CONTROLLING LEAKAGE OF BIOMETRIC INFORMATION USING DITHERING

Ileana Buhan, Jeroen Doumen, and Pieter Hartel

Distributed and Embedded Security, University of Twente
The Netherlands
email: ileana.buhan@utwente.nl

ABSTRACT

Fuzzy extractors allow cryptographic keys to be generated from noisy, non-uniform biometric data. Fuzzy extractors can be used to authenticate a user to a server without storing her biometric data directly. However, in the Information Theoretic sense fuzzy extractors will leak information about the biometric data. We propose as alternative to use a fuzzy embedder which fuses an independently generated cryptographic key with biometric data. As fuzzy extractors, a fuzzy embedder can be used to authenticate a user without storing her biometric information or the cryptographic key on a server. A fuzzy embedder will leak in the Information Theoretic sense information about both the biometrics and the cryptographic key. While both types of leakage are important, information leakage of the biometric data is critical since the cryptographic key as opposed to biometric data can be renewed. We show that constructing fuzzy embedders which leak no information about the biometrics is theoretically possible. We present a construction which allows controlling the leakage of biometric information, but which requires a weak secret at the decoder called dither. If this secret is compromised the security of the construction will degrade gracefully.

1. INTRODUCTION

A fuzzy extractor is a generic construction proposed by Dodis, *et al.* [4] which allows cryptographic keys to be generated from noisy, non-uniform data, such as biometrics. A fuzzy extractor can be used to authenticate a user to a server without storing her biometric data directly. This is important because the server may well be (partially) untrusted.

A fuzzy extractor is a pair of two functions. The first function is called the encoder, which is used once during enrollment. The second function is the decoder, which is used every time the user is authenticating to the server.
The encoder takes as input the users biometric x. It then outputs a public sketch p and a binary key k. For the same biometric x always the same pair (k,p) is output. The decoder takes as input a fresh measurement x' of the users biometric and the public sketch p, and outputs the secret key k if x and x' are similar enough (we will explain later what similar enough actually means).

A fuzzy extractor has two disadvantages. Firstly, the public sketch p and the authentication key k are extracted from the biometric and cannot be renewed. Secondly, it has been shown that it is impossible [5] to build fuzzy extractors for which the output does not leak information about the biometric input. Therefore, in [1] we propose an alternative construction to the fuzzy extractor termed a *fuzzy embedder* which takes as input an independently generated key k and the real valued (biometric data) x. Like a fuzzy extractor, a fuzzy embedder allows recovery of the binary key k, in the presence of x' (a corrupted version of x) at the decoder.

Contribution. We show that it is possible for a fuzzy embedder to make the output p statistically independent from the biometric input x or x'. We propose to use dithering techniques to break the correlation between the secret biometric information and the data that is made public. We give a practical construction based on quantization data-hiding codes [6] which requires a weak secret at the decoder. We show that if the secret is compromised, or if it is simply impossible to store secret information at the decoder, the security of the construction will degrade gracefully.

2. FUNDAMENTALS

Notation. By capital letters we denote random variables while small letters are used to denote realizations of random variables. A random variable X is endowed with a domain of definition, D_X and a probability density function $f_X(x)$. We denote the characteristic function of X by

$$F_X(u) = \int_{-\infty}^{\infty} f_X(x) e^{jux} dx.$$

In the rest of the paper we use a random variable X when referring to biometric data, P when referring to public data (the sketch) and K for binary strings that are used as cryptographic keys.

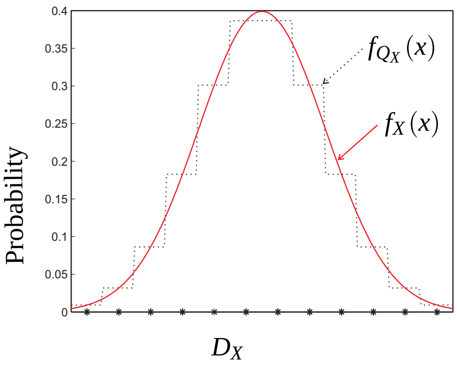

Figure 1: *By quantization, the probability density function of X $f_X(x)$ (continuous line) is transformed into $f_{Q_X}(x)$ (dotted line).*

Quantization. Quantization of variable X means sampling the probability density distribution of X and rounding the values of D_X to predefined points. By quantization the probability density function of the input X, $f_X(x)$,which is continuous, is transformed into the probability density function $f_{Q_X}(x)$, which is discrete, see *Figure* 1.

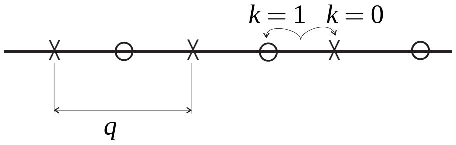

Figure 2: *Quantization of X with two scalar quantizers Q_0 and Q_1 both with step size q.*

Formally, a quantizer is a function $Q : D_X \rightarrow C_Q$ that maps each $x \in D_X$ into the closest *reconstruction point* in the set $C_Q = \{c_1, c_2, \ldots\}$ by

$$Q(x) = \arg\min_{c_i \in C} d(x, c_i)$$

where d is a suitable distance measure for the space X.

When X is one dimensional, Q is called a *scalar* quantizer. In the scalar case, the length of the decision region is called the *step size*. If all decision regions of a quantizer are equal the quantizer is *uniform*.

To measure the quality of the quantizer, the quantization error e is defined as the difference between the input of the quantizer, X, and its output $Q(X)$: $e = Q(x) - x$. The quantization error is minimized if the reconstruction point is the centroid of its decision region. The *Voronoi region* of a set of points is the subset of all points that are closer to one reconstruction point than to any other reconstruction point. If the points form a lattice the Voronoi regions of all reconstruction points are congruent. We refer then, to the Voronoi region of the lattice. The *size and shape* of the Voronoi region determines the tolerated noise between two values x and x'.

Quantization-Based Hiding Codes. Quantization based data hiding codes as introduced by Chen, *et al.* [3] (also known as quantization index modulation) can embed secret information into a real valued signal. We start with an example of the simplest case of embedding one bit of information into a single sample x.

Example. In a real value x we want to embed one bit of information, thus $k \in \{0,1\}$. For this purpose we use a scalar uniform quantizer with step size q, given by

$$Q(x) = q\lfloor \frac{x}{q} \rfloor.$$

The quantizer Q is used to generate a set of two new quantizers $\{Q_0, Q_1\}$ defined as:

$$Q_0(x) = Q(x + v_0) - v_0$$

and

$$Q_1(x) = Q(x + v_1) - v_1$$

where

$$v_0 = \frac{q}{4} \text{ and } v_1 = -\frac{q}{4}.$$

In *Figure* 2 the reconstruction points for the quantizer Q_1 are shown as circles and the reconstruction points for the quantizer Q_0 are shown as crosses.

The embedding is done by outputting the distance vector to the nearest × or ○ chosen by k. When during decoding x is perturbed by noise, the decoder will assign the received data to the closest × or ○ point, and output 0 or 1 respectively. The set of the two quantizers $\{Q_0, Q_1\}$ is called a QIM.

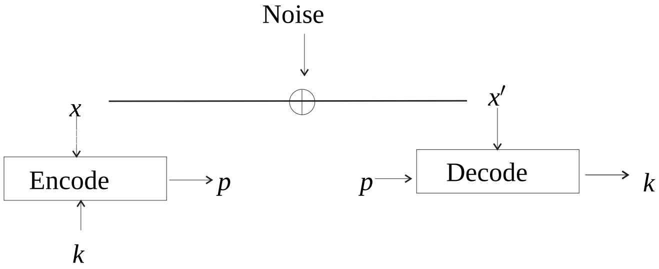

Figure 3: *A fuzzy embedder is a pair of two functions: the encoder and the decoder. The encode function, which takes as input a biometric descriptor x and a binary sequence k generated independently, is executed during enrollment. The result p is made public. The decode function, which takes as input a (possibly)corrupted biometric measurement x' and the public sketch p will output k if x and x' are close, is executed during authentication.*

General-QIM. The generalization of the one-dimensional QIM presented above is a lattice-QIM that replaces the scalar quantization by an n dimensional vector quantizer.

A $\texttt{QIM} : D_X \times K \rightarrow C_{\texttt{QIM}}$ is a set of quantizers $\{Q_1, Q_2, \ldots Q_N\}$ that maps x into one of the reconstruction points of the quantizers in the set. The quantizer is chosen by the input value $k \in K$ such that

$$\texttt{QIM}(x, k) = Q_k(x).$$

The set of all reconstruction points is $C_{\texttt{QIM}} = \bigcup_{k \in K} C_k$ where C_k is the set of reconstruction points of the quantizer Q_k.

The *number of quantizers* in the QIM determines the number of bits that can be embedded in x. By setting the number of quantizers in the QIM set and by choosing the shape and size of the decision region the performance properties can be fine tuned.

3. CONSTRUCTING A FUZZY EMBEDDER USING A QIM

We consider points in an n-dimensional universe, i.e. $D_X \subset \mathbb{R}^n$. The random binary string K is generated independently from the random variable X and K has a uniform distribution.

Definition. *A* fuzzy embedder *is a tuple (X, K, P,* Encode, Decode*), where $p =$* Encode(X,k)*, X is a random variable and $k =$* Decode(x,p) *when $x \in X$ and $p \in P$. The fuzzy embedder is ρ*-reliable *for the probability density $f_X(x)$ if*

$$P(\text{Decode}(x, \text{Encode}(X,k)) = k | X = x) \geq \rho,$$

for all $k \in K$. We say the scheme is (ε, δ)-secure *if:*

$$I(X;P) \leq \varepsilon \qquad \textit{and} \qquad I(K;P) \leq \delta.$$

Figure 3 illustrates a fuzzy embedder system. Below we give the intuition for the parameters of a fuzzy embedder. Reliability captures the capability of a fuzzy embedder to reconstruct the correct key from a noisy measurement of the biometric. Security measures the amount of secret information that is revealed by the output p. As we have two independent inputs we measure the leakage on both of them. If an attacker learns the value x she can reproduce the value k with the help of the public value p. However, if an attacker learns the secret key k, she could potentially circumvent the security altogether but cannot reproduce x. We illustrate this observation in the next example.

Example. In the fuzzy embedder example given in *Figure* 2, the attacker can choose between two different key values$\{\circ, \times\}$. Assume she learns the correct key, $\circ$. To find the correct value for x she still has to decide which of the reconstruction points of the quantizer $Q_\circ$ is closest to x. Without any other information this is an impossible task since the quantizer $Q_\circ$ has an infinite number of reconstruction points.

The public sketch p leaks information about both the random string k, denoted with δ, and the value x, denoted with ε. Since full disclosure of the string r is not enough to recover x, we conclude that $\varepsilon \leq \delta$. More details about the size of δ relative to the dimension of the parameters can be found in Buhan [1].

In the following we give a practical construction for a fuzzy embedder using `QIM` data hiding codes.

3.1 QIM fuzzy embedder basic construction

A `QIM`-fuzzy embedder is a hiding scheme where the encoder is defined as:

$$\text{Encode}(x,k) = \texttt{QIM}(x,k) - x,$$

and where the decoder is the minimum distance Euclidian decoder:

$$\text{Decode}(x',p) = \widetilde{Q}(x'+p),$$

where $\widetilde{Q} : D_X \longrightarrow D_K$, is defined as:

$$\widetilde{Q}(x') = \arg\min_{k \in D_K} d(x', C_k).$$

ρ-Reliability. Reliability is the probability with which the decode function maps x and x' to the same value k.

The public string p is the distance between x and $Q_k(x)$, the chosen reconstruction point. By adding the value p to x', $Q_k(x)$ will be detected as long as x and x' are within the bounds of the same Voronoi region. Thus, ρ is the probability that x and x' are in the same Voronoi region.

When x and x' are biometric samples collected from the same user, ρ can be seen as the *probability of detection* or the probability that two samples coming from the same user will be correctly identified as such. For a lattice quantizer we can write:

$$\rho \approx \int_V f_X(x)dx$$

where V is the Voronoi region of the lattice.

In earlier work, [2] we investigated the link between reliability and the size of the cryptographic key. It turns out that they are not independent. Increasing the number of bits in the cryptographic key k has a negative influence on the reliability.

ε-Security. To evaluate ε the statistical properties of $f_P(p)$ need to be investigated. Each p is computed as:

$$p = Q_k(x) - x, \qquad \forall x \in D_X, \qquad k \in D_K.$$

When $|K| = 1$ (or in other words `QIM` $= \{Q\}$ has only one quantizer) this simplifies to:

$$p = Q(x) - x.$$

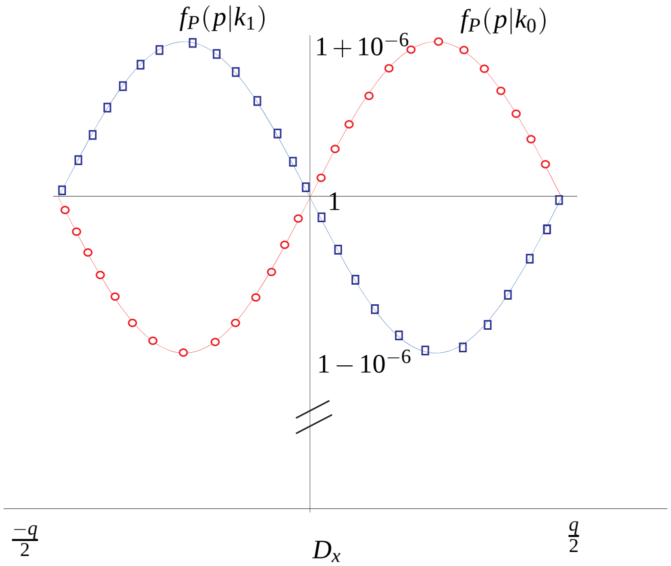

Figure 4: *Conditional probability densities functions of the public sketch P given two different keys $\{k_0, k_1\}$ that can be embedded, when $f_X(x) = N(0,1)$. We used the `QIM` construction from the example given in section 2.*

Now $f_P(p)$ is the same as the probability density $f_E(e)$ of the quantization error e. This observation makes analysis of the security properties of a `QIM` easier.

When $|K| > 1$, for each quantizer, Q_k we have a particular error probability density $f_E(e_k)$ which is equal to $f_P(p|k)$. *Figure* 4 illustrates the two error probability densities $\{f_P(p|k_1), f_P(p|k_0)\}$ of the `QIM` ensemble in the example of section 2. Conditioning on the key, we can compute $f_P(p)$ as

$$f_P(p) = \sum_{k \in D_K} f_P(p|k) \cdot f_K(k)$$

In the remainder of the paper, we analyze scalar quantization and leave lattice quantization as future work.

Widrow, [10] show how the probability density of $f_P(p|k)$ can be constructed: the value of the error results from the quantization of x falling at just the right places within all the quantization boxes. Thus when Q_k are scalar uniform quantizers with step size q and reconstruction points given by $Q_k(n \cdot q), \forall n \in \mathbb{Z}$, we can cut $f_X(x)$ into strips of length q, stacking the strips and then adding we arrive at:

$$f_P(p|k) = \begin{cases} \sum_n f_X(Q_k(nq) + p) & \text{if } |p| \leq \frac{q}{2} \\ 0, & \text{elsewhere.} \end{cases}$$

The definition of the encoder shows that there is a deterministic relationship between the input and the output of the quantizer and as a result ε cannot be zero. In spite of this deterministic relation Widrow, [10] shows that under certain circumstances (depending on the distribution of X) the quantization error can be made uniformly distributed on its support, but *not* statistically independent of X. Widrow, [10] gives sufficient conditions that $F_X(x)$ has to satisfy to make the quantization error uniform. Sripad, *et al.* [8] give necessary and sufficient conditions for the errors to be independent. Both results apply to uniform scalar quantizers, with step size q.

Proposition 1. (Sripad and Snyder) *The characteristic function of the input random variable satisfies*

$$F_X(\frac{2\pi n}{q}) = 0, \qquad \forall n \neq 0$$

if and only if the density function of the quantization error is uniform,

$$f_E(e) = \begin{cases} \frac{1}{q}, & -\frac{q}{2} \leq e < \frac{q}{2} \\ 0, & \text{otherwise.} \end{cases}$$

In our case, $f_E(e)$ in the result above can be replaced by $f_P(p|k)$, which also implies that when $f_K(k)$ is uniformly distributed also $f_P(p)$ is uniformly distributed. Unfortunately, the above result imposes conditions upon the statistics of the system input which in most practical cases cannot be controlled.

δ-Security. δ shows the amount of information that P reveals about the cryptographic key K. Information leaks whenever $f_P(p|k_i) \neq f_P(p)$, $\forall k_i \in D_K$. We look at the differences between the probability distributions of error for each quantizer.

In this paper, we focus only on the topic of ε-security, or the privacy leakage. We show, in the reminder of the paper how one can balance the private information leakage by the introduction of additional noise at the encoder. When no additional information is added to the input X of a quantizer as above, the quantizer is also known as *undithered quantization*. In the following when referring to a basic `QIM` construction we use the undithered-fuzzy embedder .

3.2 `QIM`-fuzzy embedder dithered construction

Schuchman, [7] shows how to circumvent Sripad's result by multiplying the characteristic function of the input signal, $F_X(u)$ by a desired function. A product of characteristic functions corresponds to convolution in the probability density domain. Convolution of probability densities corresponds to addition of independent random variables. Therefore according to Schuchman, [7] *any* input $f_X(x)$ can be forced to satisfy Widrow, [10] condition by adding a suitable independent variable. The independent variable is called *dither* (v).

Dithering is currently being used in processing of both digital video or audio data to reduce errors introduced by signal quantization. The premise is that quantization and requantization of digital data yields an error. If that error is repeating and correlated to the signal, the error has a determined pattern. By adding noise at the input signal the error patterns- are randomized. It was found that random errors compared to error pattern can reduce visual or audio artifacts.

For the fuzzy embedder it means that we can make the public sketch independent of the biometric data by adding an independent random variable to the input X. This means that the value of the ε parameter can be made arbitrarily small, without compromising δ, by adding to the input X, an independent source of noise with suitable statistical properties. There are two types of dithered quantization systems known in the literature. The first, is the *subtractive dither* quantization system (SD), see *Figure* 5. The dither v is added to the real valued x before it is fed into the encoder. In an SD-quantization system at the decoder, the same dither value is subtracted. The second is the *non-subtractive dither* quantization system (NSD), see*Figure* 6. The only difference between an SD and an NSD quantization system is that the dither is not available at the decoder.

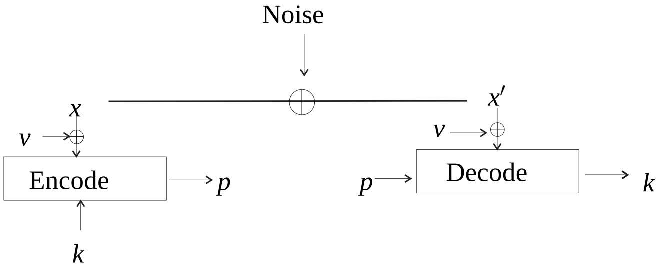

Figure 5: *A subtractive quantization fuzzy embedder system.*

SD-fuzzy embedder. When SD quantization is used the fuzzy embedder system is defined as below,

$$\text{Encode}(x+v,k) \quad = \quad \mathtt{QIM}(x+v,k)$$

where v is uniformly distributed and $-\frac{q}{2} \leq v < \frac{q}{2}$. The decoder function is defined as

$$\text{Decode}(x'-v,p) \quad = \quad \tilde{Q}(x'-v+p)$$

In this case the dither can be seen as a weak secret between the encoder and the decoder. The dither vector v is stored along with p encrypted with a key known only at the decoder.

In the proposition below Schuchman, gives a necessary and sufficient condition for the characteristic function of the dither, $F_v(u)$. All dithers that satisfy this condition render the quantization error $f_E(e)$ *uniform* and *statistically independent* of X in given by .

Proposition 2. (Schuchmans Condition) *In an subtractive quantizing system, the error will be statistically independent of the system input for* arbitrary *input distributions if and only if the characteristic function of the dither F_v satisfies the condition that*

$$F_v(\frac{n}{q}) = 0 \qquad \forall n \neq 0$$

Furthermore, the error will be uniformly distributed for arbitrary input distributions if and only if this condition holds.

It is natural to wonder which probability density functions satisfy the criterion in Schuchman result. One of the most simple candidates is a dither which is uniformly distributed on $\left(-\frac{q}{2}, \frac{q}{2}\right)$.

It was shown [9] that when subtractive dither quantization is used the properties of the public sequence p are ideal. Namely, p is statistically independent from the input sequence x and the correction capabilities are not affected by the noise introduced by the dither.

ρ-Reliability. To estimate reliability, we look at the noise tolerated between the input of the encoder $x+v$ and the input of the decoder $x'+v$.

$$\begin{aligned} \rho &= P(\text{Decode}(x'+v, \text{Encode}(X+v,k)) = k|X=x) \\ &= P(\text{Decode}(x', \text{Encode}(X,k)) = k|X=x) \end{aligned}$$

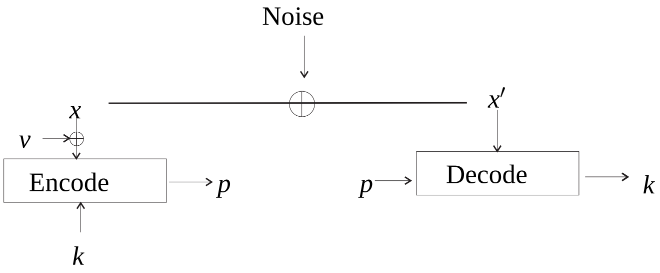

Figure 6: *A non-subtractive quantization fuzzy embedder system.*

This is exactly the same as the robustness in the case of an undithered fuzzy embedder system, section 3.1.

ε-Security. According to Schuchman's condition $f_P(p|k)$ is independent of $f_X(x)$, thus $f_P(p)$ is also independent of $f_X(x)$. We have as a result $\varepsilon = 0$.

NSD-fuzzy embedder. An SD-fuzzy embedder system might not be practical since it requires secret information to be available at the decoder. This reason would be impractical if that the decoder does not have encryption-decryption capabilities or another reason might be that the value of the dither vector v, is compromised. It is useful for practical reasons to study what happens to the reliability and security of a fuzzy embedder when the dither is not available at the decoder.
When NSD quantization is used the fuzzy embedder system is defined as below,

$$\text{Encode}(x+v,k) \quad = \quad \texttt{QIM}(x+v,k)$$

Here v is uniformly distributed and $-\frac{q}{2} \leq v < \frac{q}{2}$. The decoder function is defined as

$$\text{Decode}(x',p) \quad = \quad \tilde{Q}(x'+p)$$

ρ-Reliability. Again, we look at the noise tolerated between the input of the encoder $x+v$ and the input of the decoder x'.

$$\begin{aligned} \rho &= P(\text{Decode}(x'+v, \text{Encode}(X+v,k)) = k|X=x) \\ &= \int_q f_X(x+v)dx. \end{aligned}$$

The reliability of a NSD-fuzzy embedder is lower than both the reliability of a undithered-fe or a SD-fuzzy embedder .

ε-Security. Wannamaker, *et al.* [9] show that in an NSD quantizing system it is *not* possible to render the quantization error statistically independent or uniformly distributed for inputs of arbitrary distributions. It can render however any desired moments of the error independent of the input distribution. For many applications, controlling relevant error moments is just as good as having full statistical independence of the input and error processes.

4. CONCLUSIONS

We use the property of dithering in a novel way to reduce the correlation between information that is made public about biometric data and the biometric data itself. By dithering the biometric data we can make the published information statistically independent from the biometric data. This approach requires a weak secret to be available at the decoder. We further investigate what happens if the secret information available at the decoder is compromised. The effect of compromising the secret at the decoder is a reduction on the reliability with which the decoder finds the correct binary key, but the compromise has almost no effect on the information that is leaked about the biometric itself. As future work we intend to extend the above results to high dimensional lattice quantizers. Investigation of the exact relation between the ε and δ security is also left for future work.

REFERENCES

[1] I. Buhan. *Cryptographic Keys from Noisy Data: Theory and Applications*. PhD thesis, University of Twente, 2008(October).

[2] I. Buhan, J. Doumen, P.H Hartel, and R.N.J Veldhuis. Fuzzy extractors for continuous distributions. In R. Deng and P. Samarati, editors, *Proceedings of the 2nd ACM Symposium on Information, Computer and Communications Security (ASIACCS), Singapore*, pages 353–355, New York, March 2007. ACM.

[3] B. Chen and G.W. Wornell. Quantization Index Modulation Methods for Digital Watermarking and Information Embedding of Multimedia. *The Journal of VLSI Signal Processing*, 27(1):7–33, 2001.

[4] Y. Dodis, L. Reyzin, and A. Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. In Christian Cachin and Jan Camenisch, editors, *Advances in Cryptology - Eurocrypt 2004, International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004, Proceedings*, volume 3027 of *LNCS*, pages 523–540. Springer, 2004.

[5] Y. Dodis and A. Smith. Correcting errors without leaking partial information. In Harold N. Gabow and Ronald Fagin, editors, *Proceedings of the 37th Annual ACM Symposium on Theory of Computing(STOC), Baltimore, MD, USA, May 22-24, 2005*, pages 654–663. ACM, 2005.

[6] P. Moulin and R. Koetter. Data-hiding codes. *Proceedings of the IEEE*, 93(12):2083–2126, 2005.

[7] L. Schuchman. Dither Signals and Their Effect on Quantization Noise. *IEEE Transactions on Communications*, 12(4):162–165, 1964.

[8] A. Sripad and D. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5):442–448, 1977.

[9] R. Wannamaker, S. Lipshitz, J. Vanderkooy, and J. Wright. A theory of nonsubtractive dither. *IEEE Transactions on Signal Processing*, 48(2):499–516, 2000.

[10] B.K. Widrow, I. Kollar, and M.C. Liu. Statistical theory of quantization. *IEEE Transactions on Instrumentation and Measurement*, 45(2):353–361, 1996.