Attack Analysis for He&Wu's Joint Watermarking/Fingerprinting Scheme

Hans Georg Schaathun

University of Surrey
Department of Computing

Abstract. We introduce two novel collusion attacks against digital fingerprinting using additive spread-spectrum watermarks. These attacks demonstrate that the He-Wu fingerprinting system is considerably less secure against collusive attacks than suggested by the original paper. In addition to causing error rates above 85% at the decoder with as few as 8 colluders, one of our two attacks give copies with less distortion (measured by Euclidean distance) than the fingerprinted copies originally distributed.

1 Introduction

Unauthorised copying is a major worry for many copyright holders. As digital equipment enables perfect copies to be created on amateur equipment, many are worried about lost revenues, and steps are introduced to reduce the problem. Technology to prevent copying has been along for a long time, but it is often controversial because it not only prevents unauthorised copying, but also a lot of the legal and fair use.

A different approach to the problem is to deter potential offenders by application of forensic technology. These solutions do not prevent copying, but in the event that illegal copies are detected, they allow identification of the offenders, who can then be prosecuted. If penalties are sufficiently high, potential pirates are unlikely to accept the risk of being caught.

One forensic solution is digital fingerprinting, first proposed by Wagner [6]. Each copy of the copyrighted file is marked by hiding a fingerprint identifying the buyer. Illegal copies can then be traced back to one of the legitimate copies and the guilty user be identified. Obviously, the marking must be made such that the user cannot remove the fingerprint without ruining the file. Techniques to hide data in a file in such a way are known as robust watermarking. All references to watermarking (WM) in this paper refers to robust watermarking.

A group of users can compare their individual copies and observe differences caused by the different fingerprints embedded. By exploiting this information they can mount so-called *collusive attacks*. There is a growing literature on collusion-secure fingerprinting, both from mathematical and abstract and from practical view-points.

The goal of this paper is to make a critical review of the security of a recently proposed system by He and Wu [2], and to discuss the efficiency of various collusive attacks. In particular, we consider variations over the so-called minority choice attack, which has been largely ignored by the watermarking community and by [2] in particular. The He-Wu scheme serves as a case study for the evaluation of this attack.

We will discuss the relevant fingerprinting models in Section 2, and present the GRACE fingerprinting system of [2] in Section 3. In Section 4, we introduce our novel attacks, and in Section 5, we report our simulations and analysis. Section 6 presents our conclusions.

2 Fingerprinting Models

Digital fingerprinting is often viewed as a layered system. In the fingerprinting (FP) layer, each user is identified by a codeword $\mathbf{c}$, i.e. an n-tuple of symbols from a discrete q-ary alphabet. If there are M codewords (users), we say that they form an $(n, M)_q$ code.

In the watermarking (WM) layer, the copyrighted file is divided into n segments. When a codeword $\mathbf{c}$ is embedded, each symbol of $\mathbf{c}$ is embedded independently in one segment.

2.1 Digital Watermarking

A wide range of different WM techniques have been proposed. Following past works [2, 8], we will limit our study to non-adaptive, additive watermarks. It is commonly argued that in most fingerprinting applications, the original file will be known by the decoder, so that non-blind detection can be used [2].

We view the copyrighted file as a signal $\mathbf{x} = (x_1, \ldots, x_N)$, called the *host signal*, of real or floating-point values x_i. Since we use non-adaptive, additive watermarking, the message (or customer ID) to be embedded is mapped to a watermark signal $\mathbf{w} = (w_1, \ldots, w_N)$ which is independent of $\mathbf{x}$. The watermarked copy is the signal $\mathbf{y} = \mathbf{x} + \mathbf{w}$. The watermark $\mathbf{w}$ is designed such that $\mathbf{y}$ and $\mathbf{x}$ are perceptually equivalent.

The adversary, the copyright pirates in the case of fingerprinting, will try to disable the watermark by creating an attacked copy $\mathbf{z}$ which is perceptually equivalent to $\mathbf{y}$, but where the watermark cannot be correctly interpreted.

The watermark decoder takes the attacked signal $\mathbf{z}$. A non-blind decoder will subtract the original host signal, to obtain an attacked watermark signal $\mathbf{v} = \mathbf{z} - \mathbf{x}$. We will assume a correlation decoder, which returns the message associated with the watermark $\mathbf{w}$ which maximises the correlation $\mathbf{w} \cdot \mathbf{v}$.

A WM system can be used for fingerprinting, either alone (e.g.) [8] or in conjunction with a fingerprinting code [2]. In [8], each user u is associated with a watermark $\mathbf{w}_u$ where each sample w_i is chosen independently using a normal distribution.

In the model of [2], each user u is associated with a codeword $\mathbf{c}_u = (c_1^{(u)}, \ldots, c_n^{(u)})$ from an $(n, M)_q$ FP code. In the WM layer, q orthogonal watermarking signals $\mathbf{w}_s = (w_1^{(s)}, \ldots, w_m^{(s)})$ are used, where $w_i^{(s)} = \pm 1$ and $m \cdot n = N$. Their decoder views the two layers together, in a way equivalent to using soft-decision decoding. A watermark $\mathbf{w}'_u$ associated with user u, by concatenating the watermarks associated with the code symbols, i.e.

$$\mathbf{w}'_u = \mathbf{w}_{c_1^{(u)}} || \mathbf{w}_{c_2^{(u)}} || \mathbf{w}_{c_3^{(u)}} || \ldots || \mathbf{w}_{c_n^{(u)}}.$$

Based on this model, [2] present two novel improved solutions which we will describe in Section 3, namely group-based joint coding and embedding, and subsegment permutation.

2.2 Cut-and-Paste Attacks and Collusion-Secure Codes

One of the most studied attacks on digital fingerprinting is the *cut-and-paste* attack (aka. interleaving attack). A collusion of copyright pirates can compare their copies, and they will find that some segments differ. Such segments have ostensibly been used to hide a part of the fingerprint. By cutting and pasting segments from different copies, the colluders can produce a copy with a hybrid fingerprint distinct from those of the colluders.

Collusion-secure codes are designed to permit correct tracing in the presence of the cut-and-paste attack. A substantial literature is evolving on this topic. Usually an abstract mathematical model is used. The interface to the watermarking layer is defined as a marking assumption. The most commonly used assumption originates in [1], and says that every position (segment) of a hybrid fingerprint will match at least one of the fingerprints of the colluders.

Various strategies are available to colluders employing a cut-and-paste attack. Common strategies include the following.

Equal shares Divide the segments into groups groups of equal size, one for each colluder. All segments in a group are copied from the corresponding colluder.

Random For each segment, choose a colluder fingerprint uniformly at random and copy the segment therefrom.

Minority choice For each segment, the value occurring the fewest times in the corresponding segment of the colluders' fingerprints is used.

Majority choice For each segment, the value occurring the most times in the corresponding segment of the colluders' fingerprints is used.

Experimental work tends to focus on equal shares or random strategies. However, it is well known in the theoretical literature, that minority choice is most effective when closest neighbour or correlation decoding is used. This is because minority choice minimises the average similarity between colluder fingerprints and the hybrid fingerprint.

2.3 Attacks on joint watermarking/fingerprinting

There are important differences between the models commonly used for watermarking and for fingerprinting. Some of these differences result in different views on possible attacks, which we consider below.

Collusion-secure codes use discrete alphabets, whereas the host and watermark signals are numerical. Hence, in the joint model, a collusion attack can produce hybrid fingerprints which are numerical functions of the colluder fingerprints. A much studied example is the *averaging* attack, where the colluders produce an unauthorised copy by taking the average of the copies they have.

Example 1 *In the joint model, a collusion seeing* $+1$ *and* -1*, can produce a hybrid with a* 0*. In the (more abstract) FP model, the two symbols seen are just arbitrary symbols, say* A *and* B*, and the pirates can only output either* A *or* B*. The average of* A *and* B *is not defined.*

The Marking Assumption relies on each segment being a unit, and that the pirates cut and paste entire segments together. However, in most joint systems, the each watermark has to use a large number of samples to be sufficiently robust. Thus, the colluders could equally well cut and paste individual samples, effectively mounting a cut-and-paste attack on the WM layer.

The averaging attack, as well as a range of non-linear attacks, were studied in [8]. The attacks considered can be described as follows, where P is the set of fingerprints held by the collusions and $\mathbf{z}$ is the hybrid fingerprint produced by the attack.

$$\begin{aligned}
\text{Average:} \qquad & \bar{z}_i = \frac{1}{t} \sum_{\mathbf{y} \in P} y_i. \\
\text{Minimum:} \qquad & z_i^{\min} = \min_{\mathbf{y} \in P} y_i. \\
\text{Maximum:} \qquad & z_i^{\max} = \max_{\mathbf{y} \in P} y_i. \\
\text{Median:} \qquad & z_i^{\text{med}} = \operatorname*{median}_{\mathbf{y} \in P} y_i. \\
\text{Midpoint (MinMax):} \qquad & z_i^{\text{mid}} = (z_i^{\min} + z_i^{\max})/2. \\
\text{Modified negative:} \qquad & z_i^{\text{modneg}} = z_i^{\min} + z_i^{\max} - z_i^{\text{mid}}. \\
\text{Randomised negative:} \qquad & z_i^{\text{rndneg}} = \begin{cases} z_i^{\min} \text{with probability } p, \\ z_i^{\max} \text{with probability } 1-p, \end{cases}
\end{aligned}$$

It was assumed in [8], that p for the randomised negative attack be independent of the signals $\{y_i\}$. None of the above attacks adapt to the signal.

There are two important characteristics for the evaluation of fingerprinting attacks.

Success rate The attack succeeds when the watermark decoder does not return any of the colluders. Obviously, we want to maximise the rate of success.

Distortion The unauthorised copy has to pass as the original, so it should be as close as possible to the unknown signal $\mathbf{x}$ perceptually.

It is natural to expect low distortion from the average, median, and midpoint attacks. The pirate collusion is likely to include both positive and negative fingerprint signals. Consequently, these attacks are likely to produce a hybrid which is closer to the original sample than any of the colluder fingerprints. On the contrary, the maximum, minimum, and randomised negative attacks would tend to give a very distorted hybrid, by using the most distorted version of each sample. This is experimentally confirmed in [8].

Not surprisingly, the most effective attacks are the most distorting. The most effective attack according to [8] is the randomised negative, but the authors raise some doubt that it be practical due to the distortion.

2.4 Evaluation Methodology

The performance of existing fingerprinting schemes and joint WM/FP schemes have been analysed experimentally or theoretically. Very few systems have been studied both experimentally and theoretically. In the cases where both theoretical and experimental analysis exist, there is a huge discrepancy between the two.

It is not surprising that theoretical analyses are more pessimistic than experimental ones. An experimental simulation (e.g. [2]) have to assume one (or a few) specific attack(s). An adversary who is smarter (or more patient) than the author and analyst may very well find an attack which is more effective than any attack analysed. Thus, the experimental analyses give lower bounds on the error rate of the decoder, by identifying an attack which is good enough to produce the stated error rate.

The experimental analyses of the collusion-secure codes of [1, 5, 3] give mathematical upper bounds on the error rate under any attack provided that the appropriate Marking Assumption holds. Of course, attacks on the WM layer (which is not considered by those authors) may very well break the assumptions and thereby the system. Unfortunately, little work has been done on theoretical upper bounds for practical, joint WM/FP schemes.

In any security application, including WM/FP schemes, the designer has a much harder task than the attacker. The attacker only needs to find one attack which is good enough to break the system, and this can be confirmed experimentally. The designer has to find a system which can resist every attack, and this is likely to require a complex argument to be assuring.

This paper will improve the lower bounds (experimental bounds) for the He-Wu joint WM/FP system, by identifying adaptive, non-linear attacks, which are more effective than those originally studied. These attacks are likely to be effective against other joint schemes as well.

3 GRACE fingerprinting

We gave a brief overview of joint WM/FP in Section 2.1. The He-We solution is based on two additional features, namely Group-Based Joint Coding and Embedding Fingerprinting (GRACE), and subsegment permutations, which we introduce below.

Group-based fingerprinting assumes that one can divide the users into groups such that users are more likely to collude within a group than across several groups. A group could for instance be a limited geographical area, assuming that organising a world-wide collusion is more difficult and/or expensive than colluding with your town-mates.

3.1 The design

In the FP layer, we use a q-ary $[n, k]$ code C, with the so-called c-traceability property [4]. The code is linear of dimension k with codewords of length n. Each user is associated with one codeword (fingerprint). Let $D \subset C$ be a $[n, 1, n]$ repetition code, that is a subcode generated by a codeword of Hamming weight n. The system supports q^{k-1} disjoint groups of size q. The total number of users is $M = q^k$.

Fingerprints are assigned to users so that fingerprints in the same coset $D+\mathbf{x}$ in C belong to the same group. Consequently, the minimum Hamming distance between two fingerprints in the same group is equal to n, which is the maximum possible.

Each group is also associated with a codeword $\mathbf{c}' \in D$ over an alphabet of q^{k-1} symbols. Thus, each user will have a user fingerprint $\mathbf{c} \in C$ and a group fingerprint $\mathbf{c}' \in D$. Both these fingerprints will be embedded on top of each other in the WM layer.

The embedding uses $2q^{k-1}$ orthogonal sequences $\{\mathbf{u}_1, \ldots, \mathbf{u}_q; \mathbf{a}_1, \ldots, \mathbf{a}_{q^{k-1}}\}$. Consider a user associated with fingerprints

$$\mathbf{c} = (c_1, c_2, \ldots, c_n) \in C,$$
$$\mathbf{c}' = (c'_1, c'_2, \ldots, c'_n) \in A.$$

A WM signal $\mathbf{w}_u$ is constructed by concatenating n segments s_i defined as

$$s_i = \sqrt{1-\rho}\mathbf{u}_{c_i} + \sqrt{\rho}\mathbf{a}_{c'_i},$$

where ρ is used to adjust the relative energy of group and user information. The experiments in [2] used $\rho = 1/7$. The WM signal $\mathbf{w}$ is added to the host signal $\mathbf{x}$ as usual.

In the actual implementation tested in [2], the fingerprinting code C was a $[30, 2, 29]$ Reed-Solomon Code over $\mathsf{GF}(32)$. This code is 5-traceability code, meaning that it is collusion-secure under the Marking Assumption of [1] for collusions of size 5 or less. In the watermarking layer, 64 orthogonal sequences of length 1000 were used, requiring a total of 30 000 samples for embedding.

3.2 GRACE decoding

The GRACE decoder will first identify suspicious groups. Suppose group j is assigned $\mathbf{c}' \in D$. Define the associated group WM signal $\mathbf{g}_j$ as

$$\mathbf{g}_j = \mathbf{a}_{c'_1} || \mathbf{a}_{c'_2} || \dots || \mathbf{a}_{c'_n}.$$

A group j is deemed to be suspicious if $\mathbf{g}_j \cdot \mathbf{v} > \tau$, where τ is some threshold. Let $\mathcal{S}$ be the set of users who belong to a suspicious group. The decoder will return the user u solving

$$\max_{u \in \mathcal{S}} \mathbf{w}_u \cdot \mathbf{v}.$$

3.3 Subsegment Permutation

The second feature of the construction in [2], is that the segments are divided into subsegments, and the entire set of subsegments is permuted according to a secret key. The effect of this is that a collusion cannot mount a segment-wise cut-and-paste attack, because they have no way to identify the subsegments belonging to the same segment.

Observe that the subsegment permutation has no effect on sample-wise attacks. The only attacks it can counter are those using information about the segment structure. The only affected attack in this paper is the segment-wise cut-and-paste attack.

3.4 Original performance analysis

He and Wu analysed their scheme based on simulations of averaging and cut-and-paste attacks in combination with noise. Cut-and-paste was more effective than averaging. They did not analyse the performance absent noise. We quote their results for the cut-and-paste attack (using subsegment permutations) at a WNR of 0dB.

Drawing the pirates randomly from two groups, they had an error rate $\epsilon \approx 0.00$ up to 30 pirates. Drawing the pirates randomly across all groups, they had $\epsilon < 0.01$ up to 26 pirates and $\epsilon \approx 0.25$ at 30 pirates. The experiments without subsegment permutations had higher error rates.

4 The Novel Attacks

Let $\mathbf{w}_u$ be the watermark identifying user u, and let $\mathbf{v} = \mathbf{z} - \mathbf{x}$ be the hybrid watermark generated by the collusion. The correlation decoder calculates the heuristic

$$h_u = \mathbf{v} \cdot \mathbf{w}_u = \sum_{i=1}^{N} v_i \cdot w_i^{(u)},$$

for each u and returns the user(s) u with the largest h_u.

In order to avoid detection, the pirates should attempt to minimise $\max_{u \in P} h_u$. Without complete knowledge of the original host $\mathbf{x}$ and the watermark signals used, an accurate minimisation is intractable. However, attempting to minimise $\bar{h} = \operatorname{avg}_{u \in P} h_u$ is a reasonable approximation, and this can be done by minimising sample by sample, $\operatorname{avg}_{u \in P} v_i \cdot w_i^{(u)}$.

4.1 The minority extreme attack

If only two values occur in the watermark signals, the minority choice attack can be applied directly. The pirates seeing to different values of sample i, will use the one occurring the fewest times. Using GRACE, however, four different values occur, $\pm\sqrt{\rho} \pm \sqrt{1-\rho}$. Gaussian fingerprints [8] could give any number of distinct sample values.

Example 2 *Suppose the four possible values occur with the following frequencies*

$$\begin{aligned} a_1 &= x_i - \sqrt{1-\rho} - \sqrt{\rho} : 7 \text{ times,} \\ a_2 &= x_i - \sqrt{1-\rho} + \sqrt{\rho} : 1 \text{ times,} \\ a_3 &= x_i + \sqrt{1-\rho} - \sqrt{\rho} : 0 \text{ times,} \\ a_4 &= x_i + \sqrt{1-\rho} + \sqrt{\rho} : 2 \text{ times.} \end{aligned}$$

The minority choice a_2 will make a positive contribution to the correlation for 8 out of 10 colluder fingerprints, those with a_1 or a_2. If the pirates on the other hand choose a_4, 8 fingerprints will have negative correlation and only two will have positive correlation.

Remark 1 *In the case of GRACE, the colluders can in many cases recover x_i and remove the watermark completely from a sample. This is always true if they see four distinct values, as they can take the average of the minimum and the maximum. It is also true if they know ρ and see a_2 and a_3 or a_1 and a_4. Again x_i is the average of the two values.*

Since the remark is only applicable when there is a finite number of possible values, we focus on the idea of the example. When the average is close to the minimum, we choose the maximum, and vice versa. In mathematical notation, we write

$$\text{Minority Extreme (MX):} \qquad z_i^{\mathrm{MX}} = \begin{cases} z_i^{\min} \text{ if } z_i^{\mathrm{avg}} > z_i^{\mathrm{mid}}, \\ z_i^{\max} \text{ if } z_i^{\mathrm{avg}} < z_i^{\mathrm{mid}}, \end{cases}$$

and we shall see later that the experimental performance is good.

4.2 The moderated minority extreme attack

The above attack is expected to give distortion similar to that of the randomised negative, as it always chooses an extreme value. This problem has a simple fix.

Consider the difference $Z = z_i^{\text{avg}} - z_i^{\text{mid}}$. If $|Z|$ is small, it probably makes little difference to $\bar{h}$ whether $z_i = z_i^{\max}$ or $z_i = z_i^{\min}$. However, the distortion caused is likely to be the same regardless of Z. A solution is to use the average value when $|Z|$ is small, and the minority extreme attack when $|Z|$ is large. In other words,

$$\text{Moderated Minority Extreme (MMX):} \quad z_i^{\text{MMX}} = \begin{cases} z_i^{\min} \text{ if } Z > \theta, \\ z_i^{\text{avg}} \text{ if } \theta > Z > -\theta, \\ z_i^{\max} \text{ if } Z < -\theta, \end{cases}$$

where θ is some threshold. Again, the experimental analysis in the next section will confirm our intuition.

5 Experimental analysis

We have tried to replicate the algorithms from [2] as closely as possible. The authors did not specify how the orthogonal sequences are selected. To simplify coding, we used slightly longer sequences, 1024 bits rather than 1000, which should slightly improve the performance. These sequences were drawn randomly from a simplex code, mapping $0 \mapsto -1$.

For simplicity, we measure the distortion as the squared Euclidean distance (or power) $||\mathbf{y} - \mathbf{x}||^2$. This is intended to be a general estimate independent of the actual type of medium, and give a relative impression of distortion for the various attacks compared to the distortion in the fingerprinted copies. An exact measure would require a perceptual model, and would restrict the analysis to specific media.

We consider two different cases.

Two groups The pirate collusion is formed by t pirates drawn uniformly at random from the first two groups.

Any group The pirate collusion is formed by t pirates drawn uniformly at random from any group.

Group-based systems are only intended to have advantages in the two-group case, but [2] claim good performance in both cases, so we will discuss both.

We stress that we study exclusively the collusive attacks, and no additional noise attacks have been considered. This is for the simple reason that the two novel attacks are so effective that additional noise would not make any difference to the error rates.

We did not implement subsegment permutations, such that the performance under segment-wise attacks of our implementation is inferior to that of He-Wu's original implementation. However, the other attacks, being sample-wise, would be superfluous.

5.1 The group detection threshold

	Segment-Wise			Averaging		
t	Min.	Mean	Max.	Min.	Mean	Max.
2	5207.18	5991.31	6775.45	6014.53	6014.53	6014.53
5	1365.46	2491.35	3791.40	2322.21	2514.18	3014.23
10	353.75	1343.31	2745.63	1161.11	1362.37	2240.94
20	10.84	779.62	2181.33	580.55	783.44	1595.36
30	0.00	593.75	1913.50	387.04	592.55	1315.92

(a) Any group

	Segment-Wise			Averaging		
t	Min.	Mean	Max.	Min.	Mean	Max.
2	8347.58	8754.75	9161.91	8777.97	8777.97	8777.97
5	4255.84	6153.87	8051.89	4351.83	6165.48	7979.13
10	4026.72	5805.53	7584.35	4495.81	5840.37	7184.93
20	4507.42	5805.53	7103.65	4774.47	5805.53	6836.60
30	4629.72	5805.53	6981.35	4976.50	5805.53	6634.56

(b) Two groups

Table 1: Comparison of Group Decoding Heuristics for the attacks studied in [2]. The minimum, mean, and maximum are calculated per sample over all guilty groups. Innocent groups have a heuristic of 0 throughout. Averaging 500 samples per data point.

The first question we investigate is about the group decoding threshold τ. There is no recommendation for τ to be found in [2], so we have made a set of simulations, presenting the group decoding heuristics in Table 1.

Remark 2 *Observe that all innocent groups have a heuristic of 0. This is because all the sequences* $\mathbf{u}_i$ *and* $\mathbf{a}_i$ *are orthogonal. Calculating the correlation between the hybrid fingerprint and an innocent one, we find that each segments of the hybrid is a linear combination (either a single signal or an average of signals) of signals orthogonal to that of the innocent fingerprint.*

As we can see, a small positive threshold τ, would succeed in eliminating all innocent groups and retain almost all guilty groups under segment-wise cut-and-paste and under averaging.

	Guilty groups			Innocent groups		
t	Min.	Mean	Max.	Min.	Mean	Max.
2	5979.84	6047.89	6115.94	-254.15	-1.47	249.97
5	2168.74	2514.83	3163.88	-309.56	-1.03	303.33
10	936.94	1352.15	2287.28	-320.20	-2.00	318.50
20	317.09	784.58	1657.26	-296.10	0.55	294.48
30	112.82	597.55	1420.28	-276.91	0.52	276.96

Table 2: Group Decoding Heuristics under sample-wise cut-and-paste with equal shares. Averaging 500 samples per data point.

One of the claims in [2] was that the cut-and-paste attack is more effective when it is applied to entire segments than individual samples. On the whole, our simulations confirm this, but one point should be noted. Comparing Tables 2 and 1, we see that for 20 and 30 pirates, the gap between group decoding heuristics of innocent and guilty groups is smaller, hence less noise would have to be introduced to cause errors in the group decoding step. The group decoding heuristics of innocent groups are more spread when the attack is sample-wise.

We choose a group decoding thresholds of $\tau = 350$ for initial simulations, which is sufficient to exclude all innocent groups with high probability under the attacks studied so far. We will reconsider the threshold in Section 5.4.

5.2 The Minority Extreme Attack

	Guilty groups			Innocent groups		
t	Min.	Mean	Max.	Min.	Mean	Max.
2	5909.77	5982.16	6054.54	-249.38	0.50	250.87
5	-273.61	57.31	371.23	-258.34	60.38	652.04
7	-2118.88	-1176.22	-742.62	-326.51	94.43	680.24
10	-2889.77	-1554.41	-1005.88	-355.37	129.77	577.65
20	-5374.96	-2472.27	-1434.85	-164.68	206.50	616.22
30	-5408.57	-2247.93	-985.22	-135.92	236.44	635.33

(a) t random pirates from any group

	Guilty groups			Innocent groups		
t	Min.	Mean	Max.	Min.	Mean	Max.
5	2880.62	6169.04	9457.47	-254.89	52.38	519.69
7	3569.96	5943.76	8317.55	-228.04	72.44	504.76
10	1869.13	4576.03	7282.93	-196.99	102.61	507.94
20	913.69	4524.88	8136.07	-175.08	144.09	720.30
30	976.80	4797.54	8618.28	-193.41	161.30	823.89

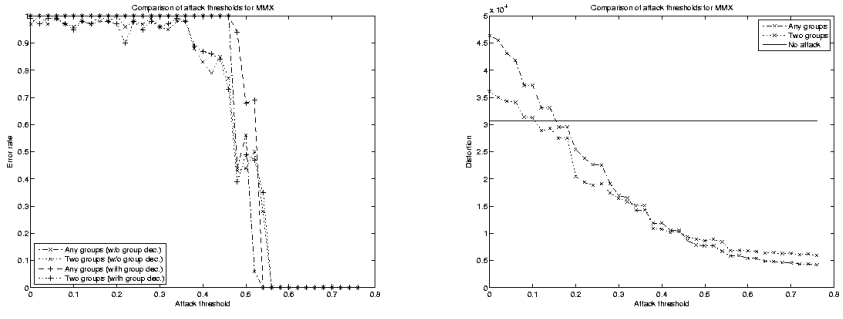
(b) t random pirates from two groups

Table 3: Group Decoding Heuristics under the Minority Extreme attack. Averaging 500 samples per data point.

Table 3 show group decoding heuristics under the minority extreme attack. The first we note is that 5-7 pirates drawn from all groups can completely wreck the group decoding. With five pirates, the mean group decoding heuristics are approximately equal for innocent and guilty groups. With seven pirates, all guilty groups have lower heuristic than any innocent group. In other words, if we require the system to be secure against collusions spanning 7 groups, we have to abandon the group decoding.

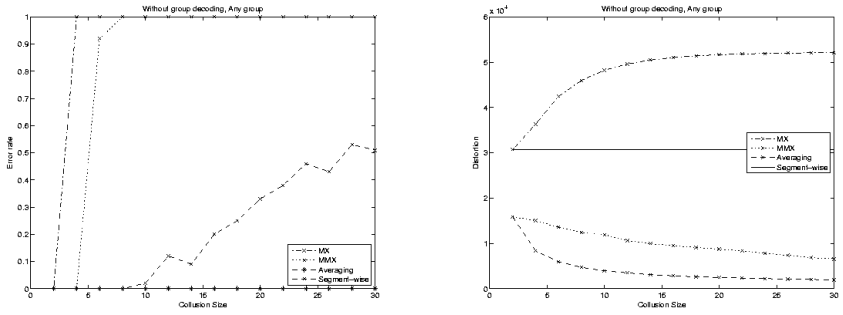
5.3 The Moderated Minority Extreme Attack

Figure 1 show a comparison of error rates and distortion levels for the MMX attack with different thresholds θ. Most interestingly, we note that for $\theta \geq 0.16$, we have less distortion than a fingerprinted copy prior to attack. For $\theta \leq 0.36$ we have more than 95% errors. If the pirates are drawn randomly across all groups, we get even more errors. Quite conservatively, we choose a threshold of $\theta = 0.4$ for our further simulations.



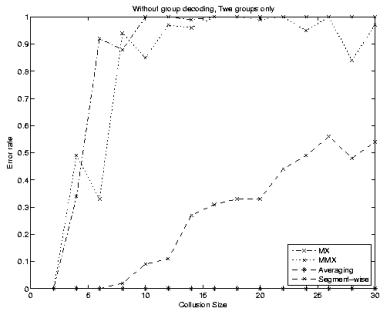
(a) Error rate (b) Distortion

Fig. 1: Comparison of different thresholds θ for the Moderated Minority Extreme attack with 10 pirates. Each data point is the average of 100 samples. Distortion is shown without group decoding. Error rates without group decoding, and with $\tau = 350$.

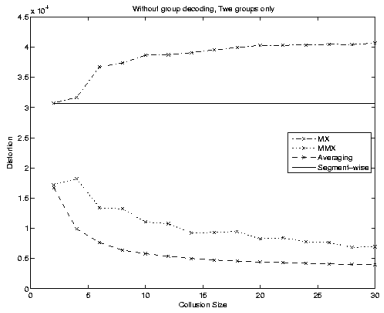


(a) Error rate (b) Distortion

Fig. 2: Comparison of different attacks with pirates drawn from any group. Averaging 100 samples per data point. The decoder uses no group decoding.

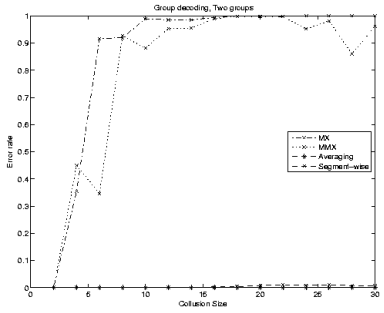


(a) Error rate

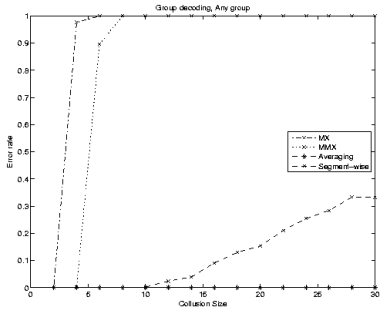


(b) Distortion

Fig. 3: Comparison of different attacks with pirates drawn from two groups only. Averaging 100 samples per data point. The decoder uses no group decoding.



(a) Two groups



(b) Any group

Fig. 4: Comparison of error rates under different attacks using group decoding with $\tau = 350$. Averaging 1000 samples per data point.

Comparison of the MX and MMX ($\theta = 0.4$) attacks with the attacks of [2] are shown in Figures 2, 3, and 4. The MX and MMX not only give much higher error rates than the old attacks, but they completely wreck the system with as few as eight pirates, giving error rates above 85%. This is a pure collusive attack, with no additional noise added.

The distortion is very much as expected. The distortion of the MX attack increase in the number of pirates, whereas the MMX and averaging attacks give decreasing distortion. Clearly the MMX attack give more distortion than the averaging attack, but yet much less than the fingerprinting caused in the first place.

5.4 Group decoding revisited

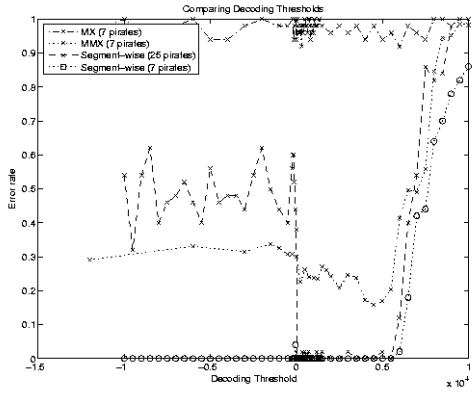
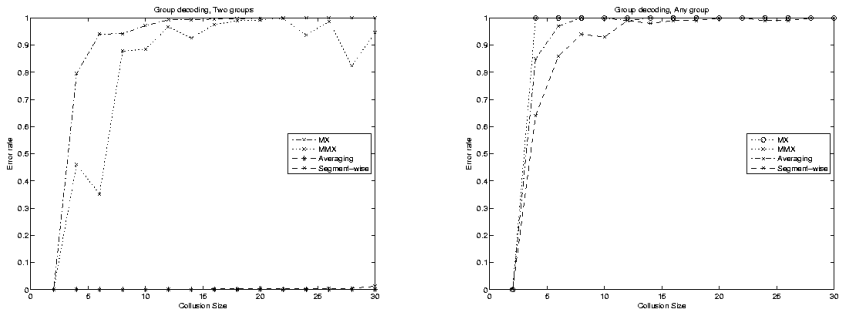


Fig. 5: Comparison of different group decoding thresholds τ under various attacks. Pirates drawn from two groups only. Using 50 samples per data point (1000 samples for MMX).

Figure 5 compares different group decoding thresholds against the MX and MMX ($\theta = 0.4$) attacks with seven pirates drawn from two groups. We observe that the group decoding threshold makes little difference under the MX attack. The sharp increase in errors starting at $\tau \approx 5500$ is due to all groups being excluded with high probability.

The figure shows that we should have $0 < \tau < 5500$, but otherwise the threshold seems to matter relatively little. There appears to be a slight dip in the error rate under the MMX attack around $\tau = 4500$, but the full simulation (Figure 6) for $t = 2 \ldots 30$ failed to confirm this. The error rates for $\tau = 350$ and $\tau = 4500$ were practically indistinguishable in the two-group case (over 1000 samples per value of t). In the any-group case, $\tau = 350$ gave clearly the better performance.



(a) Two groups (b) Any group

Fig. 6: Comparison of error rates under different attacks using group decoding with $\tau = 4500$. Averaging 1000 samples per data point.

6 Conclusion

We have taken the idea of minority choice attacks into the more practical model of joint WM/FP, and demonstrated that it can be effective. The result is an adaptive collusion attack which easily breaks the He-Wu joint WM/FP scheme (with the suggested parameters) while giving a hybrid copy which is less distorted than the original. One of the referees pointed out that He & Wu used Gaussian sequences in some of the tests. However, little detail was given for this variant, and we have not included it in our study at this stage.

The attack is quite general, and would be expected to give similar results for other fingerprinting schemes based on additive watermarking with correlation decoding (T statistic). Additional experiments would be needed to confirm this.

Later experiments have shown that the proposed attack is not effective against Gaussian fingerprints with Z statistic decoding and preprocessing [8]. However, the general point remains, that an adaptive attack more effective than the randomised attacks of [8] is likely to exist. To identify such adaptive attacks for decoding of [8] is an interesting open question.

It is interesting to note that past works on binary fingerprinting codes in abstract models, e.g. [1, 3, 5], never rely solely on closest neighbour or correlation decoders. Other combinatorial ideas are used in the inner code in order to resist attacks such as minority choice.

It would be interesting to construct joint WM/FP systems based on the collusion-secure codes from [3, 5]. It remains an open question whether it is possible to construct a joint WM/FP scheme which is both practical and secure against optimised attacks.

References

1. Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Trans. Inform. Theory*, 44(5):1897–1905, 1998. Presented in part at CRYPTO'95.
2. S. He and M. Wu. Joint coding and embedding techniques for multimedia fingerprinting. *IEEE Trans. Information Forensics and Security*, 1:231–248, June 2006.
3. Hans Georg Schaathun and Marcel Fernandez-Muñoz. Boneh-Shaw fingerprinting and soft decision decoding. In *Information Theory Workshop*, 2005. Rotorua, NZ.
4. Jessica N. Staddon, Douglas R. Stinson, and Ruizhong Wei. Combinatorial properties of frameproof and traceability codes. *IEEE Trans. Inform. Theory*, 47(3):1042–1049, 2001.
5. Gábor Tardos. Optimal probabilistic fingerprint codes. *Journal of the ACM*, 2005. `http://www.renyi.hu/~tardos/fingerprint.ps`. To appear. In part at STOC'03.
6. Neal R. Wagner. Fingerprinting. In *Proceedings of the 1983 Symposium on Security and Privacy*, 1983.
7. M. Wu, W. Trappe, Z. J. Wang, and K. J. R. Liu. Collusion resistant fingerprinting for multimedia. *IEEE Signal Processing Magazine*, 2004.
8. Hong Zhao, Min Wu, Z. June Wang, and K. J. Ray Liu. Nonlinear collusion attacks on independent fingerprints for multimedia. *IEEE Trans. Image Proc.*, pages 646–661, 2005.