How to Correct More Errors in a Secure Sketch

Y-L Lai
yenlung.lai@monash.edu

August 13, 2018

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 fuzzy min-entropy under distribution uncertain setting. In other words, our new technique offers security guarantee for all family of input distribution, as long as the sources possessing "meaningful amount" of fuzzy min-entropy over some random distributions, parametrized by a chosen error correction code.

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 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 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 work, 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 offer 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 through performing XOR operation with a noisy string. The uniform string can be reproduced by another noisy string through of error tolerance, provided the noisy 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, often represented as different genuine points. The genuine points reveal a secret which is encoded by using 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 approach has introduced to solve these issues recently, 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, 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. Conversely, an upper bound of the tolerance distance translated to a lower bound on the entropy loss of the input.

Recent works by Fuller *et al.*, (2009) [FRS16] have defined the min-entropy with maximized chances for a variable W within distance t of w', as the fuzzy min-entropy

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

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 $\mathrm{H}_\infty(W) - \log(B_t(w')) \leq$

$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W)$ minus the loss signified by the hamming ball $B_t(w')$ of radius t, due to error tolerance.

However, certain non-uniform sources come with *more error than entropy* itself are not able to sustain under this crude measurement. Lots of discussion have been given by Canetti *et al.*, (2016) [CFP+16] on how the low entropy sources must be taken into consideration when constructing a fuzzy extractor (trivially, also refer to a secure sketch). Since the source entropy rate is lower than the error rate, simply deducting the entropy loss from the sources' min-entropy always output a negative value, hence, show no security. One typical example refers to biometric, i.e., IrisCode (Daugman, 2006) [Dau06]. The IrisCode is said to provide entropy of 249 bits. Whereas, for two IrisCode generated from the same user of each 2048 bits, have shown far more than 249 bits of errors. Therefore, this more error than entropy problem is indeed restricting the usage of a secure sketch from all kind of available sources.

Distribution uncertainty: The predictability of nearby point w' is not merely entropically connected, but it is also closely tied to the distribution of the sources. Given a source under a distribution where all points are far apart (larger than t), then, one has no scruple to tolerate the errors, since, it means the probability for an adversary to predict the nearby string w' within distance t is small. However, standard secure sketch must work on all distributions (when the input distribution is unknown). Under the worst scenario, the points might be distributed very close to each other. For any variable W over this 'worse case' distribution, the sketch must lose entropy, by means of the number of similar points within distance t, which allows error tolerance. Therefore, the entropy loss of the sketch would be bounded that is proportional to this value.

Fuller *et al.*, (2013) [FMR13] have shown 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. 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 family of distribution, included the worst case distribution over error tolerance distance t, 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 then entropy problem which is intended to be solved by using fuzzy min-entropy.

Reusability: Reusability property is introduced by Boyen (2004) [Boy04]. Given a user comes with a noisy input w (i.e., biometric), the user may enrol w for different applications. Each time the user enrols 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 $ss_1, ss_2, \ldots, ss_q$. 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 means higher probability of success in predicting w' within more considerable distance around w, thus, security degradation can be seen. For this reason, 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 $B_t(\phi')$[1] 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:

[1]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

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

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 belief is mainly because of knowing some values outside the ball $B_{t'}(w')$ must add advantages of predicting a value inside the ball $B_{t'}(w')$. On the ground, the lower bound security can offer by average fuzzy min-entropy over larger tolerance distance t, renders the lower bound entropy loss of the fuzzy min-entropy over smaller tolerance distance t'. View this way, the average-fuzzy min-entropy is useful for better monitoring the loss of the fuzzy min-entropy while providing optimal resilience. This definition is not new but combined merely the average min-entropy and fuzzy min-entropy notions.

Info. theoretic secure sketch with fuzzy min-entropy: 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. Notwithstanding its security roustness, 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. For this reason, fuzzy min-entropy takes the role to offer computational security for low entropy sources, without the need for info. theoretic security. Fuller *et al.,* [FMR13] have shown that, there existing some family of distribution, where $\mathrm{H}_\infty(W) = \mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W)$, and yet the residual entropy is surprisingly low (i.e., at most 2 bits remaining) to claim meaningful security on a source. This suggested fuzzy min-entropy can be used to construct info. theoretic secure sketch subjected to the source must possess "meaningful" amount of fuzzy min-entropy under the worst case distribution.

We constructed a pair of sketching and recover algorithm that offers info. theoretical security, and free 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 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 matrix space. In fact, in our exposition, we show that the min-entropy of the resilient vectors is bounded by the fuzzy min-entropy of the sources, which is parametrized by a randomly chosen error correction code. Our works supported a statement: high fuzzy min-entropy is necessary for a source to show meaningful security. We portrayed this with an info. theoretic secure sketch.

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 noise during the sketching phase. Eventually, we find out the noise 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-enrolment of a single source with multiple providers, yet offer security assurance to each of them, as long as the noise 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 cardinality $|w| = |w'|$. 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 matrix space $\{0,1\}^n$. At the same point, we let $t' = \lfloor(\xi - \epsilon)l\rfloor$ to describe the error tolerance distance over the smaller binary matrix space $\{0,1\}^l$, with some error parameter $\epsilon \in (0, 1/2)$.

Main idea: Suppose Alice wishes to conceal a noisy non-uniform string $w \in \{0,1\}^l$. Firstly, she has to add additional random noise to w, which can be easily achieved by performing an XOR operation on the input with a noise vector $e \in \{0,1\}^l$. We here described the noise added input with $w_e = w \oplus e$. Albeit the initial step is kind of counter-intuitive, we will show later it is necessary to offer higher security and reusability. Consequently, Alice has to tolerate the noise of the input including the newly introduced random noise, for exact recovery of w. To do so, we invoke the use of error correction code. Suppose 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 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 simplistic but comes with some significant 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 type 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 noise added to the input w while sketching.

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 $1 - P = 1 - \|w \oplus w'\| l^{-1}$. 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, $\|\delta\| \sim \text{Bin}(n, 1-P)$ follows binomial distribution of expected distance $\mathbb{E}[\|\delta\|] = n(1-P)$ and s.d. $\sigma = \sqrt{nP(1-P)}$. Therefore $\mathbb{E}[\|\delta\|] = n(1-P) = n \|w \oplus w'\| \, l^{-1}$ 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 noise $e \in \{0,1\}^l$ of weight $\|e\| = l\epsilon$ to the inputs, where $\epsilon \in (0, 1/2)$ (e.g., adding the noise 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 $\epsilon \in (0, \frac{1}{2})$ be the noise parameter and $\xi > 0$ be the tolerance rate of a $[n,k,t]_2$, code $\mathcal{C}_\xi$. Suppose a resilient vector $\phi' \in \Phi$ is generated from strings $w' \in W$. For two hamming ball $B_t(\phi')$ and $B_{t'}(w')$ of radius $t' = \lfloor (\xi - \epsilon) l \rfloor$ 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 $\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \in B_{t'}(w')]\right] \geq 1 - \exp(-2n\epsilon^2)$.*

Proof. For $W \in B_{t'}(w')$, it means that any string $w \in W$ must show an error rate of $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$. Based on Theorem 2, w can be used to produce its corresponding resilient vector $\phi \in \Phi$ that shows an expected offset with ϕ' described as $\mathbb{E}[\|\phi \oplus \phi'\|] = \mathbb{E}[\|\delta\|]$ s.t. $\mathbb{E}[\|\delta\|] \leq t - n\epsilon$ (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 inequality*, one able to calculate the average minimum probability:

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

and complete the prove. □

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

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 $\|w \oplus w'\| l^{-1} > \xi - \epsilon$ which can also be interpreted as $\|w \oplus w'\| l^{-1} \geq \xi + \epsilon$, or more precisely, $\|w \oplus w'\| \geq \lfloor (\xi + \epsilon) l \rfloor > t'$. According to Theorem 2, w is capable to produce its corresponding resilient vector $\phi \in \Phi$ that will has an expected offset with ϕ' described as $\mathbb{E}[\|\delta\|] \geq t + n\epsilon$.

Thus, there will be a maximum value of $t_{\max}$ s.t. $t_{\max} = \mathbb{E}[\|\delta\|] - n\epsilon$. Therefore, By using *Hoeffding inequality*, one able to calculate the average maximum probability, by symmetry:

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

and complete the prove. □

The results obtained from Collorary 1 and Corollary 2 imply the following statement: Once the noise is introduced into the input, the probability to find any resilient vector $\phi' \in \Phi$ close to its original reading ϕ within the ball $B_t(\phi')$ will be bounded due to the noise 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

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

by definition.

2 Preliminary

In this section, we briefly highlight and recall some classical notions required 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., $\mathsf{dis}(A, C) \leq \mathsf{dis}(A, B) + \mathsf{dis}(B, C)$.

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

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

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 $\max_w \Pr[W = w \mid W' = w']$. 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 $\min_{c,c^* \in \mathcal{C}} \mathsf{dis}(c, c^*)$

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 or parity check matrix:

- a $[n, k, d]_q$ linear code $\mathcal{C}$ can be specified as the set $\{Gv : v \in \mathbb{F}_q^k\}$ for an $n \times k$ matrix 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 $\{x : x \in \mathbb{F}_q^n \text{ and } Hx = 0^n\}$ for an $(n - k) \times n$ matrix 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[2] 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.

[2]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 New Construction-LSH Secure Sketch

We hereby provide the detail of our new design and construction on a pair of sketching and recover algorithm, that incorporated with LSH-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, describe later. Generally, the hamming-hash algorithm $\Omega^{\mathsf{ham-h}} : \mathcal{M}_1 \times [l]^n \rightarrow \mathcal{M}_2$ 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.

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 $i = 1..n$ **do**
  parse $x = w(N(i))$ // x is the $(N(i))$-th bits of w
  $\phi = \phi \| x$
endfor
return ϕ

3.2 LSH-Sketching

We denote the LSH-sketching algorithm that employed the hamming-hash algorithm, Ω and a $[n, k, t]_2$ code $\mathcal{C}_\xi$ 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., $ss = c \oplus \phi$. Remark that, we have the newly introduced random noise of parameter $\epsilon \in (0, 1/2)$ in our sketching phase, the sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$ used input w, N and e described as follow:

$\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, \epsilon)$

$r \leftarrow\!\!\$\, \{0,1\}^{k-l}$ // sample r uniformly at random
$e \leftarrow\!\!\$\, \{0,1\}^{l}$ // given the weight $\|e\| = l\epsilon$
$w_\epsilon = w \oplus e$;
$v = w \| r$;
$c = Gv$; // Given G is the generator matrix of $\mathcal{C}_\xi$
$\phi \leftarrow \Omega^{\mathsf{ham-h}}(w_\epsilon, N)$
$ss = c \oplus \phi$
return ss, N

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 cardinality 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 redundant symbols required. The resultant effect of random bit padding is the minimized entropy loss from the sketch after encoding takes place.

In fact, the notable idea of using padding strategy to reduce entropy loss on a 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.

3.3 LSH-Recover

For recovery, we denote the LSH-recover algorithm that employed the hamming-hash algorithm, Ω, and a $[n, k, t]_2$ code $\mathcal{C}_\xi$ with decoding algorithm f as $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$

Suppose who wish to perform the recovery process with another string $w' \in \{0,1\}^l$. He/she has to provide another resilient vector ϕ'. The resilient vector can be generated by using the same hamming hash with input w' and the public known integer string N. The offset is manifested by the way of measuring the hamming distance on the resilient vectors pair, $\delta = \phi \oplus \phi'$. The recover algorithm $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}$ used input string ss, w' and N to recover w is described as follow:

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

$\phi' \leftarrow \Omega^{\mathsf{ham-h}}(w', N)$

$c \leftarrow \mathsf{f}(ss \oplus \phi')$ // assume $\|ss \oplus \phi'\| \leq t$

$v \leftarrow G^{-1}c$

// recover w through looking for the first l symbols of v

return w

If the final decoding process $\mathsf{f}(ss \oplus \phi')$ is successful, the algorithm returns a correct output w. Else, it will output a null result.

A brief description of the recovery mechanism is given as follow. Saying that, the recovery is allowed to success when the error rate $\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$. Suppose Bob has intercepted with a sketch $ss = c \oplus \phi$. Firstly, he has to generate a resilient vector $\phi' \leftarrow \Omega^{\mathsf{ham-h}}(w', N)$. With the previous condition applied, the number of disagreed position (i.e., the offset, δ) between ϕ and ϕ', is expected to be low as well, by means of LSH property. The hamming weight of the offset can be conveniently represented as $\|\delta\| \leq t$, with some distance t. Immediately after this, Bob can simply performs $ss \oplus \phi'$ to output the nearest codeword c'.

Given the original codeword $c \in \mathcal{C}_\xi$ of tolerance distance t, it follows the decoding process $\mathsf{f}(c \oplus \delta) = \mathsf{f}(c') = \mathsf{f}((c \oplus \phi) \oplus \phi') = \mathsf{f}(c \oplus (\phi \oplus \phi'))$. Since $\|\phi \oplus \phi'\| = \|\delta\| \leq t$, it means the offset can be tolerated through decoding and output the original codeword c subsequently. Thereafter, v can be recovered successfully and so w by looking at the first l symbols of v.

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 topic regarding resilience bound to the following section. Nevertheless, for better illustration, we will use an efficient computational class of error correction code-BCH code [Ber15] which is also considered as a $[n, k, t]_2$ liner code, to show the resilience of our proposal.

For the seek of simplicity, we combined the results from Corollary 1 and Corollary 2. Formally, we let $\beta = \mathbb{E}_{w' \leftarrow W}\left[\max_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \notin B_{t'}(w')]\right]$. Accordingly, the average minimum probability to find Φ in ball $B_t(\phi')$ can be represented as $\mathbb{E}_{w' \leftarrow W}\left[\min_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \in B_{t'}(w')]\right] = 1 - \beta$.

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 $\tilde{\mathrm{H}}^{\mathrm{fuzz}}_{t,\infty}(\Phi|W \notin B_{t'}(w')) = -\log(\beta)$. From this perspective, 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 matrix space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $w, w' \in W$, $N \in [l]^n, \epsilon \in (0, \frac{1}{2})$ and a $[n,k,t]_2$ linear code $\mathcal{C}_\xi$ with $\xi = tn^{-1}$. For a sketch ss generated through $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, \epsilon) = ss$, then the probability for $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(ss, w', N) = w$ is overwhelming when 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 the sake of practicability, we choose to use the Berlekamp-Massey decoding algorithm [MS77] to describe our decoding function, f, which is an efficient one.

Proposition 1. *Given Berlekamp-Massey decoding algorithm is used, $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$ is complete in (ξ,ϵ) 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 decoding function, f with Berlekamp-Massey decoding algorithm can decode the corrupted codeword, c' if $\|\delta\| \leq t$, described as $\mathsf{f}(c') = \mathsf{f}(c \oplus (\phi \oplus \phi')) = c$. Eventually, one has $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(w', N) = w$. The efficiency follows the decoding algorithm itself.

The remaining prove utilized the result in Corollary 1. Suppose $\|w \oplus w'\| l^{-1} \leq \xi$, with additional random noise e added to the input w, such as $\|w \oplus w' \oplus e\| l^{-1}$. The noise 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:

$$\begin{aligned} 1 - \beta &= \mathbb{E}_{w' \leftarrow W} \left[\min_{\phi'} \Pr[\Phi \in B_t(\phi') \mid W \in B_{t'}(w')] \right] \\ &\geq 1 - \exp(-2n\epsilon^2) \end{aligned}$$

Observe that $1 - \beta$ is overwhelming with negligible quantity $\beta = \exp(-2n\epsilon^2)$ when n is sufficiently large hence the proposition is prove. □

Proposition 1 concluded that given a $[n,k,t]_2$ 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.

Due to the newly introduced noise e during the sketching phase, we now have more errors need to be corrected for the exact recovery of w. For this reason, the chosen value for ϵ is preferred to be as small as possible, or one can always use a code $\mathcal{C}_\xi$ with larger value of tolerance distance t (or larger value of ξ). Precisely, the error parameter ϵ can set to minimum $\epsilon = l^{-1}$, since a single bit different in between w and w' always shifted the error rate of quantity l^{-1}.

For better illustration, it is useful to have an example to show how our results can be applied practically with a BCH code.

Example 1[3] Let $w, w' \in \{0,1\}^l$, $l = 100$. Suppose one wishes to correct some errors say 5 bits. It means $5/l = 0.05 = \|w \oplus w'\| l^{-1}$. Therefore, one needs to choose a code $\mathcal{C}_\xi$ comes with $\xi \geq 0.05 + \epsilon$ which can be easily achieved by using a $[511, 103, 61]_2$ BCH code with $\xi = 61/511 = 0.1194$. On the next, if one wishes to have overwhelming probability, i.e., 0.9 in correcting the errors, he/she must calculate s.t. $1 - \beta = 0.9 = 1 - \exp(-2(511)\epsilon^2)$. Eventually, it follows that $\epsilon = \pm 0.04$, which means that he/she can add additional random noise e of weight $\|e\| = l\epsilon = 4$ bits while sketching.

4.1 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

[3] The example itself does not mean anything about security, but merely to show resilience

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. 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 stochastic error process, or the computational power of an adversary to carry out decoding successfully. For examples, it is imprudent to assume the errors occur in a biometric always follow certain distribution. Other than this, computational hardness assumption must be applied to show meaningful fuzzy min-entropy security 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 $d = 2t + 1$. Saying so, the Plotkin bound (see [Sud01]) which has bounded maximum number of codeword in a code of blocklength n and minimum distance d. For instance, there can be only at most $2n$ codewords with $d > n/2$, which means that there have no error correction code can correct $4/n$ errors with probability one and so for a secure sketch as well.

Nevertheless, we have shown that our construction is not in the perfect correctness model but rather the slightly relaxed notion called *probabilistic correctness*. With this relaxed notion of correctness, the decoding will not succeed with probability one, rather $1 - \beta$, with some probability to fail. 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. Besides, given $\beta = 0$, which implies the average fuzzy min-entropy is infinitely large ($-\log(\beta)$); Recall the average min-entropy also reveals to the losses of fuzzy min-entropy from the sources, this will always lead to negative result to security. Therefore, slightly relaxed on the notion of correctness is relevance to show security in our case.

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 accordance with 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:

$$\frac{k}{n} < 1 - h_2(p) - \delta(n)$$

where $h_2(p) = -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function of error rate p and $\delta(n) = o(1)$. 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 p as long as p is smaller than $1/2$. Therefore, we obtain

Proposition 2. *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 $\|w \oplus w'\| l^{-1} + \epsilon < 1/2$.*

Proof. This proof is straightforward by using the Shannon noisy channel coding theorem. Summing up the effect due to the introduced random noise and the original offset between w and w', the total error rate can be described as $p = \|w \oplus w'\| l^{-1} + \epsilon$, hence, with $p < 1/2$, $\frac{k}{n} < 1 - h_2(p) - \delta(n)$ is always possible with sufficiently large n through LSH hamming strategy with n iteration. High resilience will eventually follow with sufficiently large n reasoned by the completeness itself. □

Proposition 2 concluded that the 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 capable to correct the errors of number up to Shannon bound with some code $\mathcal{C}_\xi$. Computationally efficient code achieve this bound is later 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 matrix 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 assumption is orthodox to show error tolerance up to distance t. We aim 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, $\mathcal{A} : \mathcal{M}_1 \times \mathcal{M}_2 \times [l]^n \to \mathcal{M}_1$ is just an algorithm that is computationally unbounded, purposed to recover w from a sketch $ss \in \mathcal{M}_2$, with integer string $N \in [l]^n$ and $w' \in \mathcal{M}_1$, where $\mathcal{M}_2 = \{0,1\}^n$. Meanwhile, we imposed an additional requirement for $\mathcal{A}$ in running the attack, to be specific, once it 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 $\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \mathcal{A})$ with input LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$, N, ϵ, and $\mathcal{A}$ as follow:

$\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \mathcal{A})$

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** $\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, \epsilon), w', N) = w \;\; \& \;\; \|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$
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 errors e has added to the input during the sketching phase, it should introduce some uncertainty to the total error rate between w and w' over the resilient vectors. This noise effect must be taken into account for retrospective security study. For instance, given $\|w \oplus w'\| \geq \xi$, after additional noise e of weight $\|e\| = l\epsilon$ 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 $\|w \oplus w'\| \leq \xi - \epsilon$, focusing on both cases when $\|w \oplus w'\| \geq \xi + \epsilon$ and $\|w \oplus w'\| \leq \xi - \epsilon$ should complete our security evaluation. Therefore, the step 2 and 3 of the attack is carefully designed for this purpose, to cover both scenarios when $\|w \oplus w'\| \, l^{-1} \geq \xi$ and $\|w \oplus w'\| \, l^{-1} \leq \xi$. For this cause, we have the following definition.

Definition 2. *Let β and β' be some negligible quantity. Let W and Φ be some random variable over a matrix space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $N \in [l]^n$, and $\epsilon \in (0, \frac{1}{2})$, 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 $\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \mathcal{A}) = \mathbf{true}\right] \leq \beta'$ and $\Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, \epsilon), w', N) = w\right] \leq \beta$ 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 2

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 the completeness(resilience) of the

algorithm itself. Formally, given any pair of string $w, w' \in \mathcal{M}_1$, under the case when $\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$, the offset can be tolerated with overwhelming probability at least $1 - \beta = 1 - \exp(-2n\epsilon^2)$ for negligible β if n is sufficiently large.

Due to the introduced noise 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 $\|w \oplus w'\| l^{-1} < \xi + \epsilon$. 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 $\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$. The second part of the inequality comes with slight refinement but still conserved the original interpretation, since, it embodies all possible error rates in the resilient vectors when $\|w \oplus w'\| < \xi + \epsilon$.

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

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

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

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

Proof for Part (2): 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:

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

The second line result obtained by using *Bayesian law*. For the third line result, it follows: given $t' = \lfloor(\xi - \epsilon)l\rfloor$, and let $\alpha = \Pr[\|w \oplus w'\| \leq t'] \leq \max_{w'} \Pr[W \in B_{t'}(w')]$. Then, by combining the result from Corollary 1 and 2, $\Pr[\|\delta \leq t\|] = (1-\beta)\alpha + \beta(1-\alpha)$. We also claim that $\alpha \leq \beta$ since $-\log(\alpha) = \mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W)$ and $-\log(\beta) = \mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(\Phi | W \notin B_{t'}(w')$. Recall the average fuzzy min-entropy reveals the loss of fuzzy min-entropy which supported the claim.

In the end, the maximum probability of recovering w for both part (1) and part (2) described as $\max\{\beta, \beta'\}$. Converting the above result into entropy calculation, since the sources must contain certain amount of fuzzy min-entropy, $-\log(\alpha) = \mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) \geq m$, with $m > 0$. Recall the fuzzy min-entropy can be bounded by the min-entropy minus the loss (residual entropy) described as $\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(\Phi) \geq \mathrm{H}_\infty(\Phi) - \lambda$. Thus, given $\mathrm{H}_\infty(\Phi) = \mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) \geq m$, the final results follow:

$$\begin{aligned}
&\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(\Phi) \geq \mathrm{H}_\infty(\Phi) - \lambda = \min\{-\log(\beta'), -\log(\beta)\} \\
&\geq \min\{-\log(2^{-m}/\beta), -\log(\beta)\}
\end{aligned}$$

with $\beta = \exp(-2n\epsilon^2)$. The entropy loss simply follows $\lambda \leq -\log(\beta)$ or $\lambda \leq m - (-\log(\beta))$ holds under the cases when $m/2 \leq -\log(\beta)$ and $m/2 > -\log(\beta)$ respectively. □

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 the fuzzy min-entropy notion can use to construct info. theoretic secure sketch.

Therefore, to show meaningful security, the sources must at least come with sufficient amount of fuzzy min-entropy (with tolerance distance t') that solely depends upon the system requirements. An alternative to always ensure meaningful security can be provided is to have a precise knowledge setting for 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 certain distribution W over $\mathcal{M}_1$, which has no fuzzy min-entropy (with tolerance distance t') 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 fuzzy min-entropy, we do not give up the forest for one tree. In light of this, showing security to all family of distribution is necessary 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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$.

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

Security Bound for β-Correct Secure Sketch		
Computational	Best possible security	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(1-\beta)$
Computational	FRS sketch(universal hash functions) [FRS16]	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(\frac{1}{\beta}) - \log\log(\mathsf{supp}(\mathsf{W})) - 1$
Computational	Layer hiding hash (strong universal hash function)[WCD+17]	$\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(W) - \log(\frac{1}{\beta}) - 1$
Info. theoretic	**LSH sketch**	$\min\{\mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) + \log(\beta), -\log(\beta)\}$[4]

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

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 $\|w \oplus w'\| \, l^{-1}$. 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 $\min\{\mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) + \log(\beta), -\log(\beta)\}$. Apparently, this residual entropy is always bounded by the fuzzy min-entropy of the source instead of the blocklength of the code n.

Perceivably, fuzzy 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 fuzzy 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.

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

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 3. *The best 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* $m/2$ *and the errors can be corrected with overwhelming probability at least* $1 - 2^{-m/2}$.

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:

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

It follows that, the errors can be corrected with overwhelming probability at least $1 - \beta = 1 - 2^{-m/2}$ □

5.2 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 noise content in different IrisCode of the same user lye in between $10\% - 35\%$ [FSS17].

Suppose one wishes to correct $l/3$ number of errors, given the desired security of 40 bits, therefore:

$$\begin{aligned} \beta &= 2^{-40} \\ \exp(-2n\epsilon^2) &= 2^{-40} \\ \epsilon &= \pm\sqrt{\frac{1}{3n}(13.8629)} \end{aligned}$$

We reason that, using average fuzzy min-entropy should enough to show security because we have $\min\{m + \log(\beta), -\log(\beta)\}$, where $m \approx 249$ in this example. Therefore, the best possible security $\beta \leq 2^{-249/2}$ must hold.

If one wishes to choose a $[n, k, t]_2$ code $\mathcal{C}_\xi$ with $n = 2^{13}$, then ξ must come as:

$$\begin{aligned} \xi &= tn^{-1} \geq 1/3 + \epsilon \\ t &\geq 2^{13}(1/3 + 0.0238) \geq 2926 \end{aligned}$$

The errors can be corrected with overwhelming probability of $1 - 2^{-40}$. As shown by the above example, in order to correct $1/3$ fraction of errors, one must choose a code $\mathcal{C}_\xi$ with $\xi \geq 1/3 + \epsilon$.

The security level can be increased through the increment of ϵ, but this will lead to a higher value of ξ is required for a chosen code $\mathcal{C}_\xi$. If one wishes to show higher security, then he/she has no choice to increase the blocklength n, since the errors will not be able to get corrected if the total error rate $1/3 + \epsilon \geq 1/2$ due to Shannon bound.

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-enrols. The reusability property allows the reuse/re-enrolment 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 enrolment of correlating samples $w_1, w_2, \ldots, w_q$. Each enrolment should return a sketch ss_i which possesses individual security that holds even under the existence of other sketches for $i \in \{1, \ldots, q\}$. 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 $w_1, w_2, \ldots, w_q$, 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 isometry 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 $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$. This group of adversaries implicitly defined different distributions over the published sketch $\{ss_1, \ldots, ss_q\}$. 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-enrolment of the original input with multiple providers that may not trust each other. They utilized set of functions $f_1, \ldots, f_q$ to sample $w', \ldots, w_q$ 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 particular distribution W_i. The security is defined computationally by using fuzzy min-entropy and holds for large class of family of distributions $\{W_1, \ldots, W_q\}$ 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 matrix 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 $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$ comes with unlimited computation power. To do so, we have introduced additional random noise $\{e'_1, \ldots, e'_q\}$, s.t. $|e'_i| \leq l\epsilon'$ where $\epsilon' < \epsilon$ acting as perturbation to the input w^* to sample a list of correlating reading $\{w_1, \ldots, w_q\}$. The usage of random noise is better fit to real case scenario, since any perturbation occurs during re-enrolment must cause certain amount of bits flip to the original sample w^*.

Briefly, we seek to show reusability defined in the information-theoretical sense as well. Our work is 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 $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$. Formally, each adversary $\mathcal{A}_i : \mathcal{M}_1 \times \mathcal{M}_2 \times [l]^n \to \mathcal{M}_1$ is simply an algorithm that is computationally unbounded to output w_i with a public sketch, $ss \in \mathcal{M}_2$, an integer string $N \in [l]^n$ and $w' \in \mathcal{M}_1$. 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 $\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_\mathsf{q}\})$ with input LSH-sketching algorithm $\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}$, N, ϵ, ϵ' and $\mathcal{A}_i$ as follow:

$\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_q\})$

$w^* \leftarrow\!\!\$\, \{0,1\}^l, \quad w' \leftarrow\!\!\$\, \{0,1\}^l$
if $\|w^* \oplus w'\| l^{-1} \leq \xi$, **repeat step 1 until** $\|w^* \oplus w'\| l^{-1} \geq \xi$
for $i = 1 : q$
$e'_i \leftarrow\!\!\$\, \{0,1\}^l$ // the weight $\|e'_i\| = l\epsilon'_i \leq l\epsilon'$
$w_i = w^* \oplus e'_i$
    if $\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w_i, N, \epsilon), w', N) = w_i \;\;\&\;\; \|w_i \oplus w'\| l^{-1} \leq \xi - \epsilon + \epsilon'$
    Output **true**
    else
    Output **false**
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-enrolment, with some perturbation. Reusability can only be claimed if the security holds for all adversaries corresponds to individual re-enrolment of w^* respectively. Since each re-enrolment 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 3. *Let β_2 and β'_2 be some negligible quantities. Let W and Φ be some random variable over a matrix space $\mathcal{M}_1 = \{0,1\}^l$ and $\mathcal{M}_2 = \{0,1\}^n$ respectively, where $l < n$. Given $N \in [l]^n, (\epsilon, \epsilon') \in (0, \frac{1}{2})$, and $\xi > 0$, suppose $\epsilon > \epsilon'$, 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 $(\max\{\beta_2, \beta'_2\}, \epsilon^*, q)$-reusable if the probabilities $\max_q \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_q\}) = \textbf{true}\right] \leq \beta'_2$ and $\max_i \Pr\left[\mathcal{A}_i(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w_i, N, \epsilon), w', N) = w_i\right] \leq \beta_2$ for a group of computational unbounded adversary $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$, with $\epsilon^* = \epsilon - \epsilon'$.*

Recall we have initially introduced an error e of weight $\|e\| = l\epsilon$ during sketching. Given $\epsilon' < \epsilon$, it means $\|e' \oplus e\| = l(\epsilon \pm \epsilon')$. Suppose the error rate between w^* and w' satisfies $\|w^* \oplus w'\| \geq \xi$, the total noise 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 $\epsilon^* = \epsilon + \epsilon'$ or $\epsilon^* = \epsilon - \epsilon'$, 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 noise 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 $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. Indeed, our exposition under Attack_2 utilized the second interpretation, by letting $\epsilon^* = \epsilon - \epsilon'$.

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 value $m > 0$.*

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' of weight $\|e_i\| = l\epsilon'_i \leq l\epsilon'$ to show reusability. Given the error rate for w^* and w' satisfies $\|w^* \oplus w'\| \geq \xi$, the noise (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 noise added in during sketching phase, The final error rate thus can be likely described as $\|w_i \oplus w'\| l^{-1} \geq \xi + \epsilon \pm \epsilon'_i$ and $\|w_i \oplus w'\| l^{-1} \geq \xi - \epsilon \pm \epsilon'_i$. Likewise the single adversary setting, the second scenario can be conservatively interpreted by using the inequality $\|w_i \oplus w'\| l^{-1} \leq \xi - \epsilon \pm \epsilon'_i$

Therefore, the error rate in the resilient vectors can simply analysed under these two cases, hence, our proof can be divided into only two parts, 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, $\epsilon^* = \epsilon - \epsilon'$ throughout the whole prove to picture the reduction of our result from multiple adversaries to single adversary setting.

Proof for part (1): Since we have multiple adversaries needed to consider, this part is simply finding 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:

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

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.

Proof for part (2): The main proof for part (2) is to show security hold for all adversaries $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$. Formally, it characterized by Attack_2 which can be described as:

$$\max_q \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_q\}) = \mathbf{true}\right] = \max_{w_i} \Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi - \epsilon \pm \epsilon'_i \;\middle|\; \|\delta\| \leq t\right]$$

Firstly, we have to solve for $\Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi - \epsilon \pm \epsilon'_i \;\middle|\; \|\delta\| \leq t\right]$, then only follow by its maximum value. We can use the Bayesian law, and the result from Corollary 1 and 2. The final outcome should follow the prove in Theorem 3 (part (2)) described as:

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

This result depends upon the noise parameter ϵ_i for each $\mathcal{A}_i$. Accordingly, t'_i is now described as $t'_i = \lfloor(\xi - \epsilon \pm \epsilon'_i)l\rfloor$, and $\alpha_{2,i} = \Pr[\|w_i \oplus w'\| \leq t'_i] \leq \max_{w'} \Pr\left[W \in B_{t'_i}\right]$.

With $\alpha_2 \leq \max_{w'} \Pr[W \in B_{t'}]$, where $t' = \lfloor(\xi - \epsilon^*)l\rfloor$ refers to the maximum value of t'_i (e.g., $\max_i(t'_i)$), and $\max_i \beta_{2,i} = \beta_2$ (prove in part (1)), the maximum probability is thus:

$$\max_q \Pr\left[\mathsf{Attack}_2(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \epsilon', \{\mathcal{A}_1, \ldots, \mathcal{A}_q\}) = \mathbf{true}\right] = \max_{w_i} \Pr\left[\|w_i \oplus w'\| \, l^{-1} \leq \xi - \epsilon \pm \epsilon'_i \;\middle|\; \|\delta\| \leq t\right]$$
$$\leq \max_i \frac{\alpha_{2,i}}{\beta_{2,i}} = \frac{\alpha_2}{\beta_2}$$

The maximum value of t' is reasoned as follow: we refer $+\epsilon'_i$ instead of $-\epsilon'_i$ to maximize t'_i, since $(\xi - \epsilon + \epsilon') \geq (\xi - \epsilon - \epsilon')$. Thus $\max_i(t'_i) = \max_{\epsilon'_i} \lfloor(\xi - \epsilon + \epsilon'_i)l\rfloor = \lfloor(\xi - \epsilon + \epsilon')l\rfloor = \lfloor(\xi - \epsilon^*)l\rfloor$.

Viewed this way, a true result can only obtain if the adversary $\mathcal{A}_i$ succeeded in recovered w_i and able to show the sampled pair (w_i, w') comes with error rate $\|w_i \oplus w'\| \, l^{-1} \leq \xi - \epsilon^* \leq \xi - \epsilon + \epsilon'$. Therefore, this proves itself indeed follows Attack_2 terminology without contradiction.

Lastly, we assign some value $m > 0$ to bound $\alpha_2 \leq 2^{-m}$. Since $-\log(\alpha)$ is the fuzzy min-entropy of source W with tolerance distance $t' = \lfloor(\xi - \epsilon^*)l\rfloor$, some minimum value of entropy is required to show security. Doing so eventually leads to 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 $\{\mathcal{A}_1, \ldots, \mathcal{A}_q\}$. The prove follows with $q = \infty$.

□

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-enrolment of the input w^* for $q = \infty$ number of times as long as the error (perturbation) e' has bounded weight $\|e'_i\| \leq e'$ for $i = \{1, \ldots, q\}$. 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 ϵ^* (multi-adversaries). We therefore obtain the following proposition

Proposition 4. *If a pair of LSH-sketching and recover algorithm* $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,C_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,C_\xi,\mathsf{f}} \rangle$ *is* $(\max\{\beta_2, \beta_2'\}, \epsilon^*, q)$*-reusable, it is also a* $(\mathcal{M}_2, m, \min\{-\log(\beta_2), -\log(\beta_2')\}, t)$ *secure sketch.*

7 Acknowledgement

This paper is a pre-print version. The author is welcoming any critical comments and discussion to improve the quality of this works further.

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.

[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. 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`.

[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.

[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.

[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.

[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 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). 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, $\mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}}(w', N) = w$ is overwhelming, with probability at least $1 - \beta \geq 1 - \exp(-2n\epsilon^2)$, for negligible β. This argument holds for any $[n, k, t]_2$ linear code $\mathcal{C}_\xi$ with suitable choice of error tolerance rate $\xi = tn^{-1}$, given n is sufficiently large.

Security: We now argue in the security of $\langle \mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, \mathsf{Rec}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi,\mathsf{f}} \rangle$. 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 noise 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.

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

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

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 noise 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 $\|w \oplus w'\| l^{-1} < \xi + \epsilon$. 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 $\|w \oplus w'\| l^{-1} < \xi + \epsilon$: Recall after the noise e of weight $\|e\| = l\epsilon$ is included, it may lead to either $\|w \oplus w'\| \geq \xi + \epsilon$ or $\|w \oplus w'\| \geq \xi - \epsilon$. To show security when $\|w \oplus w'\| l^{-1} < \xi + \epsilon$, it is enough to just focus on the case when $\|w \oplus w'\| l^{-1} \leq \xi - \epsilon$, since, it conserves the original interpretation and embodies all possible error rate in the resilient vectors as well. Based on the attack given above, adversary $\mathcal{A}$ is given ss, N and w'. If the adversary $\mathcal{A}$ able to carry out decoding successfully, this attack will output true only if the sampled pair (w, w') comes with error rate $\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon$. We need to measure the probability as follow:

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

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 *Bayesian law*:

$$\begin{aligned}
&\Pr\left[\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon \;\middle|\; \|\delta\| \leq t\right] = \frac{\Pr\left[\|\delta\| \leq t \;\middle|\; \|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon\right] \Pr\left[\|w \oplus w'\| \, l^{-1} \leq \xi - \epsilon\right]}{\Pr[\|\delta\| \leq t]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0}\right]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right] + \Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\right] \Pr\left[\mathsf{Event}_{b=1}\right]} \\
&= \frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right] + \Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\right] \left(1 - \Pr\left[\mathsf{Event}_{b=0}\right]\right)} \\
&= \frac{1}{1 + \frac{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}](1 - \Pr[\mathsf{Event}_{b=0}])}{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}] \Pr[\mathsf{Event}_{b=0}]}} \leq \frac{1}{\frac{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}](1 - \Pr[\mathsf{Event}_{b=0}])}{\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}] \Pr[\mathsf{Event}_{b=0}]}} \\
&\leq \frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right] \Pr\left[\mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\right] \left(1 - \Pr\left[\mathsf{Event}_{b=0}\right]\right)} = \left(\frac{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}\right]}{\Pr\left[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=1}\right]}\right) \left(\frac{\Pr\left[\mathsf{Event}_{b=0}\right]}{1 - \Pr\left[\mathsf{Event}_{b=0}\right]}\right) \quad (3)
\end{aligned}$$

Let $\alpha = \Pr[\mathsf{Event}_{b=0}]$, as $t' = \lfloor(\xi - \epsilon)l\rfloor$, α can be rewritten as:

$$\begin{aligned}\alpha = \Pr[\mathsf{Event}_{b=0}] &= \Pr[\|w \oplus w'\| \leq t'] \\ &\leq \max_{w'} \Pr[W \in B_{t'}(w')]\end{aligned}$$

Straight away, we use the results from Corollary 1 and Corollary 2 to compute the maximum and minimum probability for the events $\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}]$ and $\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}]$ to occur respectively. It follows:

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

also,

$$\Pr[\mathsf{Event}_{a=0} \mid \mathsf{Event}_{b=0}] \geq \min_{t=t_{\min}} \Pr[\|\delta\| \leq t \mid \|w \oplus w'\| \leq t'] = 1 - \beta$$

Recall the definitions:

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

Since the average fuzzy min-entropy indicate the loss of the fuzzy min-entropy of the source, represented by variable W, therefore, $\alpha \leq \beta$ must hold, if not, there will have no security to show. With this argument, Eq. (3) can further simplify as

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

Eventually, letting $\beta' = \alpha/\beta$, we obtain the final result for part (2):

$$\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \mathcal{A}) = \mathbf{true}\right] \leq \beta' = \frac{\alpha}{\beta}$$

Combining the results from part (1) and part (2), it follows that, the maximum probability to decode the codeword is described as:

$$\max\left\{\Pr\left[\mathsf{Attack}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}, N, \epsilon, \mathcal{A}) = \mathbf{true}\right], \Pr\left[\mathcal{A}(\mathsf{SS}^{\mathsf{LSH}}_{\Omega,\mathcal{C}_\xi}(w, N, \epsilon), w', N) = w\right]\right\} = \max\{\beta', \beta\}$$

The remaining prove follows by simply convert the above result into entropy calculation. Precisely, by fuzzy min-entropy definition, the residual entropy of Φ is equivalent to its fuzzy min-entropy over tolerance distant 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, thus:

$$\begin{aligned}\mathrm{H}^{\mathrm{fuzz}}_{t,\infty}(\Phi) \geq \mathrm{H}_\infty(\Phi) - \lambda &= \min\{-\log(\beta'), -\log(\beta)\} \\ &\geq \min\{-\log(\alpha/\beta), -\log(\beta)\} \\ &\geq \min\{\mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) + \log(\beta), -\log(\beta)\} \\ &\geq \min\{m + \log(\beta), -\log(\beta)\}\end{aligned} \tag{4}$$

The last line conversion argued as follow. We utilized the min-entropy notion by assigned some value to $\mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) \geq m$, where $m > 0$. This quantity of min-entropy m must cover the worst case distribution of the sources W, which is signify by the smaller tolerance distant t', parametrized by a chosen $[n, k, t]_2$ code $\mathcal{C}_\xi$. Given the case when $m/2 \leq -\log(\beta)$, through direct comparison, we have $\mathrm{H}_\infty(\Phi) - \lambda \geq m + \log(\beta)$, where $\mathrm{H}_\infty(\Phi) = \mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) \geq m$, and the entropy loss would be bounded as

$\lambda \leq -\log(\beta)$. On the other hand, if $m/2 > -\log(\beta)$, one has $\mathrm{H}_\infty(\Phi) - \lambda \geq -\log(\beta)$, hence smaller entropy loss can be seen described as $\lambda \leq m - (-\log(\beta))$.

Finally, the probability:

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

Above result has shown that, even for computationally unbounded adversary $\mathcal{A}$, he/she must at least works with entropy of $\min\{m + \log(\beta), -\log(\beta)\}$. Moreover, since $\mathrm{H}_\infty(\Phi) = \mathrm{H}^{\mathrm{fuzz}}_{t',\infty}(W) \geq m$, this illustrated the fuzzy min-entropy of the sources is equivalent to the min-entropy of the resilient vectors. Thus, one has a $(\mathcal{M}_2, m, \min\{-\log(\beta), -\log(\beta')\}, t)$ secure sketch over larger matrix space $\mathcal{M}_2$ with $\beta' = 2^{-m}/\beta$ and $\beta = \exp(-2n\epsilon^2)$ and complete the prove.

□