CONSTRUCTING SECURE CONTENT-DEPENDENT WATERMARKING SCHEME USING HOMOMORPHIC ENCRYPTION

Zhi Li, Xinglei Zhu†, Yong Lian* and Qibin Sun†*

*Department of ECE, National University of Singapore
†Institute for Infocomm Research, A-STAR, Singapore
Email: {lizhi, eleliany}@nus.edu.sg, {xzhu, qibin}@i2r.a-star.edu.sg

ABSTRACT

Content-dependent watermarking (CDWM) has been proposed as a solution to overcome the potential estimation attack aiming to recover and remove the watermark from the host signal. It has also been used for the application of content authentication. In this work, we first present an analysis on why some prior work on CDWM pose potential security problems due to their inherent cryptographic weakness. With the aim of achieving cryptographic level of security, we then propose a novel CDWM scheme based on homomorphic encryption and dirty paper precoding. The general idea is to introduce a decryption module before watermark detection to create some nonlinearity and thereby inhibit conventional watermark attacks based on linear operations. We conclude this paper by bringing up some thoughts on the integration of watermarking and cryptography.

1. INTRODUCTION

Consider a simplified model of Spread-Spectrum (SS) watermarking. Let $\mathbf{x}$ be the host signal,[1] $\mathbf{w}$ the watermark signal and $\mathbf{y}$ the signal after watermarking. Watermark embedding can be expressed in $\mathbf{y} = \mathbf{x} + \mathbf{w}$. We only consider $\mathbf{w}$ as a bipolar sequence with unit magnitude, i.e., $\mathbf{w} \in \{\pm 1\}^K$. Typically a scaling factor α on the watermark strength is necessary. In this paper we ignore α because such term can be canceled by scaling $\mathbf{y}$ by $1/\alpha$. For simplicity, we also do not consider any local scaling based on perceptual models. We detect the watermark by thresholding on the correlation coefficient $\mathcal{R}(\mathbf{y}, \mathbf{w})$.[2]

In many applications, we desire the repetitive watermark signal presented in the content, i.e.,

$$\mathbf{Y} = (\mathbf{y}^1, \dots, \mathbf{y}^N) = (\mathbf{x}^1 + \mathbf{w}, \dots, \mathbf{x}^N + \mathbf{w}). \qquad (1)$$

One possible application scenario of this embedding strategy is when we want the watermark to survive geometrical attacks by making sure that at least one copy of the watermark is detectable [1, 2, 3]. Similarly, this formulation can also be applied to the scenario of using the same watermark sequence $\mathbf{w}$ for many different pieces of content $\mathbf{y}^i$. However, this embedding approach poses vulnerability to estimation (or collusion, removal, etc.) attack, since one adversary knowing $\mathbf{Y}$ can extract the watermark signal $\mathbf{w}$ by averaging:

$$\hat{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}^i = \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^i\right) + \mathbf{w}. \qquad (2)$$

The first term goes to zero if $\mathbf{x}^i$ is zero-mean and N is large.

To overcome this attack, many researchers propose content-dependent watermarking (CDWM). The basic idea is to multiply $\mathbf{w}$ with the encrypted version of the content hash,[3] i.e.,

$$\mathbf{y}^i = \mathbf{x}^i + \mathbf{w} \cdot \mathcal{E}(\mathbf{h}^i) \;\; \text{for} \;\; i = 1, 2, \dots, N \qquad (3)$$

where $\mathbf{h}^i = \mathcal{H}(\mathbf{x}^i)$. $\mathcal{H}(\cdot) : \mathbb{Z}^K \to \{\pm 1\}^K$ is the content hash function, and $\mathcal{E}(\cdot) : \{\pm 1\}^K \to \{\pm 1\}^K$ is the encryption function. The output of $\mathcal{H}(\cdot)$ is a string of bits representing the important features of the content. For watermark detection, when the original signal is unknown at the detector, an approximate version of the content hash $\bar{\mathbf{h}}^i$ is reconstructed from $\mathbf{y}^i$ using $\bar{\mathbf{h}}^i = \mathcal{H}(\mathbf{y}^i)$, and we can detect the watermark by $\mathcal{R}(\mathbf{y}^i, \mathbf{w} \cdot \mathcal{E}(\bar{\mathbf{h}}^i))$. To facilitate this detection, the encryption function must satisfy the requirement of distance-preserving. That is, $D(\mathbf{a}, \mathbf{b}) = D(\mathcal{E}(\mathbf{a}), \mathcal{E}(\mathbf{b}))$, where $D(\cdot, \cdot)$ is any distance metric. Then there would be a strong correlation between $\mathbf{y}^i = \mathbf{x}^i + \mathbf{w} \cdot \mathcal{E}(\mathbf{h}^i)$ and $\mathbf{w} \cdot \mathcal{E}(\bar{\mathbf{h}}^i)$ if $\bar{\mathbf{h}}^i$ is close to $\mathbf{h}^i$, and therefore we can detect the watermark by correlation. Lu et al. propose the use of *permutation-only* encryption $\mathcal{P}(\cdot)$ to achieve this property [4]. However, encryption by permutation is not a very secure approach by nature and we show how a proposed cryptanalysis can break the scheme in $\mathcal{O}(K^2)$ computations.

In this paper, we present a novel approach to construct secure CDWM with the aid of homomorphic encryption and dirty paper precoding. The intuitive idea is to introduce a decryption module before watermark detection such that only those who know how to decrypt is able to detect the watermark. This paper is organized as follows. Firstly we briefly sketch the cryptanalysis of Lu's CDWM scheme in Section 2. We then present the proposed secure CDWM scheme in Section 3. We conclude this paper by bringing up some thoughts on the integration of watermarking and cryptography.

Notably the generic CDWM formulation in this paper can also be applied to *watermarking-based content authentication*.

[1] A vector is denoted in boldface, e.g., $\mathbf{a} = (a(1), a(2), a(3), \dots)$. An element of $\mathbf{a}$ is denoted by $a(i)$.

[2] Correlation coefficient of two vectors $\mathcal{R}(\mathbf{a}, \mathbf{b}) = \frac{<\mathbf{a}-m_a, \mathbf{b}-m_b>}{\sqrt{<\mathbf{a}-m_a, \mathbf{a}-m_a> \cdot <\mathbf{b}-m_b, \mathbf{b}-m_b>}}$, where m_a is the mean of $\mathbf{a}$, and $<\mathbf{a}, \mathbf{b}> = \sum_i a(i)b(i)$ is the inner product of $\mathbf{a}$ and $\mathbf{b}$.

[3] Without special notice, the operations on vectors are all element-by-element. For example, $\mathbf{a} \cdot \mathbf{b} = (a(1)b(1), a(2)b(2), a(3)b(3), \dots)$.

1-4244-1017-7/07/$25.00 ©2007 IEEE

In this scenario, the embedded content hash is used to verify the integrity of content. We can simply ignore the watermark term $\mathbf{w}$ and treat $\mathcal{E}(\mathbf{h}^i)$ directly as the embedding payload. Therefore, the analyses and proposals in this paper can also be applied to the content authentication scenario.

2. CRYPTANALYSIS OF LU'S SCHEME

Our cryptanalysis rests on two facts. *1*) The permutation-only encryption does not hide the statistics of the plaintext. *2*) It does not introduce dependency between bits and thus makes the encryption much weaker. Assume we have a large pool of images (hence enough $\mathbf{y}^i$ blocks). We start by identifying from all the $\mathbf{y}^i$ blocks those whose content hash $\bar{\mathbf{h}}^i$ have 1's exceeding a threshold τ (say, $\tau \geq 95\%$). We can similarly identify those blocks of -1's. For these blocks, we definitely know that their encrypted content hash $\mathcal{E}(\mathbf{h}^i)$ also have excessive number of 1's (or -1's). We can recover the watermark $\mathbf{w}$ based on these blocks similarly as in Eq. (2). Once we find $\mathbf{w}$, we seek to recover $\mathcal{P}(\cdot)$ by search. Apparently the full search space is $K!$. However, since for $\mathcal{P}(\cdot)$, each bit is independent of each other, the search space can be reduced to the order of $\mathcal{O}(K^2)$.

An intuitive improvement of Lu's scheme is to complement some bits to hide the statistics, besides permutation. Our analysis shows that the necessary searches can now be improved to $\mathcal{O}(2^{0.12K} \cdot K^2)$, but still a weak level of security.

3. A SECURE CDWM IMPLEMENTING HOMOMORPHIC ENCRYPTION AND DIRTY PAPER PRECODING

From the analysis above, we see that more secure encryption schemes are needed for CDWM. In the meantime, these encryption schemes must be able to tolerate the errors introduced by the content hash functions. In this section, we consider a novel approach utilizing homomorphic encryption and dirty paper precoding.

3.1. The Rationale

In [5], Cox et al. introduced a layered approach to the design of secure watermarking systems (Fig. 1 (a)). They pointed out that the watermarking layer is "almost always the weakest link" with the lowest level of security. We notice that the reasons accounting for the weak security of watermarking are manifold, and one of the most prominent is the dependence on the *linear operations* inherent from the traditional signal processing techniques (recall the computation of the correlation coefficient, and also the computations in Eq. (2)). One possible countermeasure against such problems is to try to *create some nonlinearity in the watermarking layer* such that linear operations are no longer applicable. We notice that cryptography serves as a very good candidate here. Consider reversing the order of watermark detection and decryption, such that watermark detection is only accessible to those who know how to decrypt. That is, we make sure there is no apparent correlation between the watermark and the watermarked host signal $\mathbf{y}$, and therefore for an adversary knowing $\mathbf{y}$, there is no opportunity to exploit the weakness of encryption and thereby recover the watermark. However, for some legitimate user who wants to detect the watermark, she can perform decryption on $\mathbf{y}$, and correlate with the watermark in the plaintext domain. Note that the encryption must be on the watermark signal to be embedded while the decryption must be on the watermarked host signal. This can be achieved through carefully applying the property of homomorphic encryption.

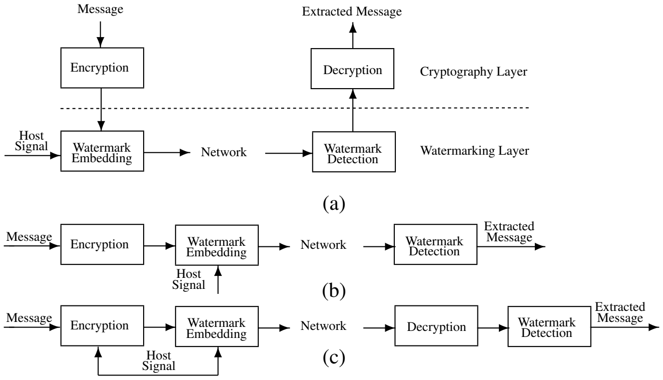

Fig. 1. (a) Layered architecture for secure watermarking design proposed by Cox et al. [5]. (b) Framework of Lu et al.'s CDWM based on distance-preserving encryption [4]. (c) The proposed secure CDWM framework.

For clarity, we have summarized the structural differences between Cox et al.'s layered architecture for watermark design [5], Lu et al.'s distance-preserving design of CDWM [4], and our proposed secure CDWM scheme in Fig. 1.

3.2. The Proposed Scheme

We use some notations slightly different from the previous sections. First we split the ith host signal $\mathbf{x}^i$ into M subvectors, i.e., $\mathbf{x}^i = (\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_M^i)$, each subvector is of length $k = \lceil \log_2 n \rceil$ where n is associated with the encryption function introduced later. Therefore, we have the length of the vector $\mathbf{x}^i$ equal to $K = M \cdot k = M \cdot \lceil \log_2 n \rceil$. From now on, we drop the superscript of $\mathbf{x}^i$ for convenience and denote by $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ the host signal of interest.

For watermark embedding, let

$$\mathbf{y}_i = \mathbf{x}_i + \mathcal{S}\left(\mathcal{E}\left(p_i(\mathbf{w}, \mathbf{x}_i)\right)\right) \quad \text{for} \quad i = 1, 2, \ldots, M \tag{4}$$

where p_i is the plaintext as a function of the watermark sequence $\mathbf{w}$ and the host signal $\mathbf{x}_i$. p_i can be expressed as (for the detail of Eq. (5), refer to Subsection 3.3):

$$p_i = (\mathcal{C}(\mathbf{w} \cdot \mathbf{h}_i) - \mathcal{D}(\mathcal{C}(\mathbf{x}_i))) \bmod n \tag{5}$$

where $\mathbf{h}_i = \mathcal{H}(\mathbf{x}_i)$. Both $\mathbf{w}$ and $\mathbf{h}_i$ are in bipolar form. Different from Lu's scheme, here we put the watermark $\mathbf{w}$ inside the encryption function. The homomorphic encryption function $\mathcal{E}(\cdot) : \mathbb{Z}_n \to \mathbb{Z}_n$ is the mapping from plaintext p_i to ciphertext c_i, and correspondingly, $\mathcal{D}(\cdot) : \mathbb{Z}_n \to \mathbb{Z}_n$ is the decryption function. The *energy spreading* function $\mathcal{S}(\cdot) : \mathbb{Z}_n \to \{\pm 1\}^k$ maps the ciphertext into a vector of bipolar form before it is added to the host signal. This function is to effectively spread the energy of the embedded signal to many host signal coefficients to enforce imperceptibility. Correspondingly, $\mathcal{C}(\cdot) : \mathbb{Z}^k \to \mathbb{Z}_n$ is some *energy collection* function which works in the opposite direction of $\mathcal{S}(\cdot)$. Note that $\mathcal{C}(\cdot)$ is not the inverse function of $\mathcal{S}(\cdot)$ since the mapping domains are different.

For watermark detection, first we want to recover c_i from $\mathbf{y}_i$ using $\mathcal{C}(\cdot)$, such that $y_i = \mathcal{C}(\mathbf{y}_i) = x_i + c_i$, where $x_i = \mathcal{C}(\mathbf{x}_i)$ is a term representing the interference of the host signal $\mathbf{x}_i$. We require the homomorphic encryption function to have two properties. *1*) The mapping between ciphertext and plaintext is bijective, such that $x = \mathcal{D}(\mathcal{E}(x)) = \mathcal{E}(\mathcal{D}(x))$. *2*) There exists additive homomorphism $\mathcal{E}((a+b) \bmod n) = (\mathcal{E}(a) + \mathcal{E}(b)) \bmod n$, or equivalently, in the decryption form,

$$\mathcal{D}((a+b) \bmod n) = (\mathcal{D}(a) + \mathcal{D}(b)) \bmod n. \tag{6}$$

In [6], Rivest et al. proposed a cryptosystem based on the Chinese Remainder Theorem which satisfies the two requirements above. Notably in [7], Brickell and Yacobi proposed an attack on [6]. However, we notice that this attack relies on the known ciphertext, which is not directly accessible to the attacker in our CDWM scheme. Although the security of this cryptosystem needs further investigation, here we use it only to demonstrate our main idea of using cryptography to enhance the security of watermarking systems.

If we perform decryption on y_i, these properties yield to the following derivation.

$$\begin{aligned}\mathcal{D}(y_i \bmod n) &= \mathcal{D}\left((x_i + c_i) \bmod n\right) \\ &= \mathcal{D}\left((x_i + \mathcal{E}(p_i)) \bmod n\right) = (\mathcal{D}(x_i) + p_i) \bmod n.\end{aligned} \tag{7}$$

In the plaintext, we first reconstruct $\bar{p}_i = \mathcal{C}(\mathbf{w} \cdot \bar{\mathbf{h}}_i)$ for $i = 1, 2, \ldots, M$, where $\bar{\mathbf{h}}_i = \mathcal{H}(y_i)$. The watermark detection is: $\mathcal{R}((\mathcal{D}(y_1 \bmod n), \ldots, \mathcal{D}(y_M \bmod n)), (\bar{p}_1, \ldots, \bar{p}_M))$.

3.3. Dirty Paper Precoding

To reduce the influence of the modulo operation (Eq. (7)) on the watermark detectibility, we seek to compensate the interference of $\mathcal{D}(x_i)$ such that $(\mathcal{D}(x_i) + p_i)$ does not fall outside $[0, n)$. We realize this problem is closely related to the Tomlinson-Harashima (T-H) implementation [8, 9] of the dirty paper coding [10]. In this problem, $\mathcal{C}(\mathbf{w} \cdot \mathbf{h})$ can be seen as the message to convey and $\mathcal{D}(x_i)$ is the known interference. The T-H precoding involves the subtraction of $\mathcal{D}(x_i)$ from $\mathcal{C}(\mathbf{w} \cdot \mathbf{h})$, plus a modulo operation (to constrain the power), as described by Eq. (5). From Eq. (5) and Eq. (7), we have:

$$\begin{aligned}\mathcal{D}(y_i) &= [\mathcal{D}(x_i) + [\mathcal{C}(\mathbf{w} \cdot \mathbf{h}) - \mathcal{D}(x_i)] \bmod n] \bmod n \\ &= [\mathcal{D}(x_i) + \mathcal{C}(\mathbf{w} \cdot \mathbf{h}) - \mathcal{D}(x_i) + jn] \bmod n \\ &= [\mathcal{C}(\mathbf{w} \cdot \mathbf{h}) + jn] \bmod n \\ &= \mathcal{C}(\mathbf{w} \cdot \mathbf{h})\end{aligned} \tag{8}$$

where j is an arbitrary integer. The last equation holds since $\mathcal{C}(\mathbf{w} \cdot \mathbf{h}) \in [0, n)$. That is, we can completely remove the interference of $\mathcal{D}(x_i)$ by precoding the message while satisfying the power constraint. We also notice that this precoding approach is well in-line with the communication with side information guideline of watermarking design [11].

3.4. Verification

We verify the performance of the proposed CDWM scheme in terms of fidelity, robustness, payload and effective keyspace.

Perhaps the most surprising fact is that the improvement of fidelity does not affect the payload and robustness of watermark. A good measure of the fidelity is the signal-to-watermark ratio (SWR). In our simple watermarking model, we keep the watermark magnitude constant and let the magnitude of x_i increase when SWR increases. However, due to decryption (particularly the modulo operation), the interference term $\mathcal{D}(x_i)$ is always bounded within $[0, n)$. By further applying the dirty paper precoding, $\mathcal{D}(x_i)$ can be compensated completely. Intuitively, our result is in accordance with Costa's prediction of the dirty paper channel capacity in [10].

The robustness is measured by how much noise the system can accommodate for a given level of detection error. The detection error can be evaluated in terms of false negative rate (FNR) and false positive rate (FPR). We assume a noise term n_i (after energy collection) is added to y_i during transmission. Decryption of the received signal leads to: $\mathcal{D}(y_i + n_i) = [\mathcal{C}(\mathbf{w} \cdot \mathbf{h}_i) + \mathcal{D}(n_i)] \bmod n$. After detection, even if n_i is small, $\mathcal{D}(n_i)$ would be amplified to any value between $[0, n)$. Therefore, we seek to eliminate the noise term before energy collection, instead of after. This can be achieved by embedding each bit of $\mathcal{S}\left(\mathcal{E}\left(p_i(\mathbf{w}, \mathbf{x}_i)\right)\right)$ into multiple (L) host signal coefficients in the embedding and later taking the average of the L coefficients in the detection. The performance can be analyzed as follows. If we assume that the watermark signal is not present, we can easily verify that FPR $= 1/n$ (assuming $M = 1$). When the watermark is present, we assume that the noise $\mathbf{n}_i$ added to $\mathbf{y}_i$ is white Gaussian with $n_i(j) \sim \mathcal{N}(0, \sigma_n^2)$. The FNR can be expressed in

$$P_e = 1 - \Pr(n_i = 0) = 1 - \left(1 - \operatorname{erfc}\left(\sqrt{\frac{L}{2\sigma_n^2}}\right)\right)^{\lceil \log_2 n \rceil}. \tag{9}$$

In Fig. 2, we plot the results and compare it with an established non-CDWM scheme – improved spread spectrum (ISS) [12]. The results show that the proposed CDWM scheme actually performs better than ISS at high SNR region. However, when the noise variance increases, the robustness drops quickly. Therefore, our system does not have good behavior at low SNR region. However, given that the additive noise is also bounded by the fidelity constraint (i.e., the quality of the attacked image still needs to be maintained), this performance can be justified. We will further address the robustness problem in our future work.

The payload of watermark is measured by the number of bits embedded per coefficient, or $R = 1/K = 1/(M \cdot L \cdot \lceil \log_2 n \rceil)$. Note that n is related to the encryption key. Therefore, the payload is related to the encryption key (p, q) used. This is rather an undesirable feature of the proposed scheme.

We assume that the underlying encryption scheme is secure. For an adversary with no knowledge of the decryption key, the best possible cryptanalysis is brute force search. The search space would be the watermark keyspace times the encryption keyspace.

4. CONCLUDING REMARKS

There have been debates in the watermarking community on the analogy between watermarking and cryptography and the necessity for integration of them for system design. In

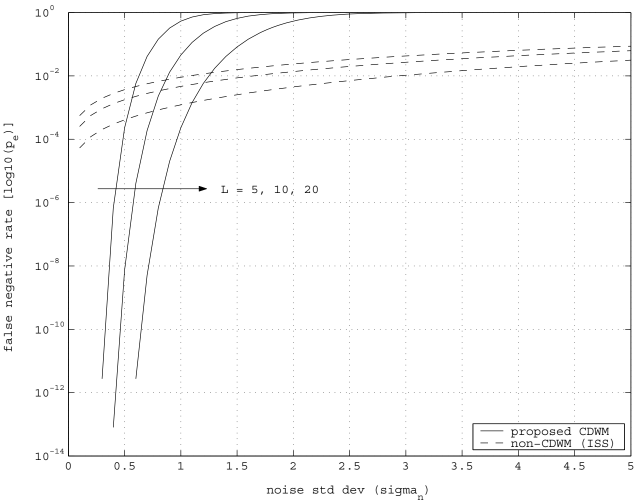

Fig. 2. FNR as a function of σ_n for the proposed CDWM scheme and a non-CDWM scheme (ISS [12]). For fair comparison, we have maintained the same fidelity and payload. The host signal is assumed i.i.d. Gaussian with $\sigma_x = 10$. We set $\lceil \log_2 n \rceil = 30$. For ISS, we set the decision threshold T at where we can mimimize the sum of FPR and FNR.

[5], Cox et al. pointed out that watermarking is fundamentally a communication problem which concerns with the reliable delivery of message over an unreliable channel, while cryptography concerns with the security of such delivery. At this level, they maintained that a layered architecture which distinguishes the roles of watermarking and cryptography is more appropriate. Based on our results, we would like to argue that besides the consideration of watermarking from a standard communication point of view, the fact that watermarking is a unique and independent problem should not be ignored. For example, besides reliable embedding and detection of watermark, one also needs to consider the potential estimation attacks from the adversary. In this paper, our example indeed shows that it would help prevent such attacks by incorporating some proper cryptographic techniques into the watermarking scheme. Besides, in [13], Kalker also showed the evidence of improving watermarking security by using homomorphic cryptography to facilitate a distributed watermark detection and centralized decryption. We also notice that by integrating watermarking with cryptography we would improve the security of the cryptographic schemes themselves (consider in our example, the ciphertext is hidden from the adversary).

In summary, the authors' point of view is that when watermarking and cryptography are designed in an integrated manner, we can achieve synergy of the two. Of course, joint consideration of the two would increase the complexity and potentially raise the risk of inappropriate designs. Therefore, careful verification is always needed to avoid design pitfalls. We hope that this paper would serve the idea-provoking purpose to motivate the use of homomorphic encryption in the watermarking research, and also that it would promote the efforts in the cryptography community on designing cryptosystems friendly to signal processing applications.

Our future works include further improving the robustness of the proposed scheme and applying it to images.

5. REFERENCES

[1] P. Bas, J. M. Chassery, and B. Macq, "Geometrically invariant watermarking using feature points," *IEEE Trans. on Image Processing*, vol. 11, no. 9, pp. 1014 – 1028, 2002.

[2] C.W. Tang and H.M. Hang, "A feature-based robust digital watermarking scheme," *IEEE Trans. on Signal Processing*, vol. 51, no. 4, pp. 950 – 959, 2003.

[3] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit digital watermarking robust against local nonlinear geometrical distortions," in *Proc. IEEE Int. Conf. on Image Processing*, 2001, pp. 999 – 1002.

[4] C.-S. Lu and C.-Y. Hsu, "Content-dependent antidisclosure image watermark," in *Proc. IWDW 2003, LNCS 2939*, 2004, pp. 61 – 76.

[5] I.J. Cox, G. Doerr, and T. Furon, "Watermarking is not cryptography," in *Proc. IWDW 2006, LNCS 4283*, 2006, pp. 1 – 15.

[6] R.L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," *R.A. DeMillo et al. eds., Foundations of Secure Computation (Academic Press, New York, 1978)*, pp. 169 – 179.

[7] E.F. Brickell and Y. Yacobi, "On privacy homomorphisms," *D. Chaum et al. eds., Advances in Cryptology-Eurocrypt'87 (Springer, Berlin, 1988)*, pp. 117 – 125.

[8] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, pp. 138 – 139, Mar. 1971.

[9] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE trans. on Commun.*, vol. 20, pp. 774–780, Aug. 1972.

[10] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. IT - 29, no. 3, Aug. 1983.

[11] I.J. Cox, M.L. Miller, and A. McKellips, "Watermarking as communications with side information," *Proceedings of the IEEE*, July 1999.

[12] H. S. Malvar and D. A. F. Florencio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 898 – 905, Apr. 2003.

[13] T. Kalker, "Secure watermark detection," in *Proc. 39th Allerton Conf. on Comm., Contr., and Comp.*, 2005.