|\, l^{-1}$, any introduced error parameter $\epsilon \in (0, 1/2)$ is possible to yield $\|w \oplus w'_G\|\, l^{-1} \leq \xi + \epsilon$, thus, the inputs' errors rate $\|w \oplus w'_G\|\, l^{-1}$ can be corrected with overwhelming probability by the correctness of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ itself.

Specifically, for some value of $\|w \oplus w'_G\|\, l^{-1} \in (0, 1/2)$, where $\|w \oplus w'_G\|\, l^{-1} = \xi + \epsilon_G$, it means one must add additional error $\|e\| = l\epsilon$ parametrized by $\epsilon = \epsilon_G$ to yield $\|w \oplus w'_G\|\, l^{-1} - \epsilon_G \leq \xi$ and show correctness. Clearly, same things applied to the imposter case as well, where any adversaries could achieve correctness by adding extra error $\|e\| = l\epsilon_I$ to yield $\|w \oplus w'_I\|\, l^{-1} \leq \xi + \epsilon_I$. In such events, one can easily show that the number of iteration required for genuine case and imposter case will be different, which can be calculated by Eq. 5, i.e., $-\log\left(\binom{l}{l\epsilon_G}/\binom{\|w \oplus w'_G\|}{l\epsilon_G}\right)$ and $-\log\left(\binom{l}{l\epsilon_I}/\binom{\|w \oplus w'_I\|}{l\epsilon_I}\right)$ respectively. From this point of view, we can further reason that without the genuine case information of $\|w \oplus w'_G\|\, l^{-1}$ and ϵ_G, any computationally bounded attacker can still succeed with overwhelming probability after $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ has run in $-\log\left(\binom{l}{l\epsilon_I}/\binom{\|w \oplus w'_I\|}{l\epsilon_I}\right)$ number of iterations. Therefore, we can conclude that introducing error e of higher weight (e.g., higher value of ϵ) will decrease our security for computational unbounded adversary (information theoretically) but lead to higher computational complexity for computational bounded adversary. In other words, correcting more errors than any code $\mathcal{C}_\xi$ could be done (i.e., $\|w \oplus w'\|\, l^{-1} \geq \xi$) implies higher computational security but lower information theoretical security. Conversely, to show security information theoretically, one is subjected to strong theoretical bound for any error correction code $\mathcal{C}_\xi$ to maximally correct error rate of 1/4 (see Section 4.2). In such a case, it is natural to ask whether the cost for information theoretical security is worth and necessary, therefore, the argument over the needs of information theoretical security is indeed an interesting open question.

References

BA11. Marina Blanton and Mehrdad Aliasgari. 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*, pages 68–77. IEEE, 2011.

BA13. Marina Blanton and Mehrdad Aliasgari. Analysis of reusability of secure sketches and fuzzy extractors. *IEEE transactions on information forensics and security*, 8(9):1433–1445, 2013.

BBR88. Charles H Bennett, Gilles Brassard, and Jean-Marc Robert. Privacy amplification by public discussion. *SIAM journal on Computing*, 17(2):210–229, 1988.

Ber15. Elwyn R Berlekamp. *Algebraic coding theory*. World Scientific Publishing Co, 2015.

BMVT78. Elwyn Berlekamp, Robert McEliece, and Henk Van Tilborg. On the inherent intractability of certain coding problems (corresp.). *IEEE Transactions on Information Theory*, 24(3):384–386, 1978.

Boy04. Xavier Boyen. Reusable cryptographic fuzzy extractors. In *Proceedings of the 11th ACM conference on Computer and communications security*, pages 82–91. ACM, 2004.

CFP$^+$16. Ran Canetti, Benjamin Fuller, Omer Paneth, Leonid Reyzin, and Adam Smith. Reusable fuzzy extractors for low-entropy distributions. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 117–146. Springer, 2016.

Cha02. Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

CS08. F Carter and A Stoianov. Implications of biometric encryption on wide spread use of biometrics. In *EBF Biometric Encryption Seminar (June, 2008)*, volume 29, 2008.

Dau06. John Daugman. Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, 94(11):1927–1935, 2006.

DRS04. Yevgeniy Dodis, Leonid Reyzin, and Adam 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*, pages 523–540. Springer, 2004.

DW09. Yevgeniy Dodis and Daniel Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 601–610. ACM, 2009.

EHMS00. Carl Ellison, Chris Hall, Randy Milbert, and Bruce Schneier. Protecting secret keys with personal entropy. *Future Generation Computer Systems*, 16(4):311–318, 2000.

FJ01. Niklas Frykholm and Ari Juels. Error-tolerant password recovery. In *Proceedings of the 8th ACM conference on Computer and Communications Security*, pages 1–9. ACM, 2001.

FMR13. Benjamin Fuller, Xianrui Meng, and Leonid Reyzin. Computational fuzzy extractors. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 174–193. Springer, 2013.

For65. G David Forney. Concatenated codes. *Phd Thesis*, 1965.

FRS16. Benjamin Fuller, Leonid Reyzin, and Adam 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*, pages 277–306. Springer, 2016.

FSS17. Benjamin Fuller, Sailesh Simhadri, and James Steel. Reusable authentication from the iris. Cryptology ePrint Archive, Report 2017/1177, 2017. `https://eprint.iacr.org/2017/1177`.

GBGRB18. Marta Gomez-Barrero, Javier Galbally, Christian Rathgeb, and Christoph Busch. General framework to evaluate unlinkability in biometric template protection systems. *IEEE Transactions on Information Forensics and Security*, 13(6):1406–1420, 2018.

GIM$^+$99. Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.

Gur04. Venkatesan Guruswami. *List decoding of error-correcting codes: winning thesis of the 2002 ACM doctoral dissertation competition*, volume 3282. Springer Science & Business Media, 2004.

Gur10. Venkatesan Guruswami. Introduction to coding theory, lecture 2: Gilbert-varshamov bound. *University Lecture*, 2010.

JNR16. Anil K Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 79:80–105, 2016.

JS06. Ari Juels and Madhu Sudan. A fuzzy vault scheme. *Designs, Codes and Cryptography*, 38(2):237–257, 2006.

JW99. Ari Juels and Martin Wattenberg. A fuzzy commitment scheme. In *Proceedings of the 6th ACM conference on Computer and communications security*, pages 28–36. ACM, 1999.

KBK+11. Emile JC Kelkboom, Jeroen Breebaart, Tom AM Kevenaar, Ileana Buhan, and Raymond NJ Veldhuis. Preventing the decodability attack based cross-matching in a fuzzy commitment scheme. *IEEE Transactions on Information Forensics and Security*, 6(1):107–121, 2011.

MS77. Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes.* Elsevier, 1977.

Por82. Sigmund N Porter. A password extension for improved human factors. *Computers & Security*, 1(1):54–56, 1982.

PPJ03. Salil Prabhakar, Sharath Pankanti, and Anil K Jain. Biometric recognition: Security and privacy concerns. *IEEE security & privacy*, (2):33–42, 2003.

PW72. William Wesley Peterson and Edward J Weldon. *Error-correcting codes.* MIT press, 1972.

Sha01. Claude E Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

STP09. Koen Simoens, Pim Tuyls, and Bart Preneel. Privacy weaknesses in biometric sketches. In *Security and Privacy, 2009 30th IEEE Symposium on*, pages 188–203. IEEE, 2009.

Sud01. Madhu Sudan. Lecture notes for an algorithmic introduction to coding theory. *Course taught at MIT*, 2001.

WCD+17. Joanne Woodage, Rahul Chatterjee, Yevgeniy Dodis, Ari Juels, and Thomas Ristenpart. A new distribution-sensitive secure sketch and popularity-proportional hashing. In *Annual International Cryptology Conference*, pages 682–710. Springer, 2017.

8 Appendix

Proof of Theorem 3:

Proof. **Correctness**: The correctness property follows the completeness and soundness of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ itself (proven in Proposition 1 and Proposition 2 respectively). Particularly, for any input string $w' \in W$ that is at most $t_{(-)} = \lfloor (\xi - \epsilon) l \rfloor$ close to its original value $w \in W$, formally, it means $\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$, then, the probability for $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(w', N, H, \epsilon) = w$ is overwhelming at least $1 - \beta \geq 1 - \exp(-2n\epsilon^2)$. On the other hand, if w' is at most $t_{(+)} = \lfloor (\xi + \epsilon) l \rfloor$ close to its original value $w \in W$, formally, it means $\|w \oplus w'\| \, l^{-1} \leq \xi + \epsilon$, after $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ run in $\mathsf{poly}(\epsilon)$ iterations, the probability for $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(w', N, H, \epsilon) = w$ is overwhelming at least $1 - \beta \geq 1 - \exp(-2n\epsilon^2)$. Both cases hold for sufficiently large value of n

Security: We now argue in the security of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. Observe that, given a sketch $ss = c \oplus \phi$, no doubt that, the best strategy to recover w is through decoding the nearest codeword. In fact, this corresponds to the well-known problem of decoding a random linear code that is considered to be NP-hard [BMVT78]. However, this statement is not sufficient for our security goal, which is to show security for computational unbounded adversary (information theoretically secure). For the seek of completeness, the proof of security can be divided into two parts:

Proof for Part (1), when $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$: Recall after error e of weight $\|e\| = l\epsilon$ is included, initially, $\|w \oplus w'\| \, l^{-1} \geq \xi$, it may lead to either $\|w \oplus w'\| \geq \xi + \epsilon$ or $\|w \oplus w'\| \geq \xi - \epsilon$. The prove for this part is to show security on the first case.

Given any pair $w, w' \in W$ with $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$, it follows that (proven in Corollary 2):

$$\begin{aligned} &\Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, H), w', N, H, \epsilon) = w\right] = \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| l^{-1} \geq \xi + \epsilon\right] \\ &\leq \max_{t = t_{\max}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| l^{-1} \geq \xi + \epsilon\right] \leq \exp\left(-2n\epsilon^2\right) = \beta \end{aligned}$$

This result depicted the upper bound advantages for $\mathcal{A}$ to decode the codeword c' when $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$, formally holds for any variable $W \notin B_{t_{(+)}}(w')$. Thus we found $\beta = \exp(2n\epsilon^2)$ and claim our security for this part.

However, since the error added is random during sketching, the condition $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$ must not holds every times. Particularly, one may also have $\|w \oplus w'\| l^{-1} \geq \xi - \epsilon$. Merely focusing on decoding the codeword might not sufficient to claim our security in this case. Therefore, we must proceed to Part (2) to complete our proof of security.

Proof for Part (2): When $\|w \oplus w'\|\, l^{-1} \geq \xi - \epsilon$, since the correctness result can be applied to the case when $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ by Proposition 2, focusing on both cases when $\|w \oplus w'\| \geq \xi + \epsilon$ and $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$ should complete our security evaluation. Therefore, the proof for this part follows the terminology in Attack. This attack will output true if the adversary $\mathcal{A}$ succeeded in recover w and able to show the sampled pair (w, w') comes with $\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon$. It should be described as follow:

$$\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \mathcal{A}) = \mathbf{true}\right] = \Pr\left[\|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon \;\middle|\; \|\delta\| \leq t\,\right]$$

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

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

$$\mathsf{Event}_b = \begin{cases} \|w \oplus w'\|\, l^{-1} \leq \xi + \epsilon, & b = 0 \\ \|w \oplus w'\|\, l^{-1} \geq \xi + \epsilon, & b = 1 \end{cases}$$

By using *Bayes' law*:

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

Straight away, we use the results from Corollary 1 to compute the maximum probability for the events $\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\,\right]$. It follows:

$$\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\,\right] \leq \max_{t = t_{\max}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| > t_{(+)}\,\right] \leq \exp(-2n\epsilon^2) = \beta$$

To obtain $\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\,\right]$, we use the result from Proposition 2. By Eq. 4:

$$\begin{aligned}
&\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\,\right] = \min_{t = t_{\min}} \Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \leq t_{(+)}\,\right] \\
&\geq 1 - \exp\left(-2n\epsilon^2\right) = 1 - \beta
\end{aligned}$$

Recall the definitions:

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

and let $\alpha = \Pr[\mathsf{Event}_{b=0}]$, as $t_{(+)} = \lfloor(\xi + \epsilon)l\rfloor$, α can be rewritten as:

$$\begin{aligned}\alpha = \Pr[\mathsf{Event}_{b=0}] &= \Pr\left[\|w \oplus w'\| \leq t_{(+)}\right]\\ &\leq \max_{w'} \Pr\left[W \in B_{t_{(+)}}(w')\right]\end{aligned}$$

Eq. (7) can further simplify as

$$\left(\frac{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\,]}{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\,]}\right)\left(\frac{\Pr[\mathsf{Event}_{b=0}]}{1 - \Pr[\mathsf{Event}_{b=0}]}\right) = \left(\frac{1-\beta}{1-\alpha}\right)\frac{\alpha}{\beta}$$

Let $\beta' = \alpha/\beta$. We used the min-entropy as lower bound, one has $\mathrm{H}^{\text{fuzz}}_{t_{(+)},\infty}(W) \geq \mathrm{H}_\infty(W) = m$, therefore, $-\log(\alpha) \geq m$, and so $\alpha \leq 2^{-m}$. In such a case, β can be any value $\geq \alpha$ to show security, thus yield:

$$\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ G, \epsilon, \mathcal{A}) = \textbf{true}\right] \leq \beta' = 2^{-m}/\beta$$

Combining the results from Part (1) and Part (2), the maximum probability to decode the codeword is:

$$\begin{aligned}&\max\left\{\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N,\ H, \epsilon, \mathcal{A}) = \textbf{true}\right], \Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N,\ H), w', N, H, \epsilon) = w\right]\right\}\\ &= \max\{\beta', \beta\} = \max\{2^{-m}/\beta, \beta\}\end{aligned}$$

hence complete the prove.

□