SENSITIVITY ANALYSIS ATTACKS AGAINST RANDOMIZED DETECTORS

Maha El Choubassi and Pierre Moulin

University of Illinois at Urbana Champaign
Beckman Inst., Coord. Sci. Lab & ECE Dept.
Emails: *cel@ifp.uiuc.edu* and *moulin@ifp.uiuc.edu*

ABSTRACT

Sensitivity analysis attacks present a serious threat to the security of popular spread spectrum watermarking schemes. Randomization of the detector is thought to increase the immunity of such schemes against cryptanalysis. In this paper, we introduce a new attack against randomized detectors. This attack is successful, which implies that spread spectrum schemes still lack security.

Index Terms: Watermarking, security, sensitivity analysis attacks, spread spectrum, randomized detectors.

1. INTRODUCTION

Information hiding is about the imperceptible embedding of information inside host data such as image signals. The application we focus on is copyright protection of digital media. In this setup, the original host signal[1] $\mathbf{s}$ is either left unchanged (unwatermarked), or a watermark signal $\mathbf{w}$ is embedded into $\mathbf{s}$, resulting in the watermarked signal $\mathbf{x} = \mathbf{s} + \mathbf{w}$. That is additive spread spectrum embedding, illustrated in Figure 1. The watermark $\mathbf{w}$ is shared between the embedder and the detector. Once a signal $\mathbf{y}$ is input to the detector, the detection function $t(\mathbf{y}, \mathbf{w})$ is evaluated and compared to a detection threshold τ to decide whether $\mathbf{y}$ is watermarked or not. If the detection function is deterministic, the set

$$\{\mathbf{y} \in \mathbb{R}^n : t(\mathbf{y}, \mathbf{w}) = \tau\}$$

is called the detection boundary for the detector $t(\cdot, \mathbf{w})$.

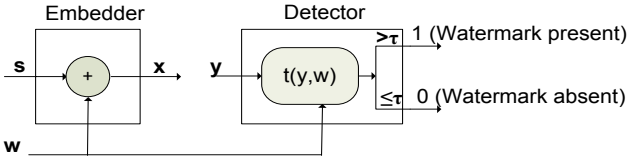

Figure 1: The watermark embedder and the watermark detector.

The results of [1]-[9] prove that spread spectrum techniques are vulnerable to sensitivity analysis attacks. In such attacks, the attacker has access to a watermark detector and a watermarked signal $\mathbf{x}$. He systematically changes $\mathbf{x}$ into auxiliary signals and inputs them to the detector. Through the leaked information about the watermark, the attacker obtains an estimate of $\mathbf{w}$. He subsequently removes it from $\mathbf{x}$ to produce the pirated copy $\hat{\mathbf{s}}$. The requirements on $\hat{\mathbf{s}}$ are to be perceptually similar to the original host signal $\mathbf{s}$ and to trigger a negative detector's response.

For a wide class of detectors, all deterministic, the number of detection operations needed by the attacker to estimate the watermark is generally linear in the size of the signal [6, 7, 8]. Since the watermark detector is the source of leakage of information about the watermark, we sought randomized detectors to improve the security of spread spectrum schemes in [10]. Linnartz and van Dijk [2], and Venkatesan and Jakubowski [9] also used randomized detection to increase security. However, this is achieved at the expense of detection performance. Assuming a generalized Gaussian distribution (GGD) on the host signal, we built in [10] a framework to control the loss in the detection performance using classical detection-theoretic tools: large deviation analysis and Chernoff bounds. However, while these techniques provide a powerful tool to evaluate the performance of the detector, there is no efficient measure to quantify the vulnerability of a watermarking scheme against sensitivity analysis attacks. In this paper, we construct attack algorithms against randomized detectors. The attacks aim at generating auxiliary signals on the *average* detection boundary and reduce the randomized detection scheme into an equivalent deterministic one. Finally, the attack algorithms built in [6, 7, 8] against deterministic watermark detectors are applied to the equivalent detector in order to estimate and then remove the watermark.

The paper is organized as follows. In Sec. 2, we describe the attack algorithm against randomized detectors. In the next three Secs. 3, 4, and 5 respectively, the approach used to average the randomized boundary of randomized threshold detectors, randomized GGD detectors mixture, and subset selection detectors is explained. To verify the properties of our new attack, we provide experimental results in Sec. 6. Finally, we conclude in Sec. 7.

2. ATTACK ALGORITHM

We consider several families of randomized detectors. The analysis is made for given watermark $\mathbf{w}$ and signal $\mathbf{y}$ and the randomness is due to one or several random parameters denoted as Θ and drawn by the detector from a probability distribution p_Θ. The watermark $\mathbf{w} \in \mathbb{R}^n$ is fixed and the randomized detection statistic for detector's input $\mathbf{y}$ is $T_\Theta(\mathbf{y}, \mathbf{w})$, to be compared against the threshold τ. We define the $p-$boundary $\mathbf{B}_p$ as the set of signals $\mathbf{y}$ characterized by

$$\mathbb{P}\left(T_\Theta(\mathbf{y}, \mathbf{w}) > \tau | \mathbf{y}, \mathbf{w}\right) = p, \tag{1}$$

where $\mathbb{P}(\mathcal{E})$ denotes the probability of an event $\mathcal{E}$. When a signal $\mathbf{y}$ on the $p-$boundary is input to the detector N times, the expected

WORK SUPPORTED BY NSF UNDER GRANT CCR 03 − 25924.

[1]Unless otherwise stated, all the signals we consider are vectors in $n-$dimensional Euclidean space.

number of positive responses is Np.

In the subsequent sections, we consider three families of randomized detectors: randomized threshold detectors [2], randomized GGD detectors [10], and subset selection detectors [9]. The $p-$boundary for each family is the same as the detection boundary of an equivalent deterministic detector. Recall from [6, 7, 8] that sensitivity analysis attacks are possible due to measurements taken on the detection boundary by the attacker. Unlike the deterministic detectors considered in [6, 7, 8], the detectors studied in this paper are randomized. In this case, the attacker generates signals on the $p-$boundary. These signals satisfy (1) with a controllable (and arbitrarily large) degree of accuracy. At the same time, these signals belong to the detection boundary of the equivalent deterministic detector. Hence, the attack algorithms of [6, 7, 8] can be used. Once an estimate $\hat{\mathbf{w}}$ of the watermark is obtained, it is subtracted from the watermarked signal $\mathbf{x}$ resulting in the pirated copy $\hat{\mathbf{s}}$.

2.1. Generating Points on the $p-$boundary

For a signal $\mathbf{z}$ and a direction $\mathbf{d}$, in order to find a scalar α such that the signal $\mathbf{y} = \mathbf{z} + \alpha\mathbf{d}$ is exactly on the $p-$boundary, an infinite number of detection probes would be needed. Since this is not possible, the attacker sets a finite number N_{probes} of detection probes to be used. For a fixed p, $\mathbf{y}$, $\mathbf{w}$, and a given randomized detection function, we define the binary function

$$d(\mathbf{y}, \mathbf{w}, N_{probes}) = \begin{cases} 1, & \text{if } T_\Theta(\mathbf{y}, \mathbf{w}) > \tau \text{ for more} \\ & \text{than } pN_{probes} \text{ times} \\ 0, & \text{else.} \end{cases}$$

The attacker can use any search algorithm, in particular the binary search algorithm, and use this function in the search queries to find signals that are approximately on the $p-$boundary. As $N_{probes} \to \infty$, these signals approach the $p-$boundary.

3. RANDOMIZED THRESHOLD

Let us consider a deterministic detection test $t(\mathbf{y}, \mathbf{w})$, for example the correlation detection test or the GGD test. As in [2], let the detection threshold be a real-valued random variable Θ distributed according to $p_\Theta(\theta)$. Therefore, the detector is viewed as a randomized detection statistic with zero threshold:

$$T_\Theta(\mathbf{y}, \mathbf{w}) = t(\mathbf{y}, \mathbf{w}) - \Theta.$$

Due to (1), signals $\mathbf{y}$ on $\mathbf{B}_p$ are characterized by

$$\mathbb{P}\left(\Theta < t(\mathbf{y}, \mathbf{w}) | \mathbf{y}, \mathbf{w}\right) = p. \tag{2}$$

For example, when Θ is a uniform random variable over an interval $[a, b]$, the $p-$boundary is given for $0 < p < 1$ as

$$\mathbf{B}_p = \{\mathbf{y} : t(\mathbf{y}, \mathbf{w}) = p(b-a) + a\}$$

In this case, $\mathbf{B}_p$ is the detection boundary of an equivalent deterministic detector

$$T_{eq}(\mathbf{y}, \mathbf{w}) = t(\mathbf{y}, \mathbf{w}),$$

with threshold $p(b-a) + a$. Therefore, any of the attack algorithms developed against deterministic detectors $T_{eq}(\mathbf{y}, \mathbf{w})$ can be applied to estimate the watermark.

4. RANDOMIZED GGD DETECTORS MIXTURE

In Sec. 3, a single parameter Θ was randomized. In [10] we studied a class of detectors with extremely large randomization space. Each time the detector is probed, the support $\{1, 2, \ldots, n\}$ of the signal is partitioned into K random subsets (Figure 2a). GGD parameters, μ_u and α_u, are assigned to each subset $u \in \{1, \ldots, K\}$. Let $\{U_1, U_2, \ldots, U_n\}$ be n independent and identically distributed (iid) random variables with alphabet $\mathcal{U} = \{1, 2, \ldots, K\}$ and probability mass function (pmf)

$$p(u) = \lambda_u, \quad \text{if } u \in \{1, 2, \ldots, K\}. \tag{3}$$

For each pixel $j \in \{1, 2, \ldots, n\}$, U_j indicates to which subset this pixel belongs. The probability that a pixel is in subset u is λ_u, and $\sum_{u=1}^K \lambda_u$ is equal to one.

In this case, $\Theta = \{U_1, U_2, \ldots, U_n\}$ and the randomized detection function is given by

$$\begin{aligned} T_\Theta(\mathbf{y}, \mathbf{w}) &= \sum_{j=1}^n V_j, \\ \text{where } V_j &= \left|\frac{y_j}{\alpha_{U_j}}\right|^{\mu_{U_j}} - \left|\frac{y_j - w_j}{\alpha_{U_j}}\right|^{\mu_{U_j}}. \end{aligned} \tag{4}$$

The $p-$boundary is the set of signals $\mathbf{y}$ that satisfy

$$\mathbb{P}\left(\sum_{j=1}^n V_j > \tau | \mathbf{y}, \mathbf{w}\right) = p. \tag{5}$$

For this purpose, we need to know the distribution of the sum of random variables $\sum_{j=1}^n V_j$. First, due to the independence of $\{U_1, U_2, \ldots, U_n\}$, $\{V_1, V_2, \ldots, V_n\}$ in (4) are also independent. The expected value and the variance of V_j conditioned on y_j and w_j are given by

$$\begin{aligned} \mathbb{E}[V_j | y_j, w_j] &= \sum_{u=1}^K \lambda_u \left(\left|\frac{y_j}{\alpha_u}\right|^{\mu_u} - \left|\frac{y_j - w_j}{\alpha_u}\right|^{\mu_u} \right), \\ \mathbb{V}ar[V_j | y_j, w_j] &= \sum_{u=1}^K \lambda_u \left(\left|\frac{y_j}{\alpha_u}\right|^{\mu_u} - \left|\frac{y_j - w_j}{\alpha_u}\right|^{\mu_u} \right)^2 \\ &\quad - \mathbb{E}[V_j | y_j, w_j]^2 \end{aligned} \tag{6}$$

Checking the conditions of Lindeberg's generalized central limit theorem [11] (CLT), we conclude that the normalized random variable

$$Z_n = \frac{\sum_{j=1}^n \left(V_j - \mathbb{E}[V_j | y_j, w_j]\right)}{\sqrt{\sum_{j=1}^m \mathbb{V}ar[V_j | y_j, w_j]}} \tag{7}$$

converges in distribution to a Gaussian random variable with mean 0 and variance 1, as $n \to \infty$. Therefore, the $p-$boundary in (5) is characterized by

$$\begin{aligned} p &= \mathbb{P}\left(Z_n > \frac{\tau - \sum_{j=1}^n \mathbb{E}[V_j | y_j, w_j]}{\sqrt{\sum_{j=1}^n \mathbb{V}ar[V_j | y_j, w_j]}} \middle| \mathbf{y}, \mathbf{w} \right) \\ &\sim Q\left(\frac{\tau - \sum_{j=1}^n \mathbb{E}[V_j | y_j, w_j]}{\sqrt{\sum_{j=1}^n \mathbb{V}ar[V_j | y_j, w_j]}} \right), \quad \text{as } n \to \infty, \end{aligned}$$

for any fixed value of p. Therefore, we approximate the $p-$boundary for the randomized detector by

$$\hat{\mathbf{B}}_p = \left\{ \mathbf{y} \in \mathbb{R}^n : \sum_{j=1}^{n} \mathbb{E}[V_j|y_j, w_j] + Q^{-1}(p) \sqrt{\sum_{j=1}^{n} \mathbb{V}ar[V_j|y_j, w_j]} = \tau \right\}. \quad (8)$$

Note that $\mathbb{E}[V_j|y_j, w_j]$ and $\mathbb{V}ar[V_j|y_j, w_j]$ are functions of y_j and w_j. As we can see, estimating the $p-$boundary reduces the randomized detector into a deterministic one with detection function

$$T_{eq}(\mathbf{y}, \mathbf{w}) = \sum_{j=1}^{n} \mathbb{E}[V_j|y_j, w_j] + Q^{-1}(p) \sqrt{\sum_{j=1}^{n} \mathbb{V}ar[V_j|y_j, w_j]} \quad (9)$$

and same threshold τ as the randomized detector.

The equivalent detection function is highly nonlinear in $\mathbf{y}$. With the choice of $p = 0.5$, $Q^{-1}(p)$ becomes zero and (9) simplifies into

$$T_{eq}(\mathbf{y}, \mathbf{w}) = \sum_{u=1}^{K} \lambda_u \sum_{j=1}^{n} \left| \frac{y_j}{\alpha_u} \right|^{\mu_u} - \left| \frac{y_j - w_j}{\alpha_u} \right|^{\mu_u}. \quad (10)$$

From (10), $T_{eq}(\mathbf{y}, \mathbf{w})$ is the expected value of the individual detection functions $\sum_{j=1}^{n} \left| \frac{y_j}{\alpha_U} \right|^{\mu_U} - \left| \frac{y_j - w_j}{\alpha_U} \right|^{\mu_U}$.

At this point the attacker can either tailor an attack against the equivalent detector in (10), or he can borrow any of the previously designed attacks. For instance, the attack proposed in [6, 7, 8] against the GGD detector can be used with $\mu = \sum_{u=1}^{K} \lambda_u \mu_u$.

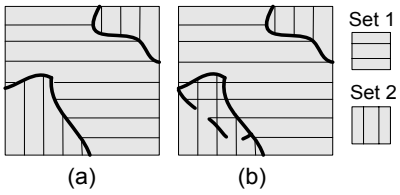

Figure 2: A partition of the image support into two sets ($K = 2$). (a) Disjoint sets. (b) Overlapping sets.

5. SUBSET SELECTION DETECTORS

In this section, we consider the randomized detector proposed by Venkatesan et al. in [9]. The scheme is based on selecting a detection function[2]

$$t(\mathbf{y}, \mathbf{w}, S) = \sum_{j \in S} t_1(y_j, w_j), \quad (11)$$

computed over a support subset $S \subseteq \{1, \ldots, n\}$. Next, the detector randomly selects *many possibly overlapping subsets* in the support of the signal, $\{1, \ldots, n\}$, and evaluates the detection statistic over each such subset according to (11), (Figure 2b). The actual detection coefficient is the median of these statistics. The authors argue that such a detector is secure against attacks that learn the watermark by introducing large changes to the value of the signal at one component. We show that the scheme is still breakable using sensitivity attacks. With the help of the concept of $p-$boundary, we derive the equivalent deterministic detector which can be attacked by already existing algorithms. Let K be the number of subsets selected and $\mathbf{M}_k \in \{0, 1\}^n$ be the mask corresponding to subset $k \in \{1, \ldots, K\}$. That is, $M_{k,j} = 1$ indicates that the j^{th} component of the signal belongs to S_k. Without loss of generality, assume that K is odd. In our model, $\mathbf{M}_k$ is a sequence of n independent and identically distributed (iid) Bernoulli random variables with probability ρ:

$$M_{k,j} = \begin{cases} 1, & \text{with probability } \rho \\ 0, & \text{with probability } 1 - \rho, \end{cases}$$

with $j \in \{1, 2, \ldots, n\}$. The masks $\mathbf{M}_k$ are mutually independent. The k^{th} statistic is defined as

$$T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) = \sum_{j=1}^{n} M_{k,j} t_1(y_j, w_j).$$

In this setting, the randomization parameters are $\Theta = \{\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_K\}$ and the resulting randomized function is given by

$$T_\Theta(\mathbf{y}, \mathbf{w}) = \text{median}_{1 \le k \le K} \left\{ T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) \right\}.$$

The $p-$boundary for this detector is characterized by

$$\begin{aligned} \mathbb{P}\left(\text{median}_{1 \le k \le K} T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) > \tau | \mathbf{y}, \mathbf{w}\right) &= p \\ \mathbb{P}\left(\text{at least } (K+1)/2 \text{ of the statistics }, \right. & \\ \left. T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) > \tau | \mathbf{y}, \mathbf{w}\right) &= p. \quad (12) \end{aligned}$$

Define p' as

$$p' \triangleq \mathbb{P}\left(T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) > \tau | \mathbf{y}, \mathbf{w}\right).$$

Note that p' is independent of k since $\mathbf{M}_k$ are iid. In this case, the binary random variables

$$R_k = \begin{cases} 1, & \text{if } T(\mathbf{y}, \mathbf{w}, \mathbf{M}_k) > \tau, \\ 0, & \text{else.} \end{cases}$$

are iid Bernoulli with probability p', i.e., $R_k \sim Be(p')$. Therefore, (12) becomes

$$\begin{aligned} p &= \mathbb{P}\left(\sum_{k=1}^{K} R_k \ge \frac{K+1}{2} \middle| \mathbf{y}, \mathbf{w} \right) \\ &= \sum_{i=(K+1)/2}^{K} \binom{K}{i} p'^i (1 - p')^{K-i}. \quad (13) \end{aligned}$$

In order to describe the $p-$boundary, the probability p' has to be computed. Let $V_j = M_{k,j} t_1(y_j, w_j)$. The random variables V_j are independent but not identically distributed. We again use the generalized CLT and conclude that as n tends to ∞, the random variable Z_n defined in the same way as in (7) converges in distribution to a Gaussian random variable $N(0, 1)$. Similarly to the derivations in Sec. 4, the approximate $p-$boundary is defined as

$$\hat{\mathbf{B}}_p = \left\{ \mathbf{y} \in \mathbb{R}^n : \rho \sum_{j=1}^{n} t_1(y_j, w_j) + Q^{-1}(p') \sqrt{\sum_{j=1}^{n} t_1^2(y_j, w_j) \rho(1 - \rho)} = \tau \right\}$$

[2]The authors in [9] choose to use a correlation detector, i.e., $t_1(y_i, w_i) = y_i w_i$, but we consider a more general setting.

Table 1: $N_{probes} = 5$ versus $N_{probes} = 100$. $\rho = \frac{\mathbf{w}\cdot\hat{\mathbf{w}}}{\|\mathbf{w}\|\|\hat{\mathbf{w}}\|}$ is the normalized correlation between $\mathbf{w}$ and $\hat{\mathbf{w}}$.

N_{probes}	$\rho > 0.65$ for	% of successful attacks
5	5.71% of the attacks	19.29%
100	70.71% of the attacks	53.57%

Note that the attacker selects p', and then calculates the corresponding value of p as in (13). If he selects $p' = 0.5$, then $Q^{-1}(p') = 0$, and $\mathbf{B}_p$ is the detection boundary of the equivalent deterministic detector

$$T_{eq}(\mathbf{y}, \mathbf{w}) = \rho t(\mathbf{y}, \mathbf{w}).$$

Hence, for Venkatesan et al.'s scheme, the equivalent deterministic detector is still the simple correlation detector but with its value scaled down by a constant factor ρ.

6. EXPERIMENTAL RESULTS

We consider a randomized GGD detector (see Sec. 4) with $K = 2$ and with GGD exponents $\mu_1 = 1.2$ and $\mu_2 = 1.9$. For an image of size 32×32, we generated 35 pseudorandom binary watermarks with equal energy, $\frac{1}{n}\|\mathbf{w}\|^2 = 6.75$, resulting in 35 watermarked signals. Against each such signal, we ran our sensitivity analysis attack algorithm with p set to 0.5, four times with $N_{probes} = 5$ and four other times with $N_{probes} = 100$. We consider an attack to be successful, when the pirated copy $\hat{\mathbf{s}}$ induces a negative response from the detector and with mean squared distortion $D_s = \|\mathbf{s} - \hat{\mathbf{s}}\|^2/n \leq 6.75$. As the N_{probes} increases, the auxiliary points we generate get closer to the $0.5-$boundary producing more successful attacks as seen in Table 1. For each the 35 watermarked signals, at least one of the 4 attacks with $N_{probes} = 100$ is successful, while this is true only for 16 watermarked signals when attacks with $N_{probes} = 5$ are used. Hence the security of the detector is severely compromised. As shown in Figures 3 and 4, attacks with $N_{probes} = 100$ result in larger correlation ρ, and smaller distortion D_s.

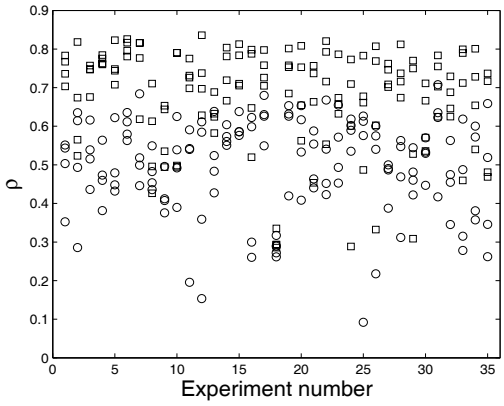

Figure 3: Normalized correlation between $\mathbf{w}$ and $\hat{\mathbf{w}}$. The squares correspond to the attacks with $N_{probes} = 100$ probes, while the circles correspond to attacks with $N_{probes} = 5$ probes.

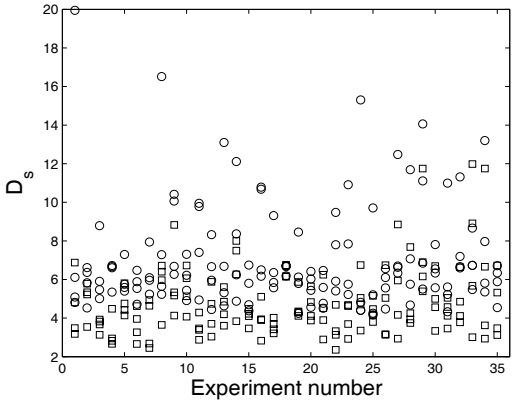

Figure 4: Mean square distortion between $\mathbf{s}$ and $\hat{\mathbf{s}}$

7. CONCLUSION

In this paper we presented a new method to launch sensitivity analysis attacks against additive spread spectrum schemes with randomized detectors. The concept of $p-$boundary is about averaging the randomized boundary of such detectors and consequently treating it as a deterministic boundary. In this case, the attacks in our previous work [6, 7, 8] are applicable. The preliminary experimental results in Sec. 6 are encouraging. In the future, we will report more extensive experiments (including larger values of N_{probes} to more precisely estimate the $p-$boundary).

8. REFERENCES

[1] I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *Proc. Int. Conference on Image Processing (ICIP)*, only CD version of proceedings available, Santa Barbara, CA, 1997.

[2] J. P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *Proc. of the Workshop of Information Hiding*, Portland, OR, 1998, pp. 258-272.

[3] T. Kalker, J. P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *Proc. ICIP*, vol. 1, pp. 425-429, Chicago, IL, 1998.

[4] A. Tewfik and M. Mansour, "LMS-based attack on watermark public detectors," in *Proc. ICIP*, Rochester, NY, 2002, pp. 649-652.

[5] P. Comesãna, L. Pérez-Freire, and F. Pérez-Gonzaléz, "The return of the sensitivity attack," in *Proc. IWDW*, Siena, Italy, 2005, pp. 260-274.

[6] M. El Choubassi and P. Moulin, "A new sensitivity analysis attack," in *Proc. SPIE*, San Jose, CA, 2005, pp. 734-745.

[7] M. El Choubassi and P. Moulin, "Noniterative algorithms for sensitivity analysis attacks," *IEEE TIFS*, vol. 2, no. 2, pp. 113-126, 2007.

[8] M. El Choubassi, "Novel algorithms for sensitivity analysis attacks," MS thesis UIUC, IL, 2005. Available from www.ifp.uiuc.edu/~cel

[9] R. Venkatesan and M.H. Jakubowski, "Randomized Detection For Spread-Spectrum Watermarking: Defending Against Sensitivity and Other Attacks," *Proc. ICASSP*, Philadelphia PA, 2005.

[10] M. El Choubassi and P. Moulin, "On the fundamental tradeoff between watermark detection," in *Proc. SPIE*, San Jose, CA, 2006, pp. 575-586.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications.* New York: Willey, 1966.