

NONLINEAR COLLUSION ATTACKS ON INDEPENDENT FINGERPRINTS FOR MULTIMEDIA

Hong Zhao, Min Wu, Z. Jane Wang, and K. J. Ray Liu

Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742

ABSTRACT

Digital fingerprinting is a technology for tracing the distribution of multimedia content and protecting them from unauthorized redistribution. Collusion attack is a cost effective attack against digital fingerprinting where several copies with the same content but different fingerprints are combined to remove the original fingerprints. In this paper, we investigate average and nonlinear collusion attacks of independent Gaussian fingerprints and study both their effectiveness and the perceptual quality. We also propose the bounded Gaussian fingerprints to improve the perceptual quality of the fingerprinted copies. We further discuss the tradeoff between the robustness against collusion attacks and the perceptual quality of a fingerprinting system.

1. INTRODUCTION

With the rapid development of multimedia and communication technologies, an increasing amount of multimedia data are distributed through networks. This introduces an urgent demand to insure the proper distribution and usage of content, especially considering the ease of manipulating digital multimedia data.

To prevent illegal duplication and redistribution of the content, a digital fingerprinting system embeds unique identification information into each distributed copy to trace customers who use their copies inappropriately. There is a cost effective attack against digital fingerprinting, known as *collusion*. In collusion attacks, several users (colluders) get together, combine information from different fingerprinted copies of the same host signal and generate a new copy where the original fingerprints are removed or attenuated [1]. Digital fingerprinting should be resistant to collusion attacks as well as to common signal processings.

An early work on digital fingerprint code and collusion attacks assumed that colluders can detect and change a specific fingerprint code bit if it has different values between several fingerprinted copies [2]. Unlike generic data, fingerprints can be seamlessly embedded into the multimedia data [3, 4] and each fingerprint code bit can be spread over the entire multimedia content. Thus, the difference between each fingerprint code bit is not easily identified and changed. Therefore, the assumption of the collusion attack in [2] is suitable mostly for generic data and the attacks in [1] are more feasible for multimedia data.

In [1], several types of collusion attacks were studied and nonlinear attacks were shown to be more effective than the average attack in removing uniformly distributed fingerprints. Simulation results in [1] also show that Gaussian fingerprints are more resistant to nonlinear collusion attacks than uniform fingerprints but

analysis of the Gaussian fingerprint's performance has not been provided. In this paper, we focus on independent Gaussian fingerprints and analyze both the effectiveness and the perceptual quality of different collusion attacks. We use digital image as example, but our results are applicable to other types of multimedia data.

The paper is organized as follows. Section 2 introduces the fingerprinting and collusion attack system model. In Section 3, we analyze the detection statistic under different collusion attacks. In Section 4, we study the resistance of unbounded Gaussian fingerprints. Section 5 proposes bounded Gaussian fingerprints to improve the perceptual quality of fingerprinted copies. We then discuss the tradeoff between the robustness against collusion attacks and the perceptual quality that a designer of a fingerprinting system has to address. Conclusions are drawn in Section 6.

2. SYSTEM MODEL

We consider a system that consists of three parts: fingerprint embedding, collusion attacks, and fingerprint detection. Spread spectrum watermark embedding [3, 4] is widely used in watermark applications where the robustness of the watermark is required. Assume that there are a total of M users in the system. Given a host signal represented by a vector \mathbf{S} with length N , the owner chooses a unique fingerprint \mathbf{W}^i of length N for user $i = 1, \dots, M$, and generates the fingerprinted copy \mathbf{X}^i by $\mathbf{X}^i = \mathbf{S} + \alpha \mathbf{W}^i$. α is the *Just-Noticeable-Difference (JND)* from human visual models [4] to guarantee the imperceptibility of \mathbf{W}^i and control the energy of the embedded fingerprints. We assume that the M fingerprints $\{\mathbf{W}^i\}$ are chosen independently.

Assume that K users collude and S_C is the set containing the indices of the colluders. We further assume that the collusion attack is in the same domain as the fingerprint embedding. With K different copies $\{\mathbf{X}^k\}_{k \in S_C}$, the colluders generate the j th ($j = 1, \dots, N$) component of the attacked copy $\mathbf{V} = [V_1, V_2, \dots, V_N]^T$ using one of the following collusion functions:

$$\begin{aligned}
 \text{Average:} \quad V_j^{ave} &= \sum_{k \in S_C} X_j^k / K, & (1) \\
 \text{Minimum:} \quad V_j^{min} &= \min_{k \in S_C} \{X_j^k\}, \\
 \text{Maximum:} \quad V_j^{max} &= \max_{k \in S_C} \{X_j^k\}, \\
 \text{Median:} \quad V_j^{med} &= \text{median}_{k \in S_C} \{X_j^k\}, \\
 \text{MinMax:} \quad V_j^{minmax} &= (V_j^{min} + V_j^{max}) / 2, \\
 \text{Modified Negative:} \quad V_j^{modneg} &= V_j^{min} + V_j^{max} - V_j^{med}, \\
 \text{Randomized Negative:} \quad V_j^{randneg} &= \begin{cases} V_j^{min} & \text{with prob. } p, \\ V_j^{max} & \text{with prob. } 1 - p. \end{cases}
 \end{aligned}$$

The authors can be reached at hzhao, minwu, wangzhen and kjr-liu@eng.umd.edu

Note that for our model, applying the collusion attacks to the fingerprinted copies is equivalent to applying the collusion attacks to the fingerprints. For example, $\mathbf{V}^{min} = \min_{k \in S_C} \{\mathbf{S} + \alpha \mathbf{W}^k\} = \mathbf{S} + \alpha \min_{k \in S_C} \{\mathbf{W}^k\}$.

In the detection process, the detector removes the host signal from \mathbf{V} and extracts the fingerprint $\mathbf{Y} = g(\{\mathbf{W}^k\}_{k \in S_C})$, where $g(\cdot)$ is one of the collusion functions defined in (1). The detector measures the similarity between \mathbf{Y} and each of the M fingerprints $\{\mathbf{W}^i\}$, compares with a threshold, and outputs the estimated colluder set. In this paper, we use the Z statistic [1] to calculate the similarity between \mathbf{Y} and $\{\mathbf{W}^i\}$ because the Z statistic is found to be more robust against nonlinear collusion attacks than other commonly used statistics [5]. The Z statistic is defined as:

$$Z^i = \frac{1}{2} \sqrt{N-3} \log \frac{1 + \rho^i}{1 - \rho^i}, \quad (2)$$

$$\text{where } \rho^i = \frac{\frac{1}{N} \sum_{j=1}^N Y_j W_j^i - \frac{1}{N} (\sum_{j=1}^N Y_j) \frac{1}{N} (\sum_{j=1}^N W_j^i)}{\sqrt{\hat{\sigma}_W^2 \hat{\sigma}_Y^2}}$$

is the estimated correlation coefficient between \mathbf{Y} and \mathbf{W}^i , N is the length of the watermark, $\hat{\sigma}_W^2 = \frac{1}{N-1} \sum_j (W_j - \frac{1}{N} \sum_{j=1}^N W_j)^2$ and $\hat{\sigma}_Y^2 = \frac{1}{N-1} \sum_j (Y_j - \frac{1}{N} \sum_{j=1}^N Y_j)^2$ are the unbiased estimates of the original fingerprint's variance and the extracted fingerprint's variance, respectively. Z^i approximately follows Gaussian distribution $\mathcal{N}(\mu^i, 1)$ with $\mu^i = \frac{1}{2} \sqrt{N-3} \log \frac{1+E[\rho^i]}{1-E[\rho^i]}$ where $E[\rho^i]$ is the mean of ρ^i . If $i \in S_C$, then $\mu^i > 0$. Otherwise, $\mu^i = 0$. (Note that $\{E[\rho^i]\}$ are the same for all $i \in S_C$, so we will drop the superscript i for simplicity.)

We use the commonly used criteria to measure the effectiveness of different attacks: the probability of capturing at least one colluder (P_d) and the probability of falsely accusing at least one innocent user (P_{fp}). We also considered other measurements like the fraction of colluders that are successfully captured and the fraction of users that are innocently accused. From the analysis in [5], they have the same tendency as P_d and P_{fp} , and therefore are not included in this paper.

When considering the perceptual quality of different attacks, among all components of the noise (which is $\{n_j = JND_j \cdot \mathbf{Y}_j\}_{j=1}^N$ in our problem), only those that exceed JND result in perceptually distinguishable distortion. The mean square error (MSE) uses the total energy of the noise, so it is not an appropriate measurement of the perceptual distortion. We redefine MSE by $MSE_{JND} \triangleq \sum_{j=1}^N n_j'^2$, where

$$n_j' = \begin{cases} n_j + JND_j & \text{if } n_j < -JND_j, \\ 0 & \text{if } -JND_j \leq n_j \leq JND_j, \\ n_j - JND_j & \text{if } n_j > JND_j. \end{cases} \quad (3)$$

3. ANALYSIS OF DIFFERENT COLLUSION ATTACKS

3.1. Analysis of $E[\rho]$ under Different Attacks

From the analysis in the previous section, in order to analyze the effectiveness of different collusion attacks, we first need to study $E[\rho]$ for $i \in S_C$. Under the assumption that $\{W_j^i\}$ are i.i.d. distributed with zero mean and variance σ_W^2 , $\{g(\{W_j^k\}_{k \in S_C})W_j^i\}_{j=1}^N$ are also i.i.d. distributed. Recall that $E[\rho]$ is the correlation coefficient between \mathbf{Y} and \mathbf{W}^i (we will drop the subscript j for simpli-

fication), since $E[W^i] = 0$, we have

$$E[\rho] = \frac{\text{cov}[g(\{W^k\}_{k \in S_C}), W^i]}{\sqrt{\sigma_W^2 \sigma_Y^2}} = \frac{E[g(\{W^k\}_{k \in S_C})W^i]}{\sqrt{\sigma_W^2 \sigma_Y^2}}.$$

Therefore, $E[g(\{W^k\}_{k \in S_C})W^i]$ and σ_Y^2 are needed for the statistical analysis of Z statistic under each attack.

For the average attack, if $i \in S_C$, then

$$E\left[\sum_{k \in S_C} \frac{W^k}{K} W^i\right] = \frac{\sigma_W^2}{K}, \text{ and } \sigma_Y^2 = \text{var}\left[\sum_{k \in S_C} \frac{W^k}{K}\right] = \frac{\sigma_W^2}{K}.$$

For the minimum attack, if $\{W^i\}$ has the pdf $f(x)$ and the cdf $F(x)$, and if the number of colluders is K , the pdf of $W^{min} = \min_{k \in S_C} \{W^k\}$ is [6]:

$$f_{W^{min}}(w) = K f(w) [1 - F(w)]^{K-1}. \quad (4)$$

σ_Y^2 can be calculated from the definition of variance. In order to calculate the correlation between W^{min} and W^i for $i \in S_C$, we can express the joint pdf of W^{min} and W^i as follows:

$$f_{W^{min}, W^i}(w', w) = \begin{cases} f(w') [1 - F(w')]^{K-1} & \text{if } W^{min} = W^i, \\ (K-1) f(w') f(w) [1 - F(w')]^{K-2} & \text{if } W^{min} < W^i. \end{cases} \quad (5)$$

From (5) and the definition of correlation, for $i \in S_C$, we have

$$\begin{aligned} E[W^{min} W^i] &= E[W^{min} W^i]_1 + E[W^{min} W^i]_2, \\ \text{where } E[W^{min} W^i]_1 &= \int_{-\infty}^{\infty} w'^2 f(w') [1 - F(w')]^{K-1} dw' \\ \text{and } E[W^{min} W^i]_2 &= \int_{-\infty}^{\infty} w' (K-1) f(w') \times \\ &\quad [1 - F(w')]^{K-2} \left(\int_{w'}^{\infty} w f(w) dw \right) dw'. \end{aligned} \quad (6)$$

For the maximum and median attacks, the analysis is similar and detailed derivation is available in [5].

For the MinMax attack with $W^{minmax} = \frac{1}{2}(W^{min} + W^{max})$,

$$\begin{aligned} \text{var}[W^{minmax}] &= \frac{1}{4} (\text{var}[W^{min}] + \text{var}[W^{max}]) \\ &\quad + \frac{1}{2} \text{cov}[W^{min}, W^{max}], \\ E[W^{minmax} W^i] &= \frac{1}{2} (E[W^{min} W^i] + E[W^{max} W^i]), \end{aligned} \quad (7)$$

where the covariance of W^{min} and W^{max} can be calculated from the joint pdf of W^{min} and W^{max} , which is

$$f_{W^{min}, W^{max}}(w', w'') = K(K-1) f(w') f(w'') [F(w'') - F(w')]^{K-2}. \quad (8)$$

The analysis of the modified negative attack (ModNeg) is similar to the MinMax attack and can be found in [5].

For the randomized negative attack (RandNeg), we assume that p is independent of $\{W^i\}$. The colluded fingerprint can be rewritten as $W^{randneg} = W^{min} B_p + W^{max} (1 - B_p)$, where B_p is a Bernoulli random variable with parameter p and is independent of $\{W^i\}$. Thus the m th moment ($m = 1, 2, \dots$) of $W^{randneg} W^i$ and of $W^{randneg}$ are

$$E[(W^{randneg} W^i)^m] = E[E[(W^{randneg} W^i)^m | B_p]]$$

$$\begin{aligned}
&= p \cdot E[(W^{\min} W^i)^m] + (1-p) \cdot E[(W^{\max} W^i)^m], \\
\text{and } E[(W^{\text{randneg}})^m] &= E[E[(W^{\text{randneg}})^m | B_p]] \\
&= p \cdot E[(W^{\min})^m] + (1-p) \cdot E[(W^{\max})^m], \quad (9)
\end{aligned}$$

from which we can calculate $E[g(\{W^k\}_{k \in S_C})W^i]$ and σ_Y^2 .

3.2. Analysis of P_d and P_{fp}

From the analysis of $E[\rho]$ in the previous section, Z^i can be approximated with the following distribution:

$$Z^i \sim \begin{cases} \mathcal{N}(0, 1) & \text{if } i \notin S_C, \\ \mathcal{N}(\frac{1}{2}\sqrt{N-3} \log \frac{1+E[\rho]}{1-E[\rho]}, 1) & \text{if } i \in S_C, \end{cases}$$

where $E[\rho] = E[g(\{W^k\}_{k \in S_C})W^i] / \sqrt{\sigma_W^2 \sigma_Y^2}$. (10)

Let us define $\mu_Z \triangleq \frac{1}{2}\sqrt{N-3} \log \frac{1+E[\rho]}{1-E[\rho]}$. Among the M statistics $\{Z^i\}_{i=1}^M$, K of them are normally distributed with $\mathcal{N}(\mu_Z, 1)$ and the others are normally distributed with $\mathcal{N}(0, 1)$. If they are uncorrelated with each other or the correlation is very small, for a given threshold h , P_d and P_{fp} can be approximated with

$$\begin{aligned}
P_d &= P[\max_{i \in S_C} Z^i > h] \approx 1 - (1 - Q(h - \mu_Z))^K \\
\text{and } P_{fp} &= P[\max_{i \notin S_C} Z^i > h] \approx 1 - (1 - Q(h))^{M-K} \quad (11)
\end{aligned}$$

where $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$ is the Gaussian tail function.

3.3. Analysis of MSE_{JND}

For our digital fingerprinting and collusion attack model, given the collusion attack $g(\cdot)$ and the number of colluders K , if the colluded fingerprint $g(\{W^k\}_{k \in S_C})$ has the pdf $f_{g,K}(w)$, then MSE_{JND} can be simplified as:

$$\begin{aligned}
MSE_{JND} &= N \left\{ \int_{-\infty}^{-1} (w+1)^2 f_{g,K}(w) dw \right. \\
&\quad \left. + \int_1^{\infty} (w-1)^2 f_{g,K}(w) dw \right\}. \quad (12)
\end{aligned}$$

4. UNBOUNDED GAUSSIAN FINGERPRINTS

It was shown in [1] that uniformly distributed fingerprints can be easily defeated by nonlinear collusion attacks. The simulation results showed that Gaussian fingerprints are more resistant to nonlinear collusion attacks than uniform fingerprints. However, analysis on the resistance of Gaussian fingerprints to nonlinear collusion attacks was not provided. In this section, we study the performance of unbounded Gaussian fingerprints. Assume that fingerprints $\{W_j^i\}$ are generated from i.i.d. normal distribution $\mathcal{N}(0, \sigma_W^2)$. Usually we take $\sigma_W \approx \frac{1}{3}$ because it is required that almost all fingerprints (e.g., $\geq 99.9\%$) are in the range of $[-1, 1]$ and do not introduce perceptual distortion. For the randomized negative attack, we take $p = 0.5$ for the Bernoulli random variable B_p and assume that it is independent of $\{W_j^i\}$.

Given the analysis in the previous section, we first calculate $E[g(\{W^k\}_{k \in S_C})W^i]$ and σ_Y^2 for Gaussian fingerprints. Since there are terms of $Q^K(\cdot)$ in the pdf (4) and the joint pdfs (5)(8), analytical expressions are not available for the integration. We use recursive adaptive Simpson quadrature [7] to numerically evaluate the integrals with an absolute error tolerance of 10^{-6} .

Our numerical results show that for a given number of colluders K , $E[g(\{W^k\}_{k \in S_C})W^i]$ of different collusion attacks are the same and equal to σ_W^2/K . Different collusion attacks have different σ_Y^2 , as shown in Figure 1: the randomized negative attack has much larger variance than the other attacks, especially when the number of colluders is large; the modified negative attack has the second largest variance followed by the minimum and maximum attacks; the variances of the average, median and MinMax attacks are similar and the smallest. Consequently, from our analysis on $E[\rho]$, P_d and P_{fp} , the average, median and MinMax attacks are the least efficient attacks followed by the minimum, maximum and modified negative attacks. The randomized negative attack is the most effective attack. Our simulation results shown in Figure 2(a) agree with the analysis. Therefore, from the colluder's point of view, the best strategy for them is to choose the randomized negative attack.

So far we have studied the detection performance of the Z statistic under different collusion attacks. As to the perceptual quality, Figure 3 shows MSE_{JND}/N of different collusion attacks with i.i.d. $\mathcal{N}(0, \frac{1}{9})$ fingerprints. We can show that the minimum, maximum and randomized negative attacks yield the same MSE_{JND} . As we can see from Figure 3, although the randomized negative attacks is more effective than other attacks studied, it also introduces much larger distortion than JND, which is proportional to the number of colluders. This is because the fingerprinted signal is not bounded and in fact, such unbounded fingerprint can introduce noticeable distortions even without collusion.

5. BOUNDED GAUSSIAN FINGERPRINTS

In order to achieve both robustness and imperceptibility of the fingerprints, one possible solution for designers is to decrease σ_W^2 . However, decreasing σ_W^2 means reducing the energy of the embedded fingerprints, so the fingerprints are more vulnerable to attacks. In order to remove the perceptual distortion without reducing the energy of the embedded fingerprints, we introduce the bounded Gaussian fingerprints and study their performance under collusion.

Assume that $f_X(\cdot)$ and $F_X(\cdot)$ are the pdf and cdf of a Gaussian random variable with zero mean and variance σ_W^2 respectively. The pdf of a bounded Gaussian distribution $f'_X(\cdot)$ is:

$$f'_X(x) = \begin{cases} \frac{f_X(x)}{F_X(1) - F_X(-1)} & \text{if } -1 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

It is easy to show that fingerprints following pdf (13) has zero mean and variance σ_W^2 , and $MSE_{JND} = 0$ for fingerprinted copies, i.e., the fingerprints introduce no perceptual distortion. By bounding the fingerprints in the range of $[-1, 1]$, we maintain the energy of the embedded fingerprints while achieving the imperceptibility.

For the bounded Gaussian fingerprints having distribution (13), the analysis of the Z statistics under different attacks is similar to the unbounded Gaussian case and is not repeated here. The simulation results for bounded Gaussian fingerprints are shown in Figure 2(b). From Figure 2(b), we find that the randomized negative attack is still the most effective attack followed by the modified negative attack; all the other attacks have similar performance.

Since the original fingerprints are bounded by JND, all attacks have $MSE_{JND} = 0$. As a consequence, none of the attacks studied introduce perceptual distortion in the case of bounded fingerprints. Therefore, the colluders can choose the most effective attack without worrying about the perceptual quality.

6. CONCLUSIONS

In this paper, we have studied the resistance of independent Gaussian fingerprints to both average and nonlinear collusion attacks. We have also introduced the bounded Gaussian fingerprints to remove the perceptual distortion introduced by the unbounded Gaussian fingerprints. Based on both our analytical and simulation results, we have found that the randomized negative attack is the most efficient attack against both unbounded and bounded Gaussian fingerprints. In the former case, perceivable distortion may exist in the fingerprinted signals even when without collusion, and the randomized negative attack can introduce larger distortion, thus the colluders may prefer not to choose the randomized negative attack if imperceptibility is required. In the latter case, both the designers and the attackers do not introduce perceptual distortion, and the attackers can choose the most effective attack without perceptual concerns. Therefore, designers of fingerprinting systems should address the tradeoff between the robustness against collusion attacks and the perceptual quality of the fingerprinted copies.

Acknowledgement The authors would like to thank Dr. Wade Trappe for his constructive suggestions and inspiring discussions.

7. REFERENCES

- [1] H. Stone, "Analysis of attacks on image watermarks with randomized coefficients," Tech. Rep. 96-045, NEC, 1996.
- [2] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. on Information Theory*, vol. 44, no. 5, pp. 1897–1905, Sept. 1998.
- [3] I. Cox, J. Killian, F. Leighton, and T. Shamos, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. on Image Proc.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.
- [4] C. Podilchuk and W. Zeng, "Image adaptive watermarking using visual models," *IEEE Journal on sel. area in Comm.*, vol. 16, no. 4, pp. 525–540, May 1998.
- [5] H. Zhao, M. Wu, Z. Jane Wang, and K. J. R. Liu, "Nonlinear collusion attacks on independent multimedia fingerprints," *submitted to IEEE Trans. on Image Proc.*, 2002.
- [6] H. A. David, *Order Statistics*, New York: John Wiley and Son, 2nd edition, 1981.
- [7] W. Gander and W. Gautschi, "Adaptive quadrature - revised," *BIT 40(1)*, pp. 84–101, March 2000.

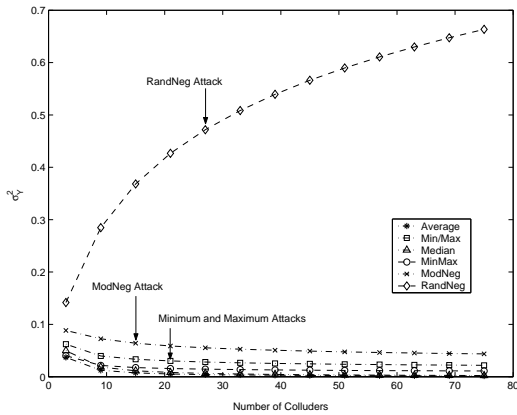
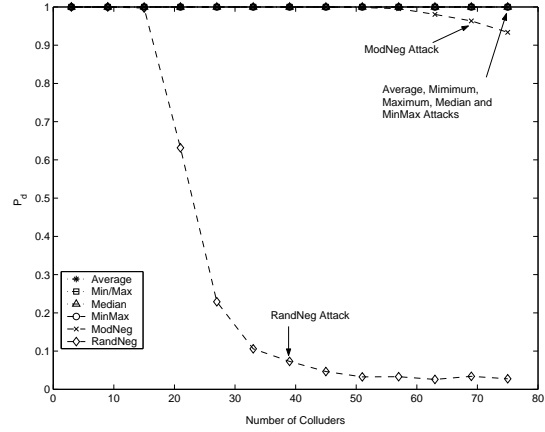
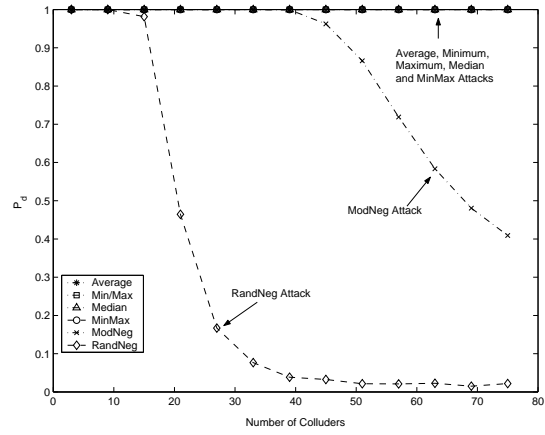


Fig. 1. σ_Y^2 of different attacks on unbounded Gaussian fingerprints with $\sigma_W^2 = 1/9$.



(a) Unbounded Gaussian fingerprints



(b) Bounded Gaussian fingerprints

Fig. 2. P_d of different attacks with $\sigma_W^2 = 1/9$ and fixed $P_{fp} = 10^{-3}$. Assume that there are a total of $M = 100$ users and the host image has $N = 10^4$ embeddable coefficients. Results are based on 2000 simulations. Simulation results on real images have the same tendency and therefore are not shown here.

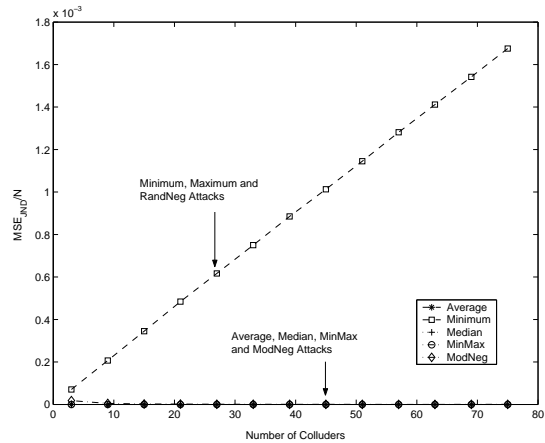


Fig. 3. MSE_{JND} of different attacks on unbounded Gaussian fingerprints with $\sigma_W^2 = 1/9$.