

A New Approach to Countering Ambiguity Attacks

Husrev T. Sencar
Computer and Information
Science Department
Six Metro Tech Center
Brooklyn, NY 11201
taha@isis.poly.edu

Qiming Li
Computer and Information
Science Department
Six Metro Tech Center
Brooklyn, NY 11201
qiming.li@ieee.org

Nasir Memon
Computer and Information
Science Department
Six Metro Tech Center
Brooklyn, NY 11201
memon@poly.edu

ABSTRACT

Watermarking based ownership dispute resolution schemes are vulnerable to a simple but effective class of attacks, called *ambiguity attacks*, which cast doubt on the reliability of resulting decision by exploiting the high false-positive rate of the watermarking scheme. To mitigate such attacks, we propose a new scheme that embeds multiple watermarks, as opposed to embedding a single watermark, while constraining the embedding distortion, and detects a randomly selected subset of them during an ownership proof. The crux of the scheme lies in both watermark generation, which deploys a family of one-way functions, and selective detection, which injects uncertainty to detection process. The reduction in false-positive probability is analyzed and compared to single watermark embedding for the additive watermarking technique through numerical solutions. Moreover, we examine the exact security level that can be achieved by our scheme with different combinations of parameters. We adopt a previous security proof of non-invertible watermarking schemes with modified notion of security that is more realistic in practice, which allows us to derive the achievable security with typical parameters.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

Keywords

Ambiguity attack, Watermarking, Security

1. INTRODUCTION

Digital watermarking schemes have been proposed to resolve ownership disputes over digital objects. In many cases, the owner of a digital object proves the ownership by showing the knowledge of a watermark that can be detected reliably from the object. Much previous work focuses on the robustness of the watermarking scheme. However, robustness is not sufficient in such ownership proofs. Consider the scenario where Alice has an original cover-object C_A , and

embeds her watermark W_A in C to obtain marked-object M , which is then distributed to the public. A malicious attacker Bob, given M , can try to find a fake original C_B and a fake watermark W_B , such that W_B is detectable in M . In this case, it is no longer clear who owns the object M . Such attack is often referred to as an *ambiguity attack*.

Craver *et al.* [1] introduced the first realization of an ambiguity attack, called *inversion attack*. They mention that these attacks are possible because the watermarking scheme can be easily *inverted* to get a fake watermark and a fake original from any marked-object. They propose a *non-invertible* watermarking scheme by requiring that the watermarks have to be generated by applying a one-way function on the original. It is asserted that an attacker would have to invert the one-way function to invert the watermarking scheme. Ramkumar *et al.* [2] pointed out that the schemes proposed in [1, 3] are not secure due to high false-positive rates, and give an improved scheme. Similar work along this line includes [4, 3, 5, 6, 7, 8].

Observing that it is difficult to design non-invertible watermarking schemes in a stand-alone setting, the involvement of a trusted third party (TTP) is considered in [9, 10, 11]. It is mentioned in [10, 11] that the security of previous non-invertible schemes are not rigorously analyzed, and that the watermarking scheme cannot be non-invertible if the false-positive is high.

In a stand-alone setting, Li and Chang [12] study the possibility of provably non-invertible watermarking schemes, and give a spread-spectrum based scheme. Their security proof relies on the security of a pseudo-random generator, which guarantees that no efficient algorithm can distinguish the output of the pseudo-random generator and that from truly random source. They show that the watermarks do not have to depend on the originals, and can be generated from any relatively short random seed. It is shown that when the false-positive of the underlying watermarking scheme is negligible, the resulting scheme is non-invertible.

In this paper, we observe that the false-positive of the underlying watermarking scheme is the key to non-invertibility, and provide a new method to reduce the false-positive rate of conventional watermarking based ownership dispute resolution schemes.¹ The scheme is based on embedding mul-

¹The proposed scheme is an improved version of the scheme described in [13] and corrects an error in its analysis.

multiple watermarks and detecting a randomly selected subset of them, as opposed to single watermark embedding, under constrained embedding distortion. We analyze the effectiveness of this scheme and apply it to additive watermark technique of [14] through analytical modeling and present a security analysis.

We further examine carefully the security of the proposed scheme. We note that although the security proof in [12] is theoretically sound, it is not straightforward to apply it in practice. Their results asserts that the success probability of any efficient attacker should be negligible if the false-positive is negligible. However, being negligible may not be necessary nor sufficient in practice. For example, an attacker that succeeds with a constant probability 2^{-100} is surely not a threat, yet a constant function is not negligible. On the other hand, an attacker that succeeds with probability 1 for all marked-objects that are not too large, but with probability 0 for all marked-objects that exceeds a certain size is not considered as a threat because the probability is negligible, yet it can be a real threat for many applications. Here we propose a modified notion of security that gives exact measure of the security level, and we analyze it using typical parameters.

The outline of the paper is as follows. We give an overall review of the literature in Section 2. In Section 3, definitions and assumptions of the watermarking model are given. We introduce and analyze multiple watermark embedding and selective detection scheme as a means to lower false-positive probability in Section 4. The security analysis is given in Section 5. We conclude in Section 6.

2. RELATED WORK

The basis of the first ambiguity attack given by Craver *et al.* [1] lies in the notion of *invertibility* of embedding operation which implies that given a marked-object it is easy to find and remove a watermark in it. As a consequence of the attack, the pirate causes an ownership dispute by making it possible to link the marked-object to two distinct originals unequivocally, at the complexity of finding a fake watermark. To cope with inversion attacks, Craver *et al.* [1] proposed imposing the *non-invertibility* requirement in watermark generation. This initiated a series of work aiming at devising *non-invertible* schemes that are built on conventional (embedding/detection) techniques [4, 3, 5, 6]. These approaches are based on the idea that the watermark cannot simply be a random signal and propose constructions that include one-way functions along the path of watermark generation. In particular, it is proposed that the watermarks should be computed by applying a cryptographic hash function on the cover-object. In this way, it is hoped that an attacker would have to invert the hash function, which is difficult.

In [2], Ramkumar *et al.* introduced another ambiguity attack on non-invertible schemes. They showed that if the false-positive rate of the underlying watermarking scheme is relatively high, the pirate does not need to invert the watermark generation process to obtain a fake original and a fake watermark. Later, their result is generalized in [7] and applied to the construction of [5]. In this attack, the pirate exploits the *diffusion* property of cryptographic constructions

by generating many watermarks from marked-object (or an attacked version) through introducing insignificant changes (e.g., tweaking bits). When the number of resultant watermarks is in the order of the false-positive rate, the pirate is very likely to obtain a fake watermark. The pirate then designates the object, that yielded the particular fake watermark, as his fake original. Since in this attack the true original and the fake one are still expected to be very *close* a fake watermark that can be detected in any of the two is very likely to be detected in the other as well. Therefore, to render attacks of this nature more difficult, Ramkumar *et al.* [8] proposed a semi-blind detection scheme in which the presence of the watermark in the original is also checked. Hence, to attack this improved detection scheme, the pirate has to ensure that his watermark cannot be reliably detected in the fake original.

To circumvent the limitations of embedding/detection schemes in resolving ownership disputes, the involvement of a trusted third party is proposed in watermark generation (e.g., [9, 10, 11]). In essence, these approaches try to achieve non-invertibility through constructions that make use of a trusted party. As compared to schemes described in [3, 8], this approach restricts brute forcing capability of the pirate since computation of each watermark requires querying the trusted party. On the other hand, when the pirate is able to make unlimited queries to the trusted party, the problem reduces to the one discussed above, and the complexity of an attack depends on the false-positive rate of the embedding/detection scheme. In [11], an alternative approach to ownership dispute resolution which requires a time-stamping service and a trusted dispute resolver is introduced. In this scheme, prior to release of the marked-object, the owner computes *commitments* (based on public-key encryption) to watermark and cover-object, which are intended for the dispute resolver, and obtains a time-stamp for these *commitments*. In this setup, an ownership dispute over a marked-object might arise only if the pirate is able to provide trusted party with a fake original, that is *similar* to the disputed object and yields a watermark that is detectable in the disputed object, and a time-stamp older than the owner's time-stamp.

Li and Chang [12] give the first provably non-invertible watermarking scheme without a trusted third party, where the underlying watermarking scheme is spread-spectrum based. Their security proof is a standard technique in cryptography. In particular, they show a reduction from the problem of breaking a secure pseudo-random generator, to that of inverting the watermarking scheme. In contrast to previous suggestions, they show that the watermarks do not have to depend on the originals, and can be generated from any relatively short random seed. In fact, it is crucial in their proof that the seed is sufficiently short, such that the number of *valid* watermarks is limited. A zero-knowledge proof protocol for their detection algorithm is given in [15].

3. MODEL

We restrict ourselves to the use of watermarking methods as a means to resolve ownership disputes. Accordingly, in this setting, the unpublished version of the newly created object will be referred to as the *cover-object* or *original*, and it will be denoted by C . Similarly, the *watermark*, which might have been produced by the owner or obtained from a trusted

party, is denoted by W , and the published version of the *original*, called the *marked-object*, by M . A watermarking system has three major components: *watermark generation*, *embedding*, and *detection*. The watermark generation algorithm yields a data string W which will be used to associate a specific object to its owner. Ideally, W is derived from C and the identity of the owner in a deterministic manner. The second component is concerned with watermark insertion. This is realized by an embedding function (embedder) \mathcal{E} which embeds the watermark W in cover-object C yielding the marked-object M where M and C are very similar with respect to a perceptual distortion measure. The third component of a watermarking system deals with watermark extraction. Depending on the design, a detection function (detector) \mathcal{D} might either extract a watermark or check the presence of a particular watermark W in a given marked-object M (or in a possibly modified version \hat{M}). Another important consideration in the design of a detector is the need for the cover-object in extraction of the watermark, *i.e.*, *blind* or *non-blind* watermark detection.

There are various embedding/detection methodologies proposed in the literature. One of the most common approach that governs the design of embedding/detection techniques is based on the principles of *linear spread-spectrum modulation* [14] and another common approach is based on *binning techniques and quantization* procedures [16, 17]. However, in practice, most ownership dispute resolution schemes are based on the former design due to its robustness properties and simplicity. In those schemes, the watermark embedding rule is linear and most typically a correlation based detector is deployed. In the case of additive embedding technique [14], the embedding rule can be described as

$$M = \mathcal{E}(C, W) = C + \alpha W \quad (1)$$

where $M, C \in \mathbb{R}^n$, $W \in \{-1, 1\}^n$, $\alpha \in \mathbb{R}$ and the distortion (per coefficient) due to embedding is α^2 . The detection of an embedded watermark is performed in a blind manner by verifying the presence or lack of the watermark. Hence, the detection function outputs a boolean value as

$$\mathcal{D}(M, W) = \begin{cases} \text{true}, & \text{if } \tau < \sum_1^n M[i] \times W[i] \\ \text{false}, & \text{otherwise.} \end{cases} \quad (2)$$

where τ is a suitably selected threshold value. On the other hand, it is essential that a watermark be derived from the cover-object and still be statistically independent from it. Moreover, given the watermark one should not be able to infer anything about the corresponding cover-object. When put together, satisfying these properties requires the involvement of one-way constructions in watermark generation. For simplicity, the watermarks are assumed to be generated by a pseudo-random generator, where the seed is taken as a one-way hash of the cover-object.

In the context of embedding and detection techniques, false-positive probability refers to possibility of detecting an unembedded watermark in a given object. In other words, for a given M , this the possibility of (2) yielding *true* value to more than one watermark. Essentially, ambiguity attacks are mainly due to high probability of detecting multiple watermarks in an object. A pirate exploits this by first searching the whole watermark space for those watermarks that can be detected in a given marked-object M and then ob-

taining a so-called original to support its claims in an ownership dispute. For this, the pirate is assumed to know exactly the details of the embedding/detection and watermark generation functions. To differentiate those watermark and originals associated with a pirate from the genuinely generated ones they will be referred to as fake watermarks and fake originals and denoted by C^* and W^* , respectively.

4. PROPOSED METHOD

To improve the achievable false-positive probability of watermarking techniques, we propose a scheme based on embedding multiple watermarks and detecting a randomly selected subset of them. The underlying idea of the scheme is to make brute-force search of fake watermarks more difficult, thereby reducing the effective false-positive probability of the overall scheme in comparison to conventional schemes, which embed a single watermark. This is realized by the use of multiple one-way functions in generating watermarks and successively embedding them into the cover-object. Hence, to mount an attack, the pirate needs to extract multiple fake watermarks (as opposed to only one) that are related through the predefined watermark generation rule. The crux of the scheme lies in its selective detection principle which refers to detecting only a randomly selected subset of the embedded watermarks. As compared to detecting all watermarks, the reduction in the number of watermarks to be detected essentially improves the probability of overall detection, thereby making detector's decision more reliable. However, due to the uncertainty as to which subset of watermarks will be detected, the pirate still needs to search for more number of fake watermarks to ensure a successful attack. That is, it is possible to improve the detection performance considerably without significantly reducing pirate's difficulty.

For this, we assume a generic hash function $h : \{0, 1\}^* \rightarrow \{0, 1\}^m$ and a family of pseudo-random generators $\mathcal{G} = \{g_1, g_2, \dots, g_s\}$ to generate the set of valid watermarks $\mathbf{W}_C = \{W_1, W_2, \dots, W_s\}$ from cover-object C such that $W_i = g_i(h(C))$, where the index of the watermark determines the particular pseudo-random generator that generated it. A sequence of watermarks are valid if and only if there exists some cover-object C from which they can be computed. Note that the parameter m is important for the security of the scheme. In particular, note that the total number of sequences of valid watermarks is at most 2^m .

To create the marked-object M the embedder takes as input the cover-object C and s watermarks in \mathbf{W}_C . When the ownership of an object has to be decided, the detector obtains $r \leq s$ (distinct) index values ranging between 1 and s from a trusted *random index generator* and forms an r -element subset of \mathbf{W}_C , by taking the watermarks with the corresponding indices. Then, it attempts to detect all the watermarks in the newly created subset in the given object. In this setup, a problem arises when a pirate is able to extract the randomly selected subset of watermarks that are computed from the fake cover-object by the pseudo-random generator associated with the designated index values.

Essentially, the associated probability of false-positives, p_{fp} , is a direct indicator of the difficulty of finding a fake watermark W^* detectable in M . In a similar manner, the proba-

bility of detecting any given set of watermarks $\{W_1^*, \dots, W_s^*\}$ in M is p_{fp}^s , as only one in every $\frac{1}{p_{fp}}$ randomly generated watermarks is likely to be detected in M . Now, when this scheme is deployed, finding a false-positive requires extracting a set of fake watermarks $\mathbf{W}_{C^*} = \{W_1^*, \dots, W_s^*\}$ from M . This in turn requires the presence of a fake original C^* satisfying $g_i(h(C^*)) = W_i^*$, $1 \leq i \leq s$. The corresponding probability for this will be in the order of p_{fp}^s , assuming each watermark is embedded at the same strength. Hence, a linear increase in the number of embedded watermarks causes an exponential drop in the overall probability of false-positives. However, since at the same time the detector has to reliably detect the s watermarks, rather than one, the detection performance also degrades exponentially. With selective detection, however, only a subset of r watermarks have to be reliably detected and, therefore, the performance degradation is exponential in $r < s$. On the other hand, the probability of finding a false-positive does not simply increase from p_{fp}^s to p_{fp}^r , as the latter does not take into account the uncertainty regarding which subset of watermarks needs to be detected. This uncertainty can be in two main forms.

1. In the first case, the r value is designated by the embedder/detector and only the actual index values of the r watermarks are randomly picked. In essence, randomly selected index values determine the r hash functions among $\binom{s}{r}$ possible combinations, that will be used for verifying the presence of corresponding watermarks. Therefore, in the uncertainty of the r indices, the probability of finding a false-positive is $\frac{1}{\binom{s}{r}} \times p_{fp}^r$ since the pirate has to both guess the right index values and search for r valid watermarks. Depending on the values of p_{fp} and s the combinatorial decay might be worse than an exponential one. That is, the $\frac{1}{\binom{s}{r}}$ term may reduce the overall false-positive probability more than p_{fp}^r term. Therefore, although ownership will be decided based on reliable detection of r randomly selected watermarks, it might be more advantageous for the pirate to extract more than r fake watermarks (because pirate's goal is to maximize the false-positive probability). For the general case, we assume the pirate searches for $l \geq r$ fake watermarks, that will be detected, at the expense of reducing p_{fp}^r to p_{fp}^l . Hence, given a specific choice of parameters r and l , false-positive probability is defined as

$$Pr[fp^{mul}|r, l] = \begin{cases} \frac{\binom{l}{r}}{\binom{s}{r}} \times p_{fp}^l, & 1 \leq r \leq l \leq s \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

It must be noted that in (3) the combinatorial term indicates the probability of the r watermark index values determined by the (trusted) random index generator to be among the l randomly selected indices by the pirate.

2. In the second case, however, a combination of indices are selected randomly with uniform probability over all possible combination of indices. That is, the random index generator provides $1 \leq r \leq s$ index values (*e.g.*,

a combination of r watermark detectors) and both the embedder/detector and the pirate are oblivious to both r value and the corresponding indices. Since there are $\sum_{j=1}^s \binom{s}{j}$ possible combinations of hash functions that the random index generator chooses from among and there are $\binom{s}{r}$ combinations with only r elements, probability of selecting a particular r value is defined as

$$Pr[r] = \frac{\binom{s}{r}}{2^s - 1} \quad (4)$$

where $\sum_{i=0}^j \binom{j}{i} = 2^j$ property is used. In the lack of any knowledge on r and the index values, the pirate extracts l fake watermarks by picking l indices with uniform probability. In fact, any particular selection of l indices yields $\sum_{j=1}^l \binom{l}{j}$ possible combinations and a false-positive arises only if the randomly selected r indices coincides with one of the possible combinations. Therefore, the probability of generating a false-positive for a given l is

$$\begin{aligned} Pr[fp^{mul}|l] &= \sum_r Pr[fp^{mul}|r, l] Pr[r] \\ &= \frac{1}{2^s - 1} \sum_{r=1}^l \binom{l}{r} p_{fp}^r \\ &= \frac{2^l - 1}{2^s - 1} \times p_{fp}^l. \end{aligned} \quad (5)$$

The overall false-positive rate then can be obtained by averaging (5) over all possible l values as

$$\begin{aligned} Pr[fp^{mul}] &= \sum_{l=1}^s Pr[fp^{mul}|l] Pr[l] \\ &= \frac{1}{s} \sum_{l=1}^s \frac{2^l - 1}{2^s - 1} \times p_{fp}^l. \end{aligned} \quad (6)$$

where $Pr[l] = \frac{1}{s}$.

4.1 Analysis of False-Positive Probability

In the context of ownership disputes, the primary concern is the detection of the presence of a particular watermark in a given object. Therefore, the detection process can be simply viewed as a procedure for statistically differentiating objects embedded with a specific watermark from the rest of the objects. Alternatively, this problem can be formulated as a binary hypothesis test. For this, let the null hypothesis \mathcal{H}_0 be “the object is not embedded with the specific watermark(s)”, and the alternative hypothesis \mathcal{H}_1 be “the object has the specific watermark(s) embedded into it.” Given an object O with unknown nature, the watermark detector tries to verify the presence of the watermark(s) by computing a test statistic ν , which is essential in making a decision to accept (or reject) one of the two hypotheses. The performance of a watermark detector is evaluated by receiver operating characteristics (ROC) analysis. This is based on two measures, namely probability of detection p_d and probability of false-positives p_{fp} . One important aspect regarding the deployment of additive watermarking technique is its linearity. Multiple watermark embedding and selective detection improves false-positive probability by making it difficult to find a valid set of (fake) watermarks. Achieving this gain requires that the pirate be permitted oracle access to the detector.

That is, the pirate may designate the input (an object and a watermark) to detector and observe the decision but cannot interfere with the operation of the detector. Otherwise, the pirate may exploit the linearity of the embedding scheme by enabling detection of the sum of watermarks $W_1 + \dots + W_s$ rather than detecting each watermark individually, thereby eliminating the improvements offered by the scheme and reducing it to single watermark embedding.

Consider the case of *single* watermark embedding where the marked-object is generated as

$$M_{one} = C + \alpha W_1. \quad (7)$$

Hence, the two hypotheses can be formulated as

$$\begin{aligned} \mathcal{H}_1 &: O = C + \alpha W_1 \text{ where } \frac{1}{n} \|O - C\| = \alpha^2 \\ \mathcal{H}_0 &: O = C, \end{aligned} \quad (8)$$

where O is an object whose type is in question. Correspondingly, the detector computes the detection statistic

$$\nu_{one} = \sum_{i=1}^{i=n} M_{one}[i] \times W_1[i] \quad (9)$$

and decides in favor of one of the hypotheses. Due to central limit theorem, test statistic ν_{one} can be shown to be a Normal distributed random variable under both hypotheses. Hence, p_{fp} and p_d are computed by comparing ν_{one} to a threshold τ as

$$p_{fp} = Q\left(\frac{\tau - E(\nu_{one}|\mathcal{H}_0)}{\sqrt{Var(\nu_{one}|\mathcal{H}_0)}}\right) \text{ and } p_d = Q\left(\frac{\tau - E(\nu_{one}|\mathcal{H}_1)}{\sqrt{Var(\nu_{one}|\mathcal{H}_1)}}\right) \quad (10)$$

where $Q(x)$ is the Gaussian error function defined as $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt$ and

$$E(\nu_{one}|\mathcal{H}_0) = 0, \quad Var(\nu_{one}|\mathcal{H}_0) = n\sigma^2, \quad (11)$$

$$E(\nu_{one}|\mathcal{H}_1) = n\alpha, \quad Var(\nu_{one}|\mathcal{H}_1) = n\sigma^2 \quad (12)$$

where σ^2 is the variance of the cover-object.

Considering a set of watermarks $\mathbf{W}_C = \{W_1, \dots, W_s\} \in \{-1, 1\}^{s \times n}$ to be embedded in C , the embedding rule takes the form of

$$M_{mul} = C + \frac{\alpha}{\sqrt{s}}(W_1 + \dots + W_s) \quad (13)$$

where the total embedding distortion is again α^2 . In selective detection, the detector extracts r watermarks (out of s embedded watermarks) designated by a trusted source. The detection statistic for each watermark is obtained as

$$\nu_{mul} = \sum_{i=1}^{i=n} M_{mul}[i] \times W_j[i] \quad 1 \leq j \leq r, \quad (14)$$

and the binary hypothesis testing of

$$\begin{aligned} \mathcal{H}_1 &: O = C + \frac{\alpha}{\sqrt{s}}(W_1 + \dots + W_s) \text{ where } \frac{1}{n} \|O - C\| = \alpha^2 \\ \mathcal{H}_0 &: O = C, \end{aligned} \quad (15)$$

is repeated for all watermarks. The probability of detection for each watermark p_d^{mul} and the false-positive probability

of p_{fp}^{mul} of selective watermark detection can be computed similar to (10) as

$$p_d^{mul} = Q\left(\frac{\tau' - E(\nu_{mul}|\mathcal{H}_1)}{\sqrt{Var(\nu_{mul}|\mathcal{H}_1)}}\right)^r \text{ and } p_{fp}^{mul} = \varphi \times (p_{fp}')^l \quad (16)$$

where

$$p_{fp}' = Q\left(\frac{\tau' - E(\nu_{mul}|\mathcal{H}_0)}{\sqrt{Var(\nu_{mul}|\mathcal{H}_0)}}\right) \text{ and } \varphi \in \left\{ \frac{2^l - 1}{2^s - 1}, \left(\frac{l}{r}\right)^s \right\}$$

depending on how r is assigned and

$$E(\nu_{mul}|\mathcal{H}_0) = 0, \quad Var(\nu_{mul}|\mathcal{H}_0) = n\sigma^2,$$

$$E(\nu_{mul}|\mathcal{H}_1) = \frac{n\alpha}{\sqrt{s}}, \quad Var(\nu_{mul}|\mathcal{H}_1) = n\sigma^2 + n\frac{s-1}{s}\alpha^2. \quad (17)$$

To compare the false-positive probability of multiple watermark embedding and selective detection to single watermark embedding, the probabilities of detecting a watermark in both cases have to be equalized by properly adjusting the threshold τ' so that $p_d = p_d^{mul}$ as

$$\tau' = Q^{-1}(p_d^{\frac{1}{r}}) \times \sqrt{n\sigma^2 + n\frac{s-1}{s}\alpha^2} + \frac{n\alpha}{\sqrt{s}}. \quad (18)$$

Correspondingly, p_{fp} and p_{fp}^{mul} can be expressed as

$$p_{fp} = Q\left(\frac{\tau}{\sigma\sqrt{n}}\right) \text{ and } p_{fp}^{mul} = \varphi \times Q\left(\frac{\tau'}{\sigma\sqrt{n}}\right)^l \quad (19)$$

where φ is as defined in (16).

Figure 4.1 provides the computed false-positive probabilities for the multiple watermark embedding and selective detection scheme obtained for $s = 25$ and varying values of r and l . Results show that when $r = l$ p_{fp}^{mul} reduces by a factor of about 10^7 throughout the range of p_{fp} as compared to single watermark embedding, see Figure 4.1 a. For the considered range of values of r and l , the results in Figures 4.1 b-g show that for relatively higher values of p_{fp} choosing $(r = 2, l = 25)$, respectively, yields the minimum achievable false-positive probability. Similarly, for the case of lower p_{fp} values, minimum false-positive probability is observed to be achieved at $(r = 7, l = 15)$. It should be noted that due to pirate's ability to choose l (as compared to fixing $r = l$) false-positive probability reduces only approximately by a factor of 10^4 .

4.2 Robustness Analysis

Another consideration is the robustness of the scheme. Since in multiple watermark embedding and selective detection scheme each watermark is embedded at a lower energy level (e.g., $\frac{\alpha^2}{s}$) and only a fraction (e.g., $\frac{r}{s}$) of total watermark energy is effectively utilized during detection, as compared to single watermark embedding and detection, the question to be answered is whether the promised performance improvements can be sustained in the presence of attacks. To determine the change in performance, we consider additive white Gaussian noise attacks.

In this setting, watermarks are detected from a distorted version of M , \hat{M} and, therefore, the changes in p_{fp} and p_{fp}^{mul}

need to be computed as a function of watermark energy to noise energy ratio (WNR) by generalizing the above analysis. This can be realized by reformulating the hypotheses in (8) and (15) as

$$\mathcal{H}_1 : O = C + \frac{\alpha}{\sqrt{s}}(W_1 + \dots + W_s) + N \text{ where } (20)$$

$$\frac{1}{n} \|O - C\| = \alpha^2 + \sigma_N^2 \text{ and } s \geq 1$$

$$\mathcal{H}_0 : O = C, (21)$$

where N is the zero mean white Gaussian noise with variance σ_N^2 and selecting $s = 1$ refers to single watermark embedding and detection case. Hence, corresponding false-positive probabilities for varying WNRs (*i.e.*, $\frac{\alpha^2}{\sigma_N^2}$ as embedding distortion is same for all $s \geq 1$) can be obtained as given in (10) and (16) via computing the statistics of the conditional distributions which can be shown to be equal to those given in (12) and (17) except for the variance under \mathcal{H}_1 . Due to independence of noise with the cover-object and the watermark(s), the two statistics are as

$$\begin{aligned} \text{Var}(\nu_{one}|\mathcal{H}_1) &= n(\sigma^2 + \sigma_N^2) \text{ and} \\ \text{Var}(\nu_{mul}|\mathcal{H}_1) &= n(\sigma^2 + \sigma_N^2) + n \frac{s-1}{s} \alpha^2. \end{aligned} (22)$$

Then, by calculating the thresholds τ in (10) and τ' in (18) necessary to achieve the designated detection probability, p_{fp} and p_{fp}^{mul} are computed. It should be noted that one can alternatively formulate \mathcal{H}_0 in (21) as $O = C + N$ rather than $O = C$. Although, at high WNR regime the two would yield similar performance, at lower WNR regime the latter poses a greater challenge to detection as the overlap between the distributions of detection statistics increases substantially with increasing s , resulting in a poor performance for the proposed scheme.

Figure 2 displays false-positive probabilities computed over a range of WNRs, from -40dB to 20dB (*e.g.*, $0.1\alpha \leq \sigma_N \leq 100\alpha$), for $s \in \{1, 6, 13, 25\}$ when detection probability ($p_d = p_d^{mul}$) is set to 0.8. Results show that when $r = s$ (*i.e.*, all watermarks are detected) single watermark embedding and detection scheme ($s = 1$) performs best for the whole range of WNR regime, and false-positive probability increases with increasing s . Essentially, the rate of increase in p_{fp}^{mul} with s , as compared to p_{fp} , determines the minimum rate of combinatorial decrease, *e.g.*, φ in (16), needed for selective detection scheme to offer an improvement. The results also indicate that for the considered values of r and l the reduction in false-positive probability is still considerable. Hence, despite the increasing noise levels false-positive probability reduces and robustness of the additive scheme is not compromised by selective detection of embedded watermarks.

4.3 Further Improvements

One way to further reduce the false-positive probability is to incorporate the approach of Ramkumar *et al.* [8] with the proposed scheme. This can be realized by verifying that the randomly selected subset of watermarks (*e.g.*, r out of s watermarks) detected in a marked-object cannot be detected in the corresponding cover-object. Since, the pirate's cover-object and the marked-object are not related through actual embedding, it is very likely that watermarks detected in one can also be detected in the other as well. In other

words, in this setting, the probability of successful detection, \hat{p}_d^{mul} , and probability of generating a false-positive, \hat{p}_{fp}^{mul} , depends on the joint probability of the two events. When $M = \mathcal{E}(C, \mathbf{W}_C)$, p_d^{mul} can be expressed in terms of p_d^{mul} and p_{fp}^{mul} , defined in (16) as

$$\begin{aligned} \hat{p}_d^{mul} &= \Pr(\mathcal{D}(M, \mathbf{W}_C) = \text{true}, \mathcal{D}(C, \mathbf{W}_C) = \text{false}) \\ &= \Pr(\mathcal{D}(M, \mathbf{W}_C) = \text{true}) \times \\ &\quad \Pr(\mathcal{D}(C, \mathbf{W}_C) = \text{false} | \mathcal{D}(M, \mathbf{W}_C) = \text{true}) \\ &= p_d^{mul} \times (1 - p_{fp}^{mul}). \end{aligned} (23)$$

It should be noted that the conditional probability in (23) is essentially the probability of not detecting the watermarks in \mathbf{W}_C in a random object which, by definition, is $1 - p_{fp}^{mul}$. Similarly for a given set of fake watermarks \mathbf{W}_C^* and a fake original C^* , \hat{p}_{fp}^{mul} can be obtained as

$$\begin{aligned} \hat{p}_{fp}^{mul} &= \Pr(\mathcal{D}(M, \mathbf{W}_C^*) = \text{true}) \times \\ &\quad \Pr(\mathcal{D}(C^*, \mathbf{W}_C^*) = \text{false} | \mathcal{D}(M, \mathbf{W}_C^*) = \text{true}) \\ &= p_{fp}^{mul} \times \epsilon. \end{aligned} (24)$$

In (24), the first term refers to possibility of generating a false-positive given M whereas the second one signifies the impossibility of two similar objects yielding opposite decisions. When the pirate obtains its original by introducing insignificant changes to M (*e.g.*, by tweaking bits of M) ϵ will approach zero, thereby making false-positive probability arbitrarily small, *i.e.*, $\hat{p}_{fp}^{mul} \rightarrow 0$, without reducing the probability of detection, *i.e.*, $\hat{p}_d^{mul} \approx p_d^{mul}$.

5. SECURITY ANALYSIS

5.1 Security Models

According to Li and Chang [12], a spread-spectrum based watermarking scheme can be non-invertible if n is sufficiently large and the valid watermarks are generated from a secure pseudo-random generator \mathcal{G} and an m -bit seed S , where m is a security parameter. They show that if an attacker can invert the watermarking scheme with a probability that is not negligible, he/she can break the pseudo-random generator \mathcal{G} with a probability that is not negligible, which contradicts with the assumption that \mathcal{G} is secure.

As noted in [18], although the security proof in [12] is theoretically sound, the asymptotic arguments may not be immediately useful for the design of practical non-invertible watermarking schemes. For example, it is not straightforward to answer questions such as the following: What is the minimum effort required by any smart attacker given a set of system parameters? How to choose system parameters if a certain level of security is required?

Here we use a security notion that focuses on the exact security of the resulting scheme, so that questions like the above can be answered. In particular, for any given security level measured by the minimum effort required by any attacker, we want to find out conditions on system parameters such that the security level can be achieved.

Our security notion is similar to that used in [18]. However, our analysis is more involved due to the selective detection procedure. In particular, as we will see in Section 5.2, the attack can be divided into two phases, namely, an *online*

phase and an *offline phase*. Our analysis focus on the success probability in the offline phase, and the actual definition of security will be given in Section 5.3.

5.2 Online and Offline Attacks

For each successful attack, we consider the pair (C^*, \mathbf{W}_{C^*}) generated by B . Let $\mathbf{W}_{C^*} = (W_1^*, W_2^*, \dots, W_s^*)$. For the attack to be successful, it must hold that (1) there are l indices $L \subseteq \{1, \dots, s\}$ such that for each $i \in L$, W_i^* is detectable in M , and (2) the set of r indices selected by the detector is a subset of L .

We observe that the attacker can find a pair (C^*, \mathbf{W}_{C^*}) with l detectable watermarks in \mathbf{W}_{C^*} “offline” in the sense that it can be done by knowing the parameters of the watermarking algorithm, and does not need to invoke the real watermark detector. On the other hand, the second condition can only satisfied while invoking the real watermark detector. Hence we consider the following hypothetical attacker, where the attacker is divided into two phases: an offline phase and an online phase.

1. (Offline). Select an l , such that with high probability $r < l$ for random r .
2. (Offline). Find a pair (C^*, \mathbf{W}_{C^*}) such that l of the watermarks in \mathbf{W}_{C^*} are detectable in C^* .
3. (Online). Send the pair (C^*, \mathbf{W}_{C^*}) to the detector.
4. (Online). Repeat the previous step if it is unsuccessful.

Let p_s be the probability that step 3 succeeds. We note that for any practical attacker, it is desirable to make sure p_s is large (e.g., 0.5). There are a few reasons for this. First, it may be very expensive to perform this step. For example, an attacker may have to send the object M to a remote server and wait for a long time before the result is returned. Secondly, the number of calls allowed to the detector may be limited. For instance, the attacker may have to go to a trusted third party (say, a judge) to perform such detections, and only a few trials are allowed. Lastly, if p_s is small, we can easily modify the ownership proof protocol such that the attacker succeeds with negligible probability. One simple way to achieve that is to require both parties (the attacker and the true owner) to perform such detections many times, and the party who succeeds more wins the ownership.

It is not difficult to see that p_s only depends on the values of l , s and r . More specifically, $p_s = \binom{l}{r} / \binom{s}{r}$. To achieve a large p_s , the attacker has to choose a large l , which can be made very close to s even when r is small. For example, when $s = 50$ and $r = 10$, l has to be at least 47 to achieve $p_s \geq 0.5$, and when $r = 5$, l has to be at least 44. Also, note that p_s is fixed when l , s , and r are fixed. Hence, we will focus on the success probability p of an attacker in the offline attack.

5.3 Exact Security in Offline Attacks

We first define an ambiguity attacker with respect to the success probability p_s in the online attacks.

Definition 1. (Ambiguity Attacker) A p_s -ambiguity attacker B is a probabilistic polynomial-time algorithm such that, given a marked object $M = \mathcal{E}(C, \mathbf{W}_C)$ for some cover-object C and valid watermarks \mathbf{W}_C , B finds a pair (C^*, \mathbf{W}_{C^*}) with probability p so that B succeed with probability at least p_s in the online attacks.

Now we define the security level of a non-invertible watermarking scheme.

Definition 2. (Non-Invertibility) A watermarking scheme is (p_s, ℓ) -non-invertible if for any p_s -ambiguity attacker B , its success probability $p \leq 2^{-\ell}$.

Our analysis is adapted from that in [12, 18]. The main idea is to show that if there is a p_s -ambiguity attacker with large success probability p (say, $p > 2^{-\ell}$ for some ℓ) and reasonably large p_s (say, $p_s \geq 0.5$), we can construct another algorithm \mathcal{T} (that makes use of B) to distinguish \mathcal{G} and a truly random source with a probability that is close to p (say, $p/2$). If \mathcal{G} is constructed in such a way that for all efficient algorithms, this probability p cannot be more than $2^{-(\ell+1)}$, we would come to a contradiction, hence such an ambiguity attacker cannot exist.

The algorithm \mathcal{T} , given input string \mathbf{W} , does the following.

1. Choose a random cover-object C .
2. Embed \mathbf{W} into C , obtaining marked object M .
3. If $\mathcal{D}(M, \mathbf{W}) = 0$, repeat from step 1.
4. Pass M to B and obtain its output.
5. If B finds a pair (C^*, \mathbf{W}_{C^*}) such that l of the watermarks in \mathbf{W}_{C^*} are detectable in C , output 1. Otherwise output 0.

The parameter l in the last step is determined by p_s , r and s as in Section 5.2. Clearly \mathcal{T} runs in polynomial time when the detection probability is high. Now consider the following two cases. In the first case, \mathbf{W} is a set of valid watermarks. In this case, B correctly outputs a pair (C^*, \mathbf{W}_{C^*}) with a probability $p > 2^{-\ell}$ by the hypothesis. The expected output of $\mathcal{T}(\mathbf{W})$ will be p .

In the second case, \mathbf{W} is uniformly random. In this case, let V be the probability that some valid watermark happens to be detectable in M . When this happens, let p' be the probability that attacker B finds a pair (C^*, \mathbf{W}_{C^*}) . We observe that the difference between p' and p must be very small, since otherwise the pseudo-random generator \mathcal{G} cannot be secure. It is safe to assume that $p/2 < p' < 2p$, considering that p is a constant $2^{-\ell}$ for some constant parameter ℓ . In this case the expected output of \mathcal{T} will be Vp' . Therefore, the difference in $\mathcal{T}(\mathbf{W})$ between these two cases is $p - Vp'$.

In the security proof in [12], the parameters of the watermarking scheme is chosen such that V is negligible, so that the above difference is not negligible. However, as shown

in [18], we only need to create a constant gap between p and Vp' , and it is sufficient to require that V to be a small constant (say, $V \leq 1/4$) for \mathcal{T} to differentiate the two cases with probability at least $p/2$.

Now, recall that the false-positive for one watermark is p_{fp} , the number of detectable watermarks that the attacker needs to find is l , and that the total number of watermark sets is no more than 2^m we have

$$V < 2^m p_{fp}^l. \quad (25)$$

Therefore, it suffices to require that $2^m p_{fp}^l \leq 1/4$ for the scheme to have ℓ bit security. In practice, the underlying pseudo-random number generator can be made so secure such that no efficient algorithm can break it much better than random guessing. Hence, we can safely assume that the success probability to break \mathcal{G} is at least 2^{m-1} . In this case, we have $\ell > m - 2$. In other words, a scheme satisfying (25) would be $(p_s, m - 2)$ -non-invertible, where p_s can be determined as in Section 5.2.

To illustrate how these results apply to the actual scheme, we give a numerical example here. Let $s = 50$, $r = 5$, and $l = 44$ (which gives $p_s > 0.5$). Let the parameters of the underlying watermarking scheme be that $\sigma = 100$, $\alpha = 7$, $n = 5000$, $p_d = 0.1$, and we choose $m = 63$. Hence $V < 2^m p_{fp}^l < 2^{-2.3}$, which satisfies the condition that $V < 1/4$. In this case, the scheme can achieve at least 61 bits of security. That is, no efficient algorithm can invert the scheme with a probability higher than 2^{-61} .

6. CONCLUSIONS

Although robustness to malicious manipulation is the core requirement for any watermarking technique, in the context of ownership proof, the problem arises due to a less intrusive type of attacks. The success of these attacks ultimately relies on the false-positive probability of the underlying watermark embedding/detection scheme. To make attacks due to high false-positive rates more difficult, we proposed and analyzed embedding multiple watermarks, rather than a single one, and selectively detecting them while constraining the embedding distortion. The security is achieved primarily due to the use of multiple one-way transformations in watermark generation and the uncertainty on the pirate's side as to which watermarks (among the embedded ones) will be detected. The proposed scheme provides an advantage over single watermark embedding only when the reduction in probability of successful detection can be compensated by the reduction in the false-positive probability due to the uncertainty inflicted by the selective detection. Moreover, a security analysis of the proposed system is performed to guide flexible implementations of similar schemes in practice.

7. REFERENCES

- [1] S. Craver, N. Memon, B. Yeo, M. Yeung: Can invisible watermarks resolve rightful ownerships. In: Technical Report RC 20509, IBM Research Institute (1997)
- [2] M. Ramkumar, A. N. Akansu: Image watermarks and counterfeit attacks: Some problems and solutions. In: Proc. of Content Security and Data Hiding in Digital Media. (1999)
- [3] S. Craver, N. Memon, B. Yeo, M. Yeung: Resolving rightful ownership with invisible watermarking techniques: Limitation, attacks, and implications. IEEE Journal on Selected Areas in Communications **16**(4) (1998) 573–586
- [4] W. Zeng, B. Liu: On resolving rightful ownerships of digital images by invisible watermarks. In: Proc. of ICIP. (1997) 552–555
- [5] L. Qiao, K. Nahrstedt: Watermarking methods for mpeg encoded video: Towards resolving rightful ownership. In: Proc. of ICMCS. (1998) 276–285
- [6] R. B. Wolfgang, E. Delp: A watermarking technique for digital imagery: Further studies. In: Proc. of SPIE: Voice, Video and Data Communications. (1997) 297–308
- [7] A. Adelsbach, S. Katzenbeisser, A. Sadegi: On the insecurity of non-invertible watermarking schemes for dispute resolving. In: Proc. of IWDW. (2003)
- [8] M. Ramkumar, A. N. Akansu: A robust protocol for proving ownership of multimedia content. IEEE Transactions on Multimedia **6**(3) (2004) 469–478
- [9] S. Katzenbeisser, H. Veith: Securing symmetric watermarking schemes against protocol attacks. In: Proc. of SPIE: Security and Watermarking of Multimedia Contents. Volume 4675. (2002) 260–268
- [10] A. Adelsbach, S. Katzenbeisser, H. Veith: Watermarking schemes provably secure against copy and ambiguity attacks. In: Proc. of ACM CCS-10 Workshop on Digital Rights Management. (2003)
- [11] A. Adelsbach, A. R. Sadeghi: Advanced techniques for dispute resolving and authorship proofs on digital works. In: Proc. of SPIE: Security and Watermarking of Multimedia Contents V. Volume 5020. (2003)
- [12] Li, Q., Chang, E.C.: On the possibility of non-invertible watermarking schemes. In: Information Hiding Workshop. Volume 3200 of LNCS. (2004) 13–24
- [13] Sencar, H.T., Memon, N.: Watermarking and ownership problem: A revisit. In: Proc. of ACM Workshop on Digital Rights Management. (2005) 93–101
- [14] I. J. Cox, F. Kilian, F. T. Leighton, T. G. Shamoan: Secure spread spectrum watermarking for multimedia. IEEE Transaction on Image Processing **6**(12) (1997) 1673–1687
- [15] Li, Q., Chang, E.C.: Zero-knowledge watermark detection resistant to ambiguity attacks. In: ACM Multimedia Security Workshop. (2006)
- [16] B. Chen, G. W. Wornell: Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Transactions on Information Theory **47**(4) (2001) 1423–1443
- [17] P. Moulin, R. Koetter: Data hiding codes. Proc. of IEEE **93** (2005) 2083–2126
- [18] Li, Q., Memon, N.: Practical security of non-invertible watermarking schemes. In: Proc. of IEEE ICIP, San Antonio, Texas (2007) To Appear.

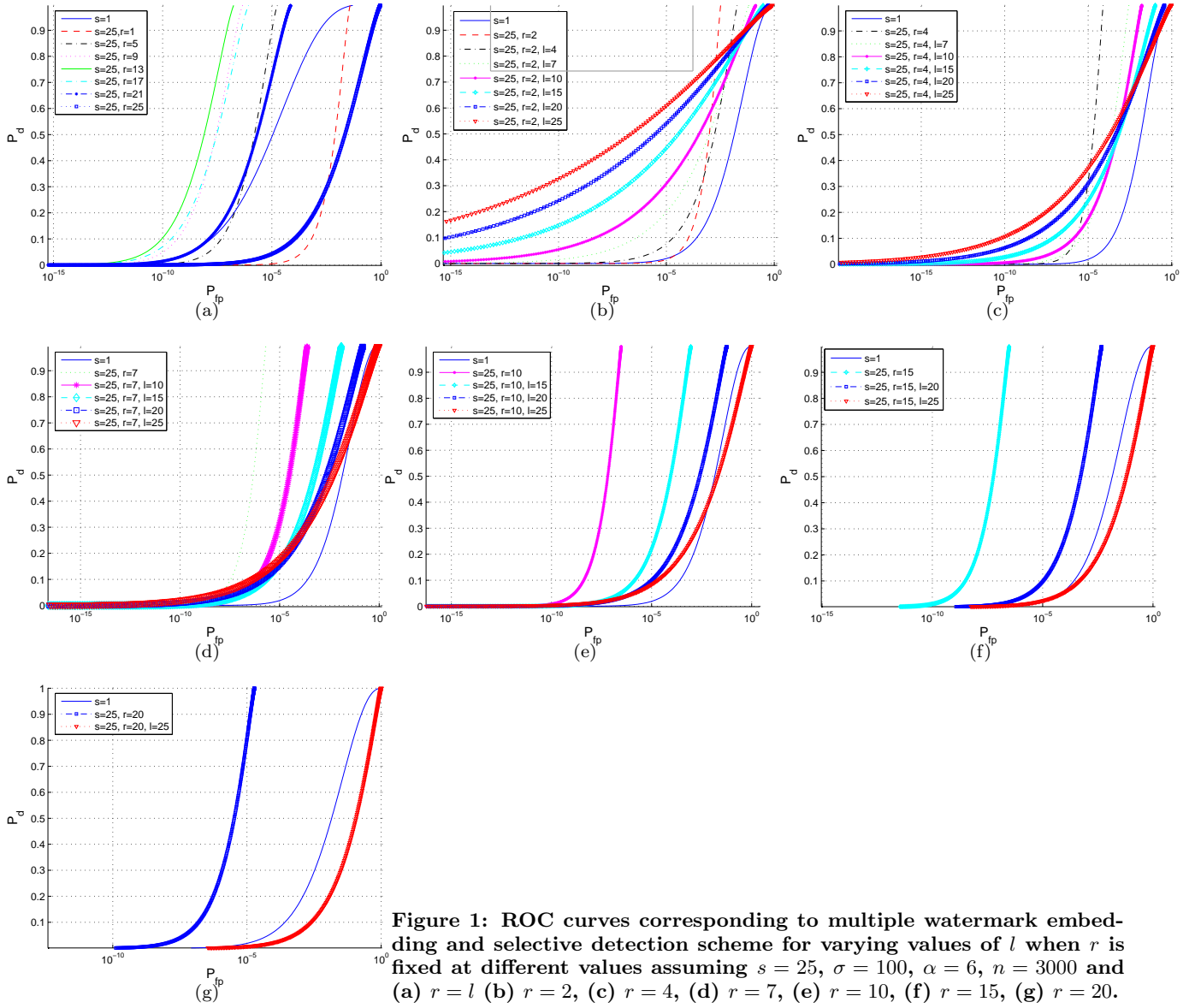


Figure 1: ROC curves corresponding to multiple watermark embedding and selective detection scheme for varying values of l when r is fixed at different values assuming $s = 25$, $\sigma = 100$, $\alpha = 6$, $n = 3000$ and (a) $r = l$ (b) $r = 2$, (c) $r = 4$, (d) $r = 7$, (e) $r = 10$, (f) $r = 15$, (g) $r = 20$.

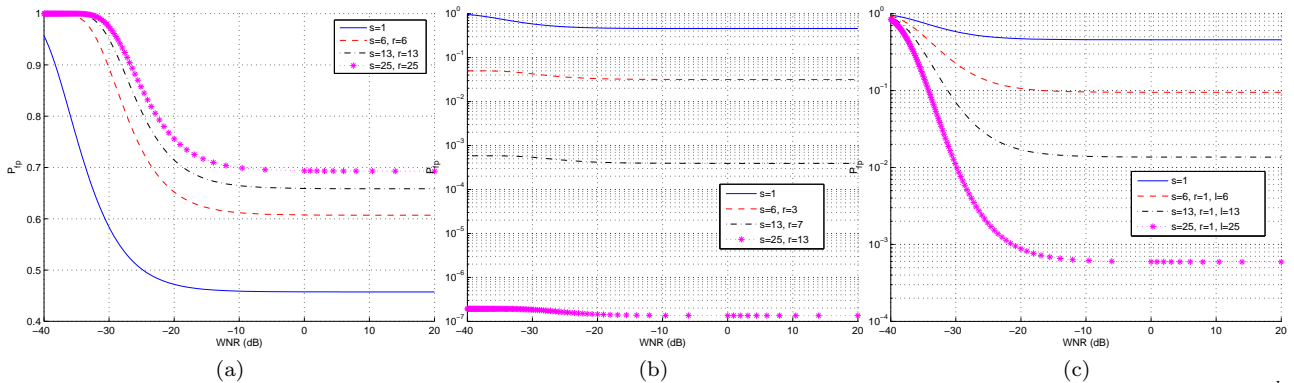


Figure 2: Change in false-positive probability as a function of WNR computed assuming $p_d = p_d^{mul} = 0.8$, $\sigma = 200$, $\alpha = 6$, $n = 1000$ and (a) $r = s$ (b) $r = l = \lfloor \frac{s}{2} \rfloor$ (c) $r = 1$, $l = s$.