On the Effectiveness of Meta-Detection for Countering Oracle Attacks in Watermarking

Benedetta Tondi [1], Pedro Comesaña-Alfaro [2], Fernando Pérez-González [2], Mauro Barni [1]

[1]*Department of Information Engineering and Mathematics, University of Siena, Siena, ITALY*

[2] *Signal Theory and Communications Department, E.E. Telecomunicación, University of Vigo, Vigo, SPAIN*

benedettatondi@gmail.com, pcomesan@gts.uvigo.es, fperez@gts.uvigo.es, barni@dii.unisi.it

***Abstract*—We evaluate the performance of smart metadetection as a way to combat oracle attacks in watermarking. In a recent work, we have shown that few queries are sufficient for a simple metadetector (namely, a metadetector based on the closeness of queries to the watermark detection boundary) to detect an oracle attack. A limitation of our prior analysis is the assumption that all the queries correspond to either honest users or malicious ones. In this paper, we address a more realistic scenario in which honest queries are interspersed with queries derived from an oracle attack. By focusing on this more general situation, we evaluate the performance of the metadetection and derive conditions under which powerful testing is possible.**

I. Introduction

Originally proposed in the 1990s by Cox and Linnartz for attacking the correlation-based detector of Additive Spread Spectrum (Add-SS) [1] (i.e., detection regions whose boundary is a hyperplane), the so-called *sensitivity attack* is the first of a number of adversarial strategies that try to exploit the information obtained by querying the detector, and that receive the general name of *oracle attacks*. The sophistication and power of those attacks have significantly improved over time [2], [3], [4], until the inception of blind algorithms [5], [6] where no *a priori* knowledge of the decision function is even needed. Blind algorithms have successfully proven to succeed in removing the watermark for a variety of watermarking algorithms, including those used in the BOWS (Break Our Watermarking System) and BOWS-2 contests [7], [8].

The effectiveness of oracle attacks is not limited to watermarking, as they can be used against any binary detector. In multimedia security, for example, this includes forensic detectors, authentication detectors, fingerprint detectors, etc. The effectiveness and generality of oracle attacks call for the development of proper countermeasures. A first step in that direction has been recently taken in [9], where the authors propose a *metadetector* that works in parallel with a standard watermark detector. While the operation of the latter is not modified, the former is specifically devoted to detecting malicious queries. Once an oracle attack has been successfully detected, effective countermeasures, including banning, and randomized or delayed answers to queries, can be enforced. In [9], two different metadetectors are proposed; one of them exploits the fact that oracle attacks generally produce a large number of queries close to the watermark detection boundary, while the other targets the line searches typically performed by those attacks. Both strategies have been theoretically analyzed and successfully used for detecting oracle attacks with just few queries.

The analysis carried out in [9] assumes that the detector is exclusively queried by either a malicious or a honest user. However, in a practical scenario, several users may be querying the system in a time window of N queries, potentially including both honest and malicious users. Hence, the metadetector wishes to discover if there is an adversary among the users querying the system. In this paper, we address this more realistic case by generalizing the analysis made in [9]. We quantify the effectiveness of the attack by bounding the number of queries which are necessary to find a point sufficiently close to the boundary (i.e., to get close enough to the boundary). We also analyze the asymptotic performance of the metatest in order to determine under which conditions it is possible to get an asymptotically powerful test, that is, to detect an oracle attack with asymptotically zero probability of error. We also derive a critical percentage of attacking queries below which no asymptotically powerful test is possible. Such critical value is directly linked to the parameters of the watermark detector.

The paper is organized as follows: in Section II we briefly introduce the simple metadetector proposed in [9] and recap the main results proved therein; then, in Section III we analyze the performance of the metadetector in the more general setup addressed in this paper. An evaluation of the impact of the parameters of the metadetector on the performance is provided in Section IV, where the asymptotic limiting performance of the test are investigated. Experimental results are reported in Section V, and conclusions provided in Section VI.

II. Metadetectors in watermarking: a higher-level of detection.

Given a sequence under test $\mathbf{y}$ and the watermark sequence $\mathbf{w}$, a watermark detector has to decide whether the sequence $\mathbf{y}$ contains the watermark $\mathbf{w}$ (hypothesis $H_{w,1}$) or not (hypothesis $H_{w,0}$). Watermark decision splits the space of sequences into two regions: $\mathcal{R}_{w,0} = \{\mathbf{y} : l(\mathbf{y}, \mathbf{w}) \leq T\}$ and $\mathcal{R}_{w,1} = \overline{\mathcal{R}_{w,0}}$, where $l(\mathbf{y}, \mathbf{w})$ is the watermark decision function and T is the decision threshold. The false positive and false negative probabilities of the watermark detection test are denoted by $P_{F,w}$ ad $P_{M,w}$, respectively. Given a vector with N queries $\mathbf{y}^N$, the metadetector defined in [9] decides whether $\mathbf{y}^N$ is a legitimate sequence of queries ($H_{q,0}$), i.e., a sequence made by honest users, or not ($H_{q,1}$), that is, $\mathbf{y}^N$ is a sequence of queries coming from a dishonest user. In this paper, we consider the Closeness-To-the-Boundary (CTB) metadetector, which is the first metadetector introduced in [9].

A. General definition of the CTB-based Metadetector

The CTB-based metadetector relies on the definition of a narrow strip across the decision boundary $\delta\mathcal{R}_w$, namely, $\mathcal{A} = \{\mathbf{y} : T - \Delta < l(\mathbf{y}, \mathbf{w}) < T + \Delta\}$, where Δ determines the width of the strip. The assumption behind the CTB metadetector is that a dishonest user will query the detector with an unusually large number of vectors falling within $\mathcal{A}$. Given a vector with N queries $\mathbf{y}^N$, the metatest is based on the number of $\mathbf{y}_i$ that belong to $\mathcal{A}$, namely $n_y^N(\mathcal{A})$. More precisely, the test is defined by the following decision function:

$$\phi_q(\mathbf{y}^N) = \begin{cases} 0 & \text{if } n_y^N(\mathcal{A}) < \alpha \cdot N \\ 1 & \text{if } n_y^N(\mathcal{A}) \geq \alpha \cdot N, \end{cases} \tag{1}$$

where α is a fixed threshold occurrence rate. The acceptance region for the metatest is $\mathcal{R}_{q,0} = \{\mathbf{y}^N : n_y^N(\mathcal{A}) < \alpha \cdot N\}$, while its

978-1-4673-6802-5/15/$31.00 ©2015 IEEE

false alarm probability is given by $P_{F,q} = P(n_y^N(\mathcal{A}) \geq \alpha \cdot N | H_{q,0})$. Similarly, $P_{M,q} = P(n_y^N(\mathcal{A}) < \alpha \cdot N | H_{q,1})$ defines the false negative probability. For a fixed number of queries N, the metadetector sets Δ and α in order to satisfy $P_{F,q} \leq P_{F,q}^*$, for a prescribed maximum $P_{F,q}^*$.

In the following we derive the performance of the CTB detector by focusing on a particularly simple, yet common, watermarking system.

B. CTB metadetector for correlation-based watermark detectors

We assume that an Add-SS watermarking scheme is used. More specifically, we have $\mathbf{x}_w = \mathbf{x} + \gamma \mathbf{w}$, where $\mathbf{x}_w$ is the watermarked signal, γ defines the watermark strength, and the watermark sequence $\mathbf{w} \in \{-1, +1\}^L$. The ML detector for Add-SS and i.i.d. Gaussian host relies on the correlation between the sequence under test $\mathbf{y}$ and the watermark sequence $\mathbf{w}$, that is, $\rho = l(\mathbf{y}, \mathbf{w}) = \langle \mathbf{y}, \mathbf{w} \rangle$. From basic watermarking theory [10], we know that for the noiseless case $\rho \sim \mathcal{N}(\mu_{\rho|i}, \sigma_\rho^2)$, where the mean under $H_{w,0}$ is $\mu_{\rho|0} = 0$, while under $H_{w,1}$ we have $\mu_{\rho|1} = \gamma L$. The variance of ρ is $\sigma_\rho^2 = L\sigma_X^2$ under both hypotheses. In this case, $\delta \mathcal{R}_w$ is a hyperplane, as in the system considered by the original sensitivity attack [1], and then the watermark detector is easy to characterize. We assume that the error probability of the watermark detector tends to zero as L grows. For this to be possible, the threshold T must satisfy the following conditions: $\lim_{L\to\infty} T/\sqrt{L} = \infty$ and $\lim_{L\to\infty}(\gamma L - T)/\sqrt{L} = \infty$.[1] With the above system, the CTB metadetector relies on the definition of a strip of width Δ across the hyperplane $\langle \mathbf{y}, \mathbf{w} \rangle = T$, namely, $\mathcal{A} = \{\mathbf{y} : T - \Delta < \langle \mathbf{y}, \mathbf{w} \rangle < T + \Delta\}$. In order to compute $P_{F,q}$ we must define a proper model for honest queries.

Definition 1 (Model of honest queries): We consider that honest users can send two kinds of queries, corresponding to watermarked and non-watermarked signals. We model the former by $\mathcal{N}(\mathbf{w}, \sigma_X^2 \mathbf{I}_{L\times L})$, where the watermark $\mathbf{w}$ is known at the detector. On the other hand, the non-watermarked signals are assumed to follow a $\mathcal{N}(\mathbf{0}, \sigma_X^2 \mathbf{I}_{L\times L})$. Query signals are assumed to be mutually independent.

We also introduce the indicator $\mathbf{s}$ vector as

$$s_i \triangleq \begin{cases} 1, & \text{if} \quad \mathbf{y}_i \quad \text{is watermarked} \\ 0, & \text{otherwise} \end{cases},$$

where $i = 1, \ldots N$; the components of the corresponding random vector $\mathbf{S}$ are independent and identically distributed.

Let us now compute $P_{F,q}$. To start with, we need to evaluate the probability p_Δ that a query $\mathbf{Y}$ made by a honest user falls inside $\mathcal{A}$. Specifically, we can write:

$$p_\Delta \triangleq P(\mathbf{Y} \in \mathcal{A} | H_{q,0}) = P(\mathbf{Y} \in \mathcal{A} | H_{q,0}, S = 0) p_S(0) + P(\mathbf{Y} \in \mathcal{A} | H_{q,0}, S = 1) p_S(1), \quad (2)$$

where $p_S(0)$ and $p_S(1)$ are the *a priori* probabilities of having a watermarked or non-watermarked signal under the null hypothesis (honest queries). The above probabilities can be redefined as a function of the correlation ρ, since $\mathbf{Y} \in \mathcal{A}$ iff $|\rho - T| \leq \Delta$. Consequently, we have (see [9] for more details)[2]:

$$P(\mathbf{Y} \in \mathcal{A} | H_{q,0}, S = 0) = Q\left(\frac{T - \Delta}{\sigma_\rho}\right) - Q\left(\frac{T + \Delta}{\sigma_\rho}\right), \quad (3)$$

and similarly,

$$P(\mathbf{Y} \in \mathcal{A} | H_{q,0}, S = 1) = Q\left(\frac{\gamma L - (T + \Delta)}{\sigma_\rho}\right) - Q\left(\frac{\gamma L - (T - \Delta)}{\sigma_\rho}\right). \quad (4)$$

The probability of having K out of N queries in $\mathcal{A}$ can be computed by resorting to the formula of repeated Bernoulli trials; that is (we assume for simplicity that αN is an integer number):

$$P_{F,q} = \sum_{K=\alpha N}^{N} \binom{N}{K} p_\Delta^K (1 - p_\Delta)^{N-K}. \quad (5)$$

For small N we can actually compute the value of (5), while, for large N, we could derive upper and lower bounds for $P_{F,q}$ by using Stirling's approximation for the binomial coefficient. By imposing $P_{F,q} \leq P_{F,q}^*$ for a prescribed maximum $P_{F,q}^*$, we can set the metadetection parameters. Clearly, there are many combinations of parameters Δ and α which lead to the same value of $P_{F,q}$.

Whenever the attack model (i.e., the model under $H_{q,1}$) is known, it can be exploited for choosing the pair (α, Δ) by searching for the couple which, among those satisfying the false positive constraint, minimizes the false negative probability. Otherwise, in order to fully define the test, we need to fix one of the two parameters, α or Δ, and use the $P_{F,q}$ constraint to set the other.

In the next sections we evaluate the performance of the CTB metadetection in the case of Add-SS watermarking.

III. Performance of the CTB metadetector in the presence of mixed queries

In this section, we evaluate the performance of the CTB-based metadetector in a more general scenario with respect to the one considered in [9]. Specifically, we assume that the oracle is shared by many users which can be honest or not. Under $H_{q,0}$, all the users are honest, and consequently, the query model does not change with respect to the one considered in Section II (Def. 1). Under $H_{q,1}$ instead, there is (at least) one malicious user hidden among the honest, so in the observation vector malicious queries are interleaved with honest ones. Since the CTB metadetector depends only on the null hypothesis, the metadetector defined in Section II is still the one used in the current scenario, although one would expect that poorer performance is achieved.

The motivation for such a generalized setting deserves a comment. Since there are many real situations in which the detector knows the origin of the queries (e.g. a server shared by multiple users), one may argue that the detector could simply run the metatest in [9] for each user separately, with no need to account for the generalized scenario. Nevertheless, by observing that the origin of the queries can be easily forged by a malicious user, it is clear that for the detector is better to resort to higher-level attack countermeasures, which makes the study of the generalized scenario relevant also in this case. On top of that, the scenario studied in this paper provides a model also for a different possible situation where the queries under $H_{q,1}$ are all made by a malicious user who, in order to hide the attack, mixes the attacking queries to honest ones.

In the following, we summarize the main steps for performance computation when all the queries under the alternative hypothesis are malicious (this analysis was originally derived in [9]). Then, in Section III-B, we generalize that analysis to the case of mixed queries.

A. Metadetection performance with malicious queries only

We derive the performance of the CTB-based metadector in the case in which all the N queries are made by the attacker. Although the CTB-based metatest does not depend on the attacking strategies

[1] Whenever T grows with the same velocity as $\sqrt{L}$ (i.e., $T \sim \sqrt{L}$), then $P_{F,w}$ is fixed as L grows, whereas, whenever T satisfies the condition $(\gamma L - T) \sim \sqrt{L}$, $P_{M,w}$ is fixed as L grows. Clearly, these choices do not lead to an asymptotically zero error probability for the watermark test.

[2] Q denotes the Q-function: $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$.

(assuming that they result in an unnaturally high number of queries close to the boundary), in order to evaluate the performance of the CTB-based metadetector, we need to specify a model for the queries under $H_{q,1}$. Due to its practical interest, we consider that the attacker performs the queries according to a binary line search. Starting from two sequences $\mathbf{z}_1$ and $\mathbf{z}_2$, the former belonging to $\mathcal{R}_{w,0}$ and the latter to $\mathcal{R}_{w,1}$, the attacker applies a bisection algorithm on the line identified by the two queries until he finds a point that is arbitrarily close to the boundary. The number of queries which fall outside $\mathcal{A}$ corresponds to the number of queries required, according to the bisection method, to reach the interval size discrimination of 2Δ, which is (at most) $\lceil \log_2(|\rho_2-\rho_1|/\Delta)\rceil$, where $\rho_i = \langle \mathbf{z}_i, \mathbf{w} \rangle$ (we take the starting pair of queries of the line search as part of the N-length observation). For sake of simplicity, in the sequel, we neglect the upper integer approximation. Then, under $H_{q,1}$, after N observations (we assume $N \geq \log_2(|\rho_2 - \rho_1|/\Delta)$), the number of queries in $\mathcal{A}$ is larger than $N - \log_2(|\rho_2 - \rho_1|/\Delta)$.

The CTB test yields a correct decision (i.e., in favor of $H_{q,1}$) whenever:

$$\log_2\left(|\rho_2 - \rho_1|/\Delta\right) \leq (1-\alpha)N, \tag{6}$$

for some chosen α. For fixed N, according to equation (6), the test succeeds if the distance between the initial queries along $\mathbf{w}$ is not too large. In order to evaluate the probability of this event, we need to determine the statistics of $\rho_2 - \rho_1$, i.e., the difference between the queries $\mathbf{z}_1$ and $\mathbf{z}_2$ in the projected domain. With regard to $\mathbf{z}_1$, the attacker will query the detector with few non-watermarked sequences until he finds one belonging to $\mathcal{R}_{w,0}$ (the search will be extremely fast, since the probability that a non-watermarked sequence belongs to $\mathcal{R}_{w,0}$ is very high). As to $\mathbf{z}_2$, this is a watermarked sequence which belongs to $\mathcal{R}_{w,1}$ from which the attacker tries to estimate a point on (very closed to) the boundary. Then, the statistics of ρ_1 are obtained by conditioning to $H_{w,0}$ and to the event that a non-watermarked sequence belongs to $\mathcal{R}_{w,0}$, that is $f_{\rho_1}(\rho_1) = f_\rho(\rho | H_{w,0}, \rho \leq T)$. By construction, the probability that ρ_1 is smaller than T under $H_{w,0}$ is very close to 1 ($P_{F,w}$ is small); hence we can bound the conditioning event and assume that $\rho_1 \sim \mathcal{N}(0, \sigma_\rho^2)$. By the same token, we can state that $\rho_2 \sim \mathcal{N}(\gamma L, \sigma_\rho^2)$. Thanks to the independence of $\mathbf{z}_1$ and $\mathbf{z}_2$, $\rho_2 - \rho_1$ is still a Gaussian random variable, with mean value equal to γL and variance equal to $2\sigma_\rho^2$.

Given a pair (α, Δ) set by the metadetector, for large values of L, we can compute $P_{M,q}$ as follows:

$$\begin{aligned} P_{M,q} &= 1 - P\left((\rho_2 - \rho_1) \in \left[-\Delta \cdot 2^{(1-\alpha)N}, +\Delta \cdot 2^{(1-\alpha)N}\right]\right) \\ &\approx Q\left(\frac{\Delta \cdot 2^{(1-\alpha)N} - \gamma L}{\sqrt{2L\sigma_X^2}}\right), \end{aligned} \tag{7}$$

where we exploited the fact that $\gamma L \gg \sigma_\rho = \sqrt{L}\sigma_X$. As expected, doubling the width of the strip has the same effect on $P_{M,q}$ as decreasing α by $1/N$. From the metadetector side, equation (7) enables to optimize the value of the pair (α, Δ) among those ensuring that $P_{F,q} \leq P_{F,q}^*$. Be aware, however, that while for the computation of the false positive error probability we did not make any additional assumption on the behavior of the attacker, the expression of the false negative in (7), is valid only under the line search model.

It has been shown in [9] that, despite its generality (it is not tailored to counter the line search attack), the CTB metadetector uncovers oracle attacks by observing very few queries.

B. Metadetection performance with mixed queries

Let us consider the more general situation in which, under $H_{q,1}$, the queries are made partly by attackers and partly by honest users, and investigate how much the metadetection performance is reduced. For simplicity, as a first step, we study the metadetection problem by assuming that all the honest queries lie outside the metadetection region; then, we complicate the analysis by removing such assumption and considering the distribution of honest queries. Let $N - D$ be the number of malicious queries in the N-length vector of observations (equivalently, there are D honest queries).

1) Worst case scenario: We assume that the queries from honest users never fall inside the metadetection region $\mathcal{A}$. From the perspective of the metadetector, this corresponds to a worst case assumption. Given an observation vector of length N with D queries coming from honest users, the false negative error probability is simply[3].

$$P_{M,q} = P\left(n_y^{N-D}(\mathcal{A}) \leq \alpha N | H_{q,1}\right). \tag{8}$$

For the case of a binary line search attack, the metadetector yields a correct decision, regardless of the honest queries, if

$$\log_2\left(|\rho_2 - \rho_1|/\Delta\right) \leq (N - D) - \alpha N = (1-\alpha)N - D, \tag{9}$$

where, as before, ρ_1 and ρ_2 denote the correlation of the two starting queries with the watermark. The missed detection probability then becomes (for large L)

$$P_{M,q} \cong Q\left(\frac{\Delta \cdot 2^{(1-\alpha)N-D} - \gamma L}{\sqrt{2L\sigma_X^2}}\right). \tag{10}$$

From (10), it is easy to argue the impact of the honest queries on the missed detection probability. Due to the exponential inverse dependence of the argument of Q with D, each honest query (in place of a malicious query) has on $P_{M,q}$ the same effect of halving the size of the strip Δ or increasing α by $1/N$.

Throughout the paper, we denote by β the fraction of honest queries in the observation vector, i.e., $\beta = D/N$. Conversely, $(1-\beta) = (N-D)/N$ denotes the fraction of malicious queries. For a given metadetector with parameters (α, Δ), and a fixed maximum missed detection probability $P_{M,q}^*$ for the metatest, we derive the minimum fraction of malicious queries which still guarantee the prescribed $P_{M,q}^*$:

$$(1-\beta)^* = \alpha + \frac{1}{N}\log_2\left(\left(\sqrt{2\sigma_\rho^2}Q^{-1}(P_{M,q}^*) + \gamma L\right)/\Delta\right). \tag{11}$$

Notice that relation (11) corresponds to fixing a maximum distance between the starting pairs of queries for the attack, which is $\sqrt{2\sigma_\rho^2}Q^{-1}(P_{M,q}^*) + \gamma L$ in the projected domain.

2) Real scenario: Generally, honest queries may fall inside the metadetection region. Therefore, we relax the simplifying assumption made in the previous section and find the exact expression of the false negative probability $P_{M,q}$. We also derive conditions under which it is possible to approximate this expression with the one found above. Let $\mathbf{y}_L^{N-D}$ denote the vector of the queries made by the attacker: $\mathbf{y}_L^{N-D} = \{\mathbf{y}_i\}_{i \in \mathcal{S}}$, where $\mathcal{S}$ is the set of the indexes of the malicious queries in the observation vector, with cardinality $N - D$. Similarly, let $\mathbf{y}_H^D = \{\mathbf{y}_i\}_{i \in \bar{\mathcal{S}}}$ be the vector with the honest queries. Notice that, according to our model for honest queries, in order to compute $P_{M,q}$, the fact that the queries are not consecutive in the observation vector

[3] We are assuming $D \leq (1-\alpha)N$. A larger number of honest queries will always lead to $P_{M,q} = 1$.

does not affect the computations. We can write:

$$\begin{aligned}P_{M,q} =& P\left(n_y^N(\mathcal{A}) \leq \alpha N | H_{q,1}\right)\\ =& P\left(n_{y_L}^{N-D}(\mathcal{A}) \leq \alpha N\right) \cdot P\left(n_{y_H}^{D}(\mathcal{A}) = 0\right) +\\ &+ P\left(n_{y_L}^{N-D}(\mathcal{A}) \leq \alpha N - 1\right) \cdot P\left(n_{y_H}^{D}(\mathcal{A}) = 1\right) + ...\\ =& \sum_{i=0}^{\alpha N} P\left(n_y^{N-D}(\mathcal{A}) \leq \alpha N - i | H_{q,1}\right) \cdot\\ & \qquad P\left(n_y^{D}(\mathcal{A}) = \min\{i, D\} | H_{q,0}\right), \qquad (12)\end{aligned}$$

where $n_{\mathbf{y}_L}^{N-D}(\mathcal{A})$ ($n_{\mathbf{y}_H}^{D}(\mathcal{A})$) denotes the number of queries inside $\mathcal{A}$ among those in vector $\mathbf{y}_L^{N-D}$ ($\mathbf{y}_H^D$). In this case, given a target $P_{M,q}^*$, we are not able to derive β^* in an explicit form.

Expression $P\left(n_{\mathbf{y}}^{D}(\mathcal{A}) = i | H_{q,0}\right)$ can be computed by resorting to the formula of the repeated Bernoulli trials with parameter p_Δ, while $P\left(n_{\mathbf{y}}^{N-D}(\mathcal{A}) \leq \alpha N - i | H_{q,1}\right)$ depends on the specific attack. For the binary line search case, it can be computed similarly to (8).

Let us inspect the difference between the value of $P_{M,q}$ in (12) and that in (8). For the first term of the sum in (12) we have $P\left(n_y^D(\mathcal{A}) = 0 | H_{q,0}\right) = (1 - p_\Delta)^D$, while for the second term $P\left(n_y^D(\mathcal{A}) = 1 | H_{q,0}\right) = Dp_\Delta(1 - p_\Delta)^{D-1}$, $P\left(n_y^D(\mathcal{A}) = 2 | H_{q,0}\right) = (D(D-1)/2)p_\Delta^2(1-p_\Delta)^{D-2}$ and so on[4]. The two expressions (12) and (8) are very close to each other if it holds that $(1 - p_\Delta)^D \approx 1$ and then $Dp_\Delta(1 - p_\Delta)^{D-1} + ... \ll 1$. The validity of the approximation depends on the relation between p_Δ and D. Roughly speaking, this occurs when $(1 - p_\Delta)^D \gg Dp_\Delta(1 - p_\Delta)^{D-1}$, that is

$$p_\Delta \ll \frac{1}{D+1} \quad \left(\text{or } D \ll \frac{(1 - p_\Delta)}{p_\Delta}\right). \qquad (13)$$

When such relation is satisfied, we can approximate (12) by (8).

IV. Parameter analysis and asymptotic performance for the CTB metadetector

A question that arises when defining the metadetector is how to choose the observation length N. Since the metadetector does not know in principle whether it is queried by honest users or attackers, choosing N is not an easy task: with a too large N we might risk that an oracle attack takes place and ends up within the time interval of N queries, thus failing to detect it; on the other hand, a too small N might cause that only few attacking queries are made in the time interval N, and that the attack is spread over several vectors of N queries. To address this problem, it is interesting to investigate the behavior of the metatest for increasing values of N. In order to do that, it is useful to quantify the effectiveness of an oracle attack.

To keep things easy, we assume that the attacker is interested in estimating only a point of (sufficiently close to) the boundary and succeeds when such a point is found. The analysis can be extended to the cases in which the goal of the attacker is more general (e.g., he wants to learn the decision boundary and hence he needs to find at least L points of the boundary).

A. Quantifying the effectiveness of the attack

It is interesting to evaluate the performance of the attack in terms of the distance between the point obtained at the end of the search and the boundary of the watermarked region. Given a vector of N malicious queries, the worst case accuracy with which the attacker estimates a point on the boundary (with the line search attack) is described by the random variable $d_\varepsilon = \frac{|\rho_1 - \rho_2|}{2^{N-2}}$, which quantifies the worst case distance of the final query to the boundary (we subtract 2 for the initial pair of queries). Since $(\rho_2 - \rho_1) \sim \mathcal{N}(\gamma L, 2\sigma_\rho^2)$, $|\rho_2 - \rho_1|$ follows a folded normal distribution, which for large L can be approximated by the same Gaussian $\mathcal{N}(\gamma L, 2\sigma_\rho^2)$.

Then, for the case of N malicious queries, d_ε is distributed as follows:

$$d_\varepsilon \sim \mathcal{N}\left(\frac{\gamma L}{2^{N-2}}, \frac{2\sigma_\rho^2}{2^{2(N-2)}}\right). \qquad (14)$$

We can evaluate the attack performance by relying on the measure of quantiles. For a fixed probability q, the q-quantile is the mininum value d_ε^q which satisfies $P\left(d_\varepsilon < d_\varepsilon^q\right) \geq q$. For large q (close to 1), it quantifies the accuracy of boundary estimation that the attacker reaches with confidence $q \times 100\%$. From (14), we can compute d_ε^q by solving the following equation:

$$Q\left(\frac{2^{N-2}d_\varepsilon^q - \gamma L}{\sqrt{2\sigma_\rho^2}}\right) = 1 - q, \qquad (15)$$

which yields

$$d_\varepsilon^q = \frac{\sqrt{2\sigma_\rho^2}Q^{-1}(1-q) + \gamma L}{2^{N-2}}. \qquad (16)$$

Similarly, for the case in which a certain fraction β of honest queries are mixed with the malicious queries, the value of the q-quantile is $d_{\varepsilon,\beta}^q = d_\varepsilon^q 2^{\beta N}$.

From the point of view of the attacker, he might be interested in fixing a target accuracy and determining the number of malicious queries which are necessary to get close to the boundary with a certain accuracy. Reasonably, such number depends on the initial pair of queries. Let $d_{\varepsilon,s}^*$ be the target accuracy fixed by the attacker. Since the attacker does not know the watermark parameters, he will consider $d_{\varepsilon,s}^* = \|\mathbf{z}_1 - \mathbf{z}_2\|/2^{N-D-2}$, i.e. the accuracy in the original (spatial) domain. Let d_ε^* be the corresponding distance measured in the projected domain. The attacker succeeds in inducing an incorrect metadecision (i.e., a metadecision in favor of $H_{q,0}$) whenever

$$\log_2(\Delta/d_\varepsilon^*) < \alpha N - 2, \qquad (17)$$

(where $\log_2 \frac{\Delta}{d_\varepsilon^*} + 2$ is the number of queries which must fall inside the strip to get the accuracy d_ε), which implies

$$d_\varepsilon^* > \Delta/2^{\alpha N - 2}. \qquad (18)$$

The quantity on the right-hand side of (18) is the minimum error that the attacker has to admit (maximum accuracy that the attacker can reach) in order not to be discovered. Since the attacker cannot compute d_ε^*, he does not know if its attack will be detected or not. However, he can consider $d_\varepsilon^{*,ub} = \sqrt{L}d_{\varepsilon,s}^*$ (which is always larger than or equal to d_ε^*). Whenever $d_\varepsilon^{*,ub}$ does not satisfy (18), the attacker can conclude that his attack will surely be detected.

B. Sufficient and necessary conditions for an asymptotically powerful metatest

We now evaluate the performance as L and N increase (we assume that $N = N(L)$ and that N tends to infinity when L increases). This allows us to find conditions under which *asymptotically powerful testing* is possible, that is $P_{F,q} + P_{M,q} \to 0$ as $L \to \infty$.

We observe that, with regard to the metadetection parameters, it is reasonable to consider $\Delta < \min\{T, \gamma L - T\}$ (otherwise, the statistical mode of one of the two distributions (watermarked or non-watermarked) falls inside $\mathcal{A}$, which implies that a significant part of the corresponding density falls inside the strip). Given the

[4] We neglect the contribution of the subsequent terms ($K \geq 1$) since it gets smaller and smaller …

metadetection parameters, $(\alpha(L), \Delta(L))$ and $N(L)$, expressed as a function of L, we consider the sequence of metadetection tests as L grows. We can state the following:

Theorem 1: A necessary condition for asymptotic powerful testing is:

$$\lim_{L\to\infty} \frac{T \cdot 2^{N-D}}{L} > \gamma. \tag{19}$$

Proof: If equation (19) does not hold, it is easy to see that $P_{M,q} \nrightarrow 0$ as $L \to \infty$. Indeed, from the argument of the Q function in (10) we argue that:

$$\frac{\Delta \cdot 2^{(1-\alpha-\beta)N}}{\sqrt{L}} - \gamma\sqrt{L} < \frac{T \cdot 2^{(1-\beta)N}}{\sqrt{L}} - \gamma\sqrt{L} \to -\infty, \tag{20}$$

yielding $P_{M,q} \to 1$ [5]. ■

We now focus on the sufficient condition. Let $\alpha N = K$ denote the critical number of queries in $\mathcal{A}$, where $K(L) \to k$, for any positive $k > 0$, as $L \to \infty$ (then, $\alpha \geq 1/N$).

Theorem 2: A sufficient condition for asymptotic powerful testing is:

$$\lim_{L\to\infty} \frac{T \cdot 2^{N-D-2K}}{L} > \gamma. \tag{21}$$

Proof: We show that, if (21) holds, there exists a proper choice for the metadetection parameters that asymptotically leads to zero error probabilities. Let us take, for instance, $\Delta = T/2^{\alpha N} = T/2^{K}$. Let N be chosen in such a way that $p_\Delta = o(1/N)$ (see the appendix). In this way, it is easy to check that $P_{F,q} \to 0$. In fact, by exploiting a known upper bound on the binomial terms [11], as $L \to \infty$:

$$P_{F,q} \leq \frac{\alpha p_\Delta}{(\alpha - p_\Delta)^2 N} = \frac{p_\Delta}{(1 - \frac{p_\Delta}{\alpha})^2 \alpha N} \to 0. \tag{22}$$

where we exploited the fact that $p_\Delta = o(\alpha)$.

To show that the false negative probability also tends to zero, we observe that $\Delta \cdot 2^{(1-\alpha)N-D} = T \cdot 2^{N-D-2K}$. Then, by exploiting (21), it is easy to argue that the argument of the Q function in (10) tends to infinity and hence $P_{F,q} \to 0$. ■

Note that, in the above theorems, we considered the expression for the false negative error probability given in (8), that holds when no honest queries fall within the metadetection region (worst case situation from the point of view of the metadetector). It is easy to argue that nothing changes by considering the general expression for the false negative probability provided in (12). Indeed, by looking at the expression in (12), it is not possible to get a zero false negative probability asymptotically if (20) holds (Theorem 1). Besides, for the choice of the metadetection parameters made in the proof of Theorem 2, we have that $p_\Delta = o(1/N)$; then, constraint (13) is trivially satisfied and the value for the false negative error probability in (12) does not substantially differ from the one in (8).

As a consequence of Theorem 1, whenever [6]

$$(N - D)_L \leq \left\lceil \log_2\left(\frac{\gamma L}{T}\right) \right\rceil, \tag{23}$$

there is no choice of the parameters which leads to an asymptotically powerful test. Then, from the point of view of the attacker, we can interpret (23) as the maximum number of queries that the attacker can make to the oracle, in place of a honest user, to avoid being discovered. From Theorem 2, we argue that, whenever

$$(N - D)_L > \left\lceil \log_2\left(\frac{\gamma L}{T}\right) \right\rceil + 2K, \tag{24}$$

the metadetector asymptotically succeeds in detecting the oracle attack (asymptotically powerful test). Interestingly, if we limit the possible choices for the metadetection parameters and assume that $\Delta = T/2^K$, relation (21) provides a sufficient and necessary condition for having an asymptotically powerful metatest, and (24) provides the critical value for the attacking number of queries. This is reflected in our experiments in Section V.

From equation (24), we deduce that it is convenient for the metadetector to keep the critical number of queries in $\mathcal{A}$ as small as possible, and then choose $K(L) = k$, constant with L.[7] Therefore, from the perspective of the metadetector, relation (24) is an interesting result, stating that the number of queries *available* to the attacker, i.e., the number of queries that the attacker can make without being discovered, increases (only) logarithmically with L. We point out that equations (23) and (24) state a relationship between the performance of the metadetector and the watermarking system.

As a final remark, we observe that

$$\left\lceil \log_2\left(\frac{\gamma L}{T}\right) \right\rceil + 2K \geq \log_2\left(\frac{\gamma L}{T}\right) + 2K; \tag{25}$$

if we divide the latter by $\log_2(\gamma L)$, we obtain

$$1 - \frac{\log_2(T)}{\log_2(\gamma L)} + \frac{2K}{\log_2(\gamma L)}, \tag{26}$$

whose limit when $L \to \infty$ is larger than or equal to 1, as $(\gamma L - T)/\sqrt{L} \to \infty$. Consequently, if the attacker does not want to be detected he must consider a $(N - D)_L$ verifying $\lim_{L\to\infty}(N - D)/\log_2(\gamma L) \leq 1$. In that case, and considering a fixed q, $0 < q < 1$, from (16) we have that

$$\lim_{L\to\infty} \frac{\log_2(d^q_{\varepsilon,\beta})}{\log_2(\gamma L)} = 1 - \lim_{L\to\infty} \frac{(N - D)}{\log_2(\gamma L)} \geq 0, \tag{27}$$

providing a bound on the target accuracy that the attacker should adopt if he does not want to be detected.

V. Experimental results

By resorting to Montecarlo simulations, we measured the performance of the CTB metadetector by considering a typical additive spread spectrum watermarking system. Specifically, we considered the following setup: $L = 2 \cdot 10^4$, $\gamma^2 = 10^{-2}\sigma_X^2$ (Document-To-Watermark Ratio DWR = 20dB) and $T = 2.5\sigma_X L^{3/5}$. For the metadetector, we set: $\alpha = k/N$, with $k = 2$, and $\Delta = T/2^{\alpha N}$, and various values of the observation length N. For each setting, we performed 10^6 simulations. First, we simulated the querying process with N honest queries and evaluated the false positive error probability $\hat{P}_{F,q}$ by counting the fraction of wrong decisions (more than αN queries fall inside $\mathcal{A}$). Then, we simulated the querying process in the case where $N - D$ queries are made by an attacker according to the binary line search method, and evaluated the false negative probability $\hat{P}_{M,q}$ by counting the fraction of wrong decisions (less then αN queries fall outside $\mathcal{A}$).

Table I depicts the numerical results for various values of $N - D$, with $N = 1000$. These results show that the critical value in this setup is 6. Indeed, from the theoretical analysis the sufficient value for $N - D$ is $\lceil \log_2\left(\frac{\gamma L}{T}\right) \rceil + 4 = 6$ (note that since $\Delta = T/2^{\alpha N}$, 6 is also the theoretical critical necessary value). On the other hand, in Table II we consider the impact of different values of N on system performance (in all cases 6 attacking queries are used). The probability $\hat{P}_{M,q}$ is less than 10^{-6} for each N considered in the table (in line with (10)). Accordingly, 6 attacking queries in a block of 10000 queries are

[5] In the limiting case in which (19) holds with the equality, the argument of $Q(\cdot)$ is lower than or equal to zero, and then $P_{M,q} \geq 1/2$.

[6] We add the subscript L to explicit the dependence of $(N - D)$ on L.

[7] Note, in any case, that for finite L a trade-off between $P_{M,q}$ and $P_{F,q}$ must be achieved by choosing k.

sufficient to detect a threat with very small error probability. This suggests that the attacker has to spread his attack over a very large number of query blocks, and thus wait a considerable time to succeed if he does not want to be detected. Finally, experiments show that for the same setup, but with $L = 2 \cdot 10^5$, the critical number of queries is $M = 7$ (which confirms the logarithmic growth predicted by theory).

	N - D						
	2	3	4	5	6	7	8
$\hat{P}_{M,q}$	1	1	1	$6.845 \cdot 10^{-1}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$

TABLE I
$\hat{P}_{M,q}$ FOR VARIOUS VALUES OF $N - D$ (N = 1000, $\hat{P}_{F,q} < 10^{-6}$).

	N				
	10^2	10^3	10^4	$2 \cdot 10^4$	10^5
$\hat{P}_{F,q}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$2 \cdot 10^{-6}$	$6 \cdot 10^{-5}$

TABLE II
$\hat{P}_{F,q}$ FOR VARIOUS VALUES OF N ($N - D = 6$, $\hat{P}_{M,q} < 10^{-6}$).

VI. Conclusions

In this work we took a step towards the analysis of oracle attack metadetectors. Specifically, the assumption of a previous work [9] on the watermark detector to be exclusively used by a malicious attacker has been removed. By doing so, we got closer to a real scenario wherein multiple users (some of them legal, some of them malicious) query the same detector. The same analysis can be applied to the case in which the attacker himself mixes honest-looking queries with malicious ones, in order to reduce the probability of being detected. The performance of the CTB metadetector is analyzed in this new setup in terms of missed detection and false alarm error probabilities. As a result, we derived some useful performance bounds, that allow to determine the maximum attacking rate that a malicious user can afford without being discovered. Indeed, this maximum attacking rate is shown to depend on both the watermark detector and the metadetector parameters. These results illustrate the power of metadetectors as countermeasures against oracle attacks, since even in the realistic framework studied in this paper very few queries are necessary for a successful detection.

Future research will focus on the extension of our results to the case of targeted metadetectors, as the one based on line search originally proposed in [9].

Acknowledgment

Research supported by the Illegal use of Internet (INT) call within the Prevention of and Fight against Crime (ISEC) programme of the Home Affairs Department of the European Commission under project NIFTy (Project Number HOME/2012/ISEC/AG/INT/4000003892), the European Regional Development Fund (ERDF) and the Galician Regional Government for funding the AtlantTIC center, the Spanish Government and the ERDF under projects TACTICA and COMPASS, and the Galician Regional Government under project "Consolidation of Research Units".

References

[1] I. J. Cox and J. P. M. G. Linnartz, "Public watermarks and resistance to tampering," in *IEEE International Conference on Image Processing, ICIP'97*, vol. 3, (Santa Barbara, California, USA), pp. 3–6, October 1997.

[2] J. P. M. G. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," in *2nd International Workshop on Information Hiding, IH'98* (D. Aucsmith, ed.), vol. 1525 of *Lecture Notes in Computer Science*, (Portland, OR, USA), pp. 258–272, Springer Verlag, April 1998.

[3] M. F. Mansour and A. H. Tewfik, "LMS-based attack on watermark public detectors," in *IEEE International Conference on Image Processing, ICIP'02*, vol. 3, (Rochester, NY, USA), pp. 649–652, September 2002.

[4] M. E. Choubassi and P. Moulin, "Noniterative Algorithms for Sensitivity Analysis Attacks," *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 113–126, June 2007.

[5] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," in *4th International Workshop, IWDW'05* (M. Barni, I. Cox, T. Kalker, and H. J. Kim, eds.), vol. 3710 of *Lecture Notes in Computer Science*, (Siena, Italy), pp. 260–274, Springer Verlag, September 2005.

[6] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," in *IEE Proceedings on Information Security*, vol. 153, pp. 115–125, IET, September 2006.

[7] P. Comesaña and F. Pérez-González, "Breaking the BOWS watermarking system: key guessing and sensitivity attacks," *EURASIP Journal on Information Security*, vol. 2007, 2007.

[8] http://bows2.ec-lille.fr/index.php?mode=VIEW&tmpl=resPrevEp#ResEp2.

[9] M. Barni, P. Comesaña-Alfaro, F. Pérez-González, and B. Tondi, "Are you threatening me?: Towards smart detectors in watermarking," 2014.

[10] M. Barni and F. Bartolini, *Watermarking Systems Engineering*. Signal Processing and Communications, Marcel Dekker, 2004.

[11] W. Feller, *An introduction to probability theory and its applications. Vol. II*. New York: John Wiley & Sons Inc., second ed., 1966.

Appendix

In order to prove sufficiency of condition (21) we assumed $p_\Delta = o(1/N)$. We now check which conditions must be satisfied by N (growing rate) in order for p_Δ to satisfy such condition. The probability that a query $\mathbf{Y}$ falls inside $\mathcal{A}$ under $H_{q,0}$ goes to zero if both the two terms $P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0)$ and $P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 1)$ go to zero too. Let us focus on the behavior of the term $P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0)$. For large values of L we can write:

$$\begin{aligned} P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0) &\approx e^{-\frac{1}{2}\left(\frac{T-\Delta}{\sigma\rho}\right)^2} - e^{-\frac{1}{2}\left(\frac{T+\Delta}{\sigma\rho}\right)^2} \\ &\approx e^{-\frac{T^2+\Delta^2-2\Delta T}{2\sigma_\rho^2}} \left(1 - e^{-\frac{2\Delta T}{\sigma_\rho^2}}\right). \quad (28) \end{aligned}$$

The first term tends to 0, while the second term is non-negative and smaller than or equal to 1. Then, a sufficient condition ensuring that $P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 0) = o(1/N)$ is given by:

$$\lim_{L\to\infty} \left(Ne^{-\frac{T^2}{L}}\right) = 0, \quad (29)$$

that is,

$$\lim_{L\to\infty} \sqrt{T^2/L - \log(N)} = \infty. \quad (30)$$

Threshold T is a parameter of the watermarking system that we assume to be fixed. The metadetector chooses N in such a way that condition (30) is satisfied. Notice that the condition can always be satisfied for some positive $N(L)$, thanks to the assumption on the growing rate of T, made in Section II-B. As to $P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 1)$, by exploiting the exponential approximation of the Q function for large arguments, we have

$$P(\mathbf{Y} \in \mathcal{A}|H_{q,0}, S = 1) \approx e^{-\frac{1}{2}\left(\frac{\gamma L-(T+\Delta)}{\sigma\rho}\right)^2} - e^{-\frac{1}{2}\left(\frac{\gamma L-(T-\Delta)}{\sigma\rho}\right)^2}.$$

By reasoning as before, it is easy to argue that, if the condition $\lim_{L\to\infty} \frac{\gamma\sqrt{L}}{\sqrt{\log N}}$ holds, this probability goes to zero at a rate larger than $1/N$. This is a less strict condition w.r.t. to the one holding for the case $S = 0$, and it is then verified for the values of N which satisfy (30).