

Fingerprinting with Minimum Distance Decoding

Shih-Chun Lin, Mohammad Shahmohammadi and Hesham El Gamal*

S. C. Lin is with Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan 10617. The work of S. C. Lin was supported by “Graduate Students Study Abroad Program” of National Science Council, Taiwan, R.O.C.

M. Shahmohammadi and H. El Gamal are with Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, 43210. This work was partly performed while Hesham El Gamal was visiting Nile University, Cairo, Egypt. The authors acknowledge the generous funding of the National Science Foundation, USA

E-mail: {lins, shahmohm, helgamal}@ece.osu.edu.

Fingerprinting with Minimum Distance Decoding

Abstract

This work adopts an information theoretic framework for the design of collusion-resistant coding/decoding schemes for digital fingerprinting. More specifically, the minimum distance decision rule is used to identify 1 out of t pirates. Achievable rates, under this detection rule, are characterized in two distinct scenarios. First, we consider the averaging attack where a random coding argument is used to show that the rate $1/2$ is achievable with $t = 2$ pirates. Our study is then extended to the general case of arbitrary t highlighting the underlying complexity-performance tradeoff. Overall, these results establish the significant performance gains offered by minimum distance decoding as compared to other approaches based on orthogonal codes and correlation detectors which can support only a sub-exponential number of users (i.e., a zero rate). In the second scenario, we characterize the achievable rates, with minimum distance decoding, under any collusion attack that satisfies the marking assumption. For $t = 2$ pirates, we show that the rate $1 - H(0.25) \approx 0.188$ is achievable using an ensemble of random linear codes. For $t \geq 3$, the existence of a *non-resolvable* collusion attack, with minimum distance decoding, for any non-zero rate is established. Inspired by our theoretical analysis, we then construct coding/decoding schemes for fingerprinting based on the celebrated Belief-Propagation framework. Using an explicit repeat-accumulate code, we obtain a vanishingly small probability of misidentification at rate $1/3$ under averaging attack with $t = 2$. For collusion attacks which satisfy the marking assumption, we use a more sophisticated accumulate repeat accumulate code to obtain a vanishingly small misidentification probability at rate $1/9$ with $t = 2$. These results represent a marked improvement over the best available designs in the literature.

EDICS WAT-FING

I. INTRODUCTION

Digital fingerprinting is a paradigm for protecting copyrighted data against illegal distribution [1]. In a nutshell, a *distributor*, i.e., the provider of copyrighted data, wishes to distribute its data \mathbb{D} among a number of licensed *users*. Each licensed copy is identified with a mark, which will be referred to as a *fingerprint* in the sequel, composed of a set of redundant digits embedded inside the copyrighted data. The locations of the redundant digits are kept *hidden* from the users and are only known to the distributor. Their positions, however, remain the same for all users.

If any user re-distributes its data in an unauthorized manner, it will be easily identified by its fingerprint. However, several users may collude to form a *coalition* enabling them to produce an unauthorized copy which is difficult to trace. In the literature, the colluding members are typically referred to as *pirates* or *colluders*. Hence, the need arises for the design of collusion-resistant digital fingerprinting techniques. Our work develops an information theoretic framework for the design of low complexity *pirate-identification* schemes.

To enable a succinct development of our results, we first consider the widely studied *averaging attack* [2]. The colluders, in this strategy, average their media contents to produce the forged copy. An explicit fingerprinting code construction for this attack was proposed in [2]. In this construction, however, the maximum number of users M , grows only polynomially with the fingerprinting code-length n (more precisely $M = O(n^2)$). Clearly, this rate of growth corresponds to a zero rate in the information theoretic sense. This motivates our pursuit for a fingerprinting scheme which supports an exponentially growing number of users, with the code-length, while allowing for low complexity pirate-identification strategies. Towards this goal, we use a random coding argument to establish the existence of a rate 0.5 *linear* fingerprinting code which achieves a vanishingly small probability of misidentification when 1) Only $t = 2$ pirates are involved in the averaging attack and 2) The low complexity minimum distance (MD) decoder is used to identify one of the two pirates. The enabling observation is the intimate connection between the scenario under consideration and the binary erasure channel (BEC). This result is then extended to the general case with an arbitrary coalition size t where the tradeoff between complexity and performance is highlighted.

Building on our analysis for the averaging attack, we then proceed to fingerprinting strategies which are resistant to more general forging techniques. More specifically, we adopt the *marking assumption* first proposed in [1]. In this framework, the pirates attempt to identify the positions occupied by the fingerprinting digits by comparing their copies. Afterwards, they can *only* modify the identified coordinates, in any desired way, to minimize the probability of traceability. The validity of the marking assumption hinges on the assumption that any modification to the data content \mathbb{D} will damage it permanently. This prevents the users from modifying any location in which they do not identify as a fingerprinting digit since it *may be* a data symbol. Boneh and Shaw [1] were the first to construct fingerprinting codes that are resistant to attacks that satisfy the marking assumption. This approach was later extended in [3] using the idea of separating

codes [4]. To the best of our knowledge, the best available explicit binary fingerprinting codes are the *low rate* codes presented in [3]. For example, for $t = 2$, the best available code has a rate ≈ 0.0092 . More recently, upper and lower bounds on the binary fingerprinting capacity for $t = 2$ and $t = 3$ were derived in [5]. The decoder used in [5], however, was based on exhaustive search, and hence, would suffer from an exponentially growing complexity in the code length. This prohibitive complexity motivates our proposed approach. In this paper, we show that using linear fingerprinting codes and MD decoding, one can achieve rates less than 0.188 when the coalition size is $t = 2$. Unfortunately, the proposed approach does not scale for $t \geq 3$. This negative result calls for a more sophisticated identification technique inspired by the analogy between our set-up and multiple access channels. Our results in this regard will be reported elsewhere.

Since the complexity of the *exact* MD decoder can be prohibitive when the code-length is long, we develop a low complexity belief-propagation (BP) identification approach [6][7]. This detector only requires a linear complexity in n , and offer remarkable performance gain over the best known explicit constructions for fingerprinting [3][2]. For example, we propose a modified iterative decoder tailored for the averaging attack with $t = 2$. Using this decoder along with an explicit repeat-accumulate (RA) fingerprinting code, we achieve a vanishingly small probability of misidentification for rates up to $1/3$. For the marking assumption set-up, we achieve a vanishingly small misidentification probability for rates up to $1/9$ using the recently proposed class of low rate accumulate repeat accumulate (ARA) codes [8]. It is worth noting that these results represent a marked improvement over the state of the art in the literature. Furthermore, one would expect additional performance enhancement by optimizing the degree sequences of the codes (which is beyond the scope of this work).

The rest of the paper is organized as follows. In Section II, we introduce the mathematical notations and formally define our problem setup. Then we explore the theoretical limits of fingerprinting using the MD decoder in Sections III and IV. The simulation results based on the BP framework are presented in Section V. Finally, Section VI offers some concluding remarks.

II. NOTATIONS AND PROBLEM STATEMENT

Throughout the paper, random variables and their realizations are denoted by capital letters and corresponding smaller case letters, respectively. Deterministic vectors are denoted by bold-face

letters. We denote the entropy function by $H(\cdot)$, with the argument being the probability mass function. Furthermore, for simplicity, we abbreviate $H(p, 1 - p)$ by $H(p)$, where $1 \geq p \geq 0$. For two functions of n , we write $a(n) \doteq b(n)$ if: $\lim_{n \rightarrow \infty} \frac{a(n)}{b(n)} = 1$, for example, $\binom{n}{d} \doteq 2^{nH(\frac{d}{n})}$. The Hamming distance between two vectors $\mathbf{x}_1, \mathbf{x}_2$ is denoted by $d_H(\mathbf{x}_1, \mathbf{x}_2)$. Without loss of generality, we assume that the number of users is M , and hence, a coalition U of size t is a subset of $\{1, 2, \dots, M\}$ where $|U| = t$. The goal of the coalition, in a nutshell, is to produce a forged fingerprint, \mathbf{y} , such that the distributor will not be able to trace it back to any of its members. In the following, we first introduce the notation that will be used for a general attack satisfying the marking assumption and then specify our notations for the averaging attack scenario. It should be noted that our formulation follows in the footsteps of [5]. For completeness, however, we repeat it here. As mentioned in [1], deterministic fingerprinting under the marking assumption is not possible in general. Therefore, the distributor needs to employ some kind of randomization which leads to a collection of binary codes (F, G) composed of K pairs of encoding and decoding functions as:

$$\begin{aligned} f_k &: \{1, 2, \dots, M\} \rightarrow \{0, 1\}^n \\ g_k &: \{0, 1\}^n \rightarrow \{1, 2, \dots, M\} \\ k &= 1, 2, \dots, K, \end{aligned} \tag{1}$$

where the code rate R is $\frac{\log_2 M}{n}$ and the secret key, k is a random variable employed to randomize the codebook. This way, the exact codebook utilized for fingerprinting is kept hidden from the users. It should be noted that, adhering to common conventions in cryptography, the family of encoding and decoding functions as well as the probability distribution of the secret key, $p(k)$, are known to all users. Finally, it is clear from the definition of g_k that the objective of the distributor, in our formulation, is to identify only one of the colluders correctly.

For simplicity of presentation, let's assume that $t = 2$ then the fingerprints corresponding to the coalition of users (also referred to as pirates or colluders), u_1, u_2 are denoted by $\{\mathbf{x}_1, \mathbf{x}_2\}$. The marking assumption implies that position i is *undetectable* to the two colluders if $x_{1i} = x_{2i}$, otherwise it is called *detectable* [1]. Those undetectable coordinates can not be changed by the pirates, and hence, the set of all possible forged copies is give by

$$E(U) = \{\mathbf{y} \in \{0, 1\}^n \mid y_i = x_{1i}, \forall i \text{ undetectable}\}. \tag{2}$$

In general, a coalition U may utilize a random strategy that satisfies the marking assumption to produce \mathbf{y} . That is, if $V(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2)$ is the probability that \mathbf{y} is created, given the coalition $\{\mathbf{x}_1, \mathbf{x}_2\}$, then we have:

$$V(\mathbf{y} \mid \mathbf{x}_1, \mathbf{x}_2) = 0 \quad \text{for all } \mathbf{y} \notin E(U). \quad (3)$$

In this paper, we focus on the maximum probability of misidentification over the set of all strategies which satisfy (3) (denoted by \mathcal{V} in the sequel). Similar to [5], we average the probability of misidentification over all possible coalitions leading to the following performance metric:

$$\bar{P}_m(F, G) := \frac{1}{\binom{M}{t}} \sum_U \max_{V \in \mathcal{V}} P_m(U, F, G, V), \quad (4)$$

where

$$P_m(U, F, G, V) := \mathbb{E}_K \left(\sum_{\mathbf{y} \in E(U), g_k(\mathbf{y}) \notin U} V(\mathbf{y} \mid f_k(U)) \right).$$

In the case of an averaging attack, we employ the typical assumption of mapping the binary fingerprints into the antipodal alphabets $\{-1, 1\}$ where the encoder now is defined as [2]

$$f : \{1, 2, \dots, M\} \rightarrow \{-1, +1\}^n. \quad (5)$$

As anticipated from the name, the forged copy is now given by:

$$\mathbf{y} = \frac{1}{t} \sum_{i=1}^t \mathbf{x}_i, \quad (6)$$

where the addition is over real field. The decoder is now defined as

$$g : \{\mathcal{A}_y\}^n \rightarrow \{1, 2, \dots, M\}, \quad (7)$$

where \mathcal{A}_y is the alphabets of \mathbf{y} , for example, it is $\{-1, 0, +1\}$ when $t = 2$. Misidentification will happen if $g(\mathbf{y}) \notin U$. Note that for $t = 2$, if $g(\mathbf{y}) \in U$, i.e., we trace one colluder correctly then we can always trace another colluder correctly according to (6). In this special case, the performance metric in (4) reduces to

$$\bar{P}_m^a := \frac{1}{\binom{M}{t}} \sum_U (g(\mathbf{y}) \notin U). \quad (8)$$

III. THE AVERAGING ATTACK

In this section, we investigate the theoretical achievable rate of fingerprinting code with the minimum distance (MD) decoder under the averaging attack. First, we need the following definition.

Definition 1: We say that the capacity of of an ensemble of fingerprinting codebooks \mathcal{E} is $R_{\mathcal{E}}$ under MD decoding if

- 1) For $M = 2^{nR}$ with $R < R_{\mathcal{E}}$, the average probability of misidentification over the ensemble P_m using MD decoding goes exponentially to zero as the codelength n goes to infinity.
- 2) Conversely, for $M = 2^{nR}$ with $R > R_{\mathcal{E}}$, there exists a constant $\delta > 0$ such that $P_m > \delta$ for sufficiently large block lengths.

Note that this converse in the previous definition is applicable only to a specific family of codes similar to the approach taken in [6], [7]. We also call a rate is MD-achievable if only the first part in Definition 1 is met. We are now ready to prove our first result.

Theorem 1: The fingerprinting capacity of the i.i.d codebook ensemble when $t = 2$ is $R_{\mathcal{E}} = 0.5$ (under the averaging attack and the MD decoder).

Proof: The encoder and decoder come as follows.

Encoder: The encoder chooses codewords uniformly and independently from all 2^n different vectors belonging to $\{0, 1\}^n$, transfers the fingerprinting codeword alphabets from $\{0, 1\}$ to $\{-1, +1\}$, and assigns the fingerprints to the users.

Decoder: With the given forged fingerprint \mathbf{y} , the decoder treats the position i where $\mathbf{y}_i = 0$ as an erased position, and the others as unerased positions. Let \mathcal{E} be the set of erasure positions and $\bar{\mathcal{E}} := [1 : n] \setminus \mathcal{E}$. Also let $\mathbf{y}_{\bar{\mathcal{E}}}$ denote those components of \mathbf{y} which are indexed by $\bar{\mathcal{E}}$. The decoder will search the codebook to find the codeword which agrees with \mathbf{y} in all unerased positions $\mathbf{y}_{\bar{\mathcal{E}}}$. Once the decoder finds such a codeword, the decoder declares it as the pirate. A misidentification occurs when the codeword of an innocent user \mathbf{z} is consistent with \mathbf{y} .

Achievability: For a small ε , we say the assigned fingerprints $\mathbf{x}_1, \mathbf{x}_2$ are *close* if $d_H(\mathbf{x}_1, \mathbf{x}_2) \leq n(\frac{1}{2} + \varepsilon)$, here the fingerprinting alphabets are $\{0, 1\}$ before transformation. As shown in Appendix

I-A, we know that with high probability, $(\mathbf{x}_1, \mathbf{x}_2)$ are a close pair. Thus, given a small $\epsilon > 0$,

$$|\mathcal{E}| \leq n\left(\frac{1}{2} + \epsilon\right), \quad (9)$$

since the erasures happen when the bits of $(\mathbf{x}_1, \mathbf{x}_2)$ are different. For the given forged fingerprint \mathbf{y} , \mathbf{z} must agree with \mathbf{y} in all $n - |\mathcal{E}|$ unerased positions, and can be -1 or $+1$ in the rest $|\mathcal{E}|$ erased positions. The probability of choosing such codeword is upper-bounded by

$$2^{n*(1/2+\epsilon)}/2^n. \quad (10)$$

By using the union bound, we know that for $R < 1/2 - \epsilon$, the probability of misidentification P_m tends to zero exponentially fast for sufficiently large codeword length n .

Converse: From (23) in the Appendix, we know that $P(|\mathcal{E}| \geq n/2) > P(|\mathcal{E}| = n/2) = \delta$, where δ is non-vanishing with respect to codeword length n . For a fingerprinting codeword \mathbf{x} , we form $\mathbf{x}_{\overline{\mathcal{E}}}$ as the components of \mathbf{x} which are indexed by $\overline{\mathcal{E}}$. And we arrange all $\mathbf{x}_{\overline{\mathcal{E}}}$ in the fingerprinting codebook as rows of a $2^{nR} \times (n - |\mathcal{E}|)$ array $\mathbf{X}_{\overline{\mathcal{E}}}$. The misidentification happens if $\mathbf{y}_{\overline{\mathcal{E}}}$ equals to more than two rows of $\mathbf{X}_{\overline{\mathcal{E}}}$. With $R > 1/2$, $|\mathcal{E}| \geq n/2$, and sufficiently large n ,

$$2^{nR} - 2 > 2^{(n-|\mathcal{E}|)} - 1. \quad (11)$$

And the misidentification will happen with probability at least $1/3$. From above, we know that if $R > 1/2$, the misidentification probability will be larger than $\delta/3$ for sufficiently large n which concludes the proof. □

Intuitively, the i.i.d generated codebook will result in $|\mathcal{E}| \approx n/2$ number of erased positions with high probability [5]. Then the ‘‘channel’’ between one of the pirates \mathbf{x}_1 and the forged fingerprint \mathbf{y} can be approximated by a binary erasure channel (BEC) with erasure probability $1/2$. From [9], we know that the capacity using the MD decoder of this channel is $1/2$. However, in the two-pirate fingerprinting system, there are always two codewords \mathbf{x}_1 and \mathbf{x}_2 in the codebook which meet the MD decoding criteria. This is the fundamental difference between this system and the classical BEC channel. In the BEC channel, with high probability, only one codeword will meet the MD decoding criteria. As will be presented in Section V-A, this difference will have an important implication on the design of Belief Propagation decoders for fingerprinting. The following result shows that restricting ourselves to the class of linear fingerprinting does

not entail any performance loss (at least from an information theoretic perspective)

Theorem 2: The fingerprinting capacity of the binary linear ensemble with $t = 2$ is $R_{\mathcal{E}} = 0.5$ (under the averaging attack and the MD decoder).

Proof: We consider the ensemble of binary linear codes of length n and dimension $n - l$ defined by the $l \times n$ parity check matrix H , where each entry of H is an i.i.d Bernoulli random variable with parameter $1/2$. The code rate $R = 1 - l/n$.

Encoder: The encoder chooses one codebook from this linear code ensemble, transfers the fingerprinting codeword alphabets from $\{0, 1\}$ to $\{-1, +1\}$, and assigns the fingerprints to the users.

Decoder: With the given forged fingerprint \mathbf{y} , again the decoder treats the position i where $y_i = 0$ as an erased position, and the others as unerased positions. The decoder will also transfer the alphabets of unerased positions from $\{-1, +1\}$ back to $\{0, 1\}$. Let $H_{\mathcal{E}}$ denote the submatrix of H that consists of those columns of H which are indexed by the set of erasures \mathcal{E} . In a similar manner, let $\mathbf{x}_{\mathcal{E}}$ denote those components of the pirate's fingerprint which are indexed by \mathcal{E} , and $\mathbf{x}_{\overline{\mathcal{E}}}$ denote those components which are indexed by $\overline{\mathcal{E}}$. In the following, we assume that the fingerprinting codeword alphabets are transferred back to $\{0, 1\}$ and the addition is module-2. Note that the true pirates \mathbf{x}_1 and \mathbf{x}_2 will result in the same $\mathbf{x}_{\overline{\mathcal{E}}} = \mathbf{y}_{\overline{\mathcal{E}}}$, where $\mathbf{y}_{\overline{\mathcal{E}}}$ is defined as in Theorem 1. From the parity check equations,

$$H_{\mathcal{E}}\mathbf{x}_{\mathcal{E}}^{\text{T}} = \mathbf{s}^{\text{T}}, \quad (12)$$

where $\mathbf{s}^{\text{T}} := H_{\overline{\mathcal{E}}}\mathbf{y}_{\overline{\mathcal{E}}}^{\text{T}}$ is called the syndrome. The syndrome is known at the decoder. The decoder solves these linear equations to find $\mathbf{x}_{\mathcal{E}}$, combines it with the known $\mathbf{x}_{\overline{\mathcal{E}}} = \mathbf{y}_{\overline{\mathcal{E}}}$, and declares one of the results as the pirate.

Achievability: We know that (12) has at least two solutions corresponding to the true pirates \mathbf{x}_1 and \mathbf{x}_2 . The rank of $l \times |\mathcal{E}|$ matrix $H_{\mathcal{E}}$ must equal to $|\mathcal{E}| - 1$ to make sure that there is only two solutions. The decoder will declare an innocent user as the pirate if there are more than two

solutions, iff $H_{\mathcal{E}}$ has rank less than $|\mathcal{E}| - 1$. This happens with probability

$$1 - \frac{M_b(l, |\mathcal{E}|, |\mathcal{E}| - 1)}{2^{|\mathcal{E}|} - M_b(l, |\mathcal{E}|, |\mathcal{E}|)}, \quad (13)$$

where $M_b(l_1, m_1, k_1)$ denote the number of binary matrices with dimension $l_1 \times m_1$ and rank k_1 .

To make (13) approach zero as n increases, the second term in (13) must approach one as n goes up. To show this, we first assume that $|\mathcal{E}| + n\epsilon_1 \leq l$, where $\epsilon_1 > 0$ is a small number. And according to (28) in Appendix II and [10], the second term in (13) equals

$$\frac{M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 1)(2^{|\mathcal{E}|} - 1)}{2^{|\mathcal{E}|} - M_b(|\mathcal{E}|, l, |\mathcal{E}|)}. \quad (14)$$

From [10], for $j = 0 \dots |\mathcal{E}| - 1$

$$M_b(|\mathcal{E}| - j, l, |\mathcal{E}| - j) = \prod_{p=0}^{|\mathcal{E}|-j-1} (2^l - 2^p). \quad (15)$$

Using this formula in (14) and dividing the nominator and denominator by $M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 1)$, this term equals

$$\frac{2^{|\mathcal{E}|} - 1}{2^{(|\mathcal{E}|-1)} + 2^l [-1 + \prod_{p=0}^{|\mathcal{E}|-2} 1/(1 - 2^{p-l})]}. \quad (16)$$

Note that $n\epsilon_1 \leq l - |\mathcal{E}|$, each 2^{p-l} approaches zero exponentially fast with n . By using Taylor series on $1/(1 - 2^{p-l})$, and with some simplifications, the denominator becomes

$$2^{(|\mathcal{E}|-1)} + \sum_{p=0}^{|\mathcal{E}|-2} 2^p + 2^{|\mathcal{E}|} * h.o.t. = 2^{|\mathcal{E}|} * (1 + h.o.t.) - 1, \quad (17)$$

where the higher order terms of the Taylor series are denoted by *h.o.t* and approach zero exponentially fast. Using this result in (16), our claim is valid and (13) approaches zero as $n \rightarrow \infty$ if $|\mathcal{E}| + n\epsilon_1 \leq l$.

As shown in Appendix I-B, $|\mathcal{E}| \leq n(1/2 + \epsilon)$ with high probability, we know that if $n(1/2 + \epsilon) + n\epsilon_1 \leq l$, or $R < 1/2 - (\epsilon + \epsilon_1)$, the probability of misidentification can be made arbitrary small.

Converse: From (26) in Appendix, we know that $P(|\mathcal{E}| \geq n/2) > P(|\mathcal{E}| = n/2) = \delta$, where δ is non vanishing with respect to codeword length n . With $R > 1/2$ and sufficiently large n , $P(|\mathcal{E}| - 1 > l) \geq \delta$. In this case, the rank of $H_{\mathcal{E}}$ is less than $|\mathcal{E}| - 1$ and the syndrome decoder will find at least three solutions of equation (12). The misidentification will happen with probability at least $1/3$ since. From above, we know that if $R > 1/2$, the probability will

be larger than $\delta/3$ for sufficiently large n and it concludes the proof. □

Next, our approach is generalized to coalitions with $t > 2$. The key to the following corollary is to treat all alphabets other than ± 1 in \mathcal{A}_y of (7) as erasures.

Corollary 1: The rate $\frac{1}{2^{(t-1)}}$ is MD-achievable for fingerprinting under average attack with a coalition of size t .

Proof: The encoder/decoder are the same as the ones in Theorem 1 except for the choices of erasure positions as described previously. Note that $y_i \neq \pm 1$ whenever the pirates' fingerprints bits are not the same at position i . Similar to [5], we know that with high probability, the i.i.d generated codebooks will meet

$$|\mathcal{E}| \leq n \left\{ 1 - \frac{1}{2^{(t-1)}} + \epsilon \right\}.$$

Then, following in the footsteps of the proof of Theorem 1 we obtain our result. □

The advantage of the MD decoder, used to obtain the previous result, is the universality for all t . However, for each t , we can obtain higher rates by tailoring our encoder/decoder to this specific case. To illustaret the idea, let's consider the $t = 3$ case. Now, $\mathcal{A}_y = \{\pm 1, \pm \frac{1}{3}\}$ and one can achieve better performance by exploiting the information contained in the positions with $y_i = \pm \frac{1}{3}$.

Theorem 3: The rate $H(\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8}) - H(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}) = 0.3113$ is achievable for fingerprinting under average attack with $t = 3$.

Proof: The encoder is the same as Theorem 1. As for the decoder, we first define X as a random variable with $P(X = \pm 1) = 1/2$, and the random variable $Y = (X + X_2 + X_3)/3$, where X_2, X_3 has the same distribution as X and (X, X_2, X_3) are independent. The transition matrix of $P(Y|X)$ is

Typically, we need a maximum likelihood (ML) decoder designed for the transition matrix $P(Y|X)$. Note that when $t = 2$, this decoder reduces to the one specified in Theorem 1. However,

$x \backslash Y$	-1	-1/3	1/3	1
-1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0
1	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

it is hard to investigate the performance of the ML decoder, and we use the jointly-typical decoder defined in [9] as a lower-bound for the achievable rate of this decoder. Given a forged fingerprint \mathbf{y} , the decoder search the codebook to find the codeword such that this codeword and \mathbf{y} are jointly-typical with respect to $P(X, Y)$. Once the decoder finds such a codeword, the decoder declares it as the pirate.

Achievability : Without loss of generality, we can assume that the pirates indices are $(1, 2, 3)$. An event E_i occurs when the i th codeword and \mathbf{y} are jointly typical, and the event E_i^c is its complement. Then the probability of misidentification P_m is upper-bounded by

$$P_m \leq P(E_1^c) + P(E_2^c) + P(E_3^c) + \sum_{i \neq 1,2,3} P(E_i).$$

From [9, Theorem 15.2.1], the first three terms can be made less than any arbitrary small $\epsilon > 0$ for sufficiently large n . And the last term is upper-bounded by

$$(M - 3)2^{-n(I(X;Y) - 4\epsilon)},$$

So if $R < I(X;Y) - 4\epsilon$, P_m can be made arbitrary small for sufficiently large n . According to the transition matrix of $P(Y|X)$, we know that

$$I(X;Y) = H\left(\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8}\right) - H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right),$$

which concludes the proof □

IV. THE MARKING ASSUMPTION

Having studied the special case of averaging attack, we now proceed to the case when the coalition can employ *any* strategy as long as the marking assumption is satisfied. The following result establishes the achievable rate of random fingerprinting codes with MD decoding

Theorem 4: For all rates less than $1 - H(0.25)$ there exists an MD-achievable fingerprinting code, when $t = 2$.

Proof: We use a random coding argument to prove our result. We construct the following ensemble of binary random codes as in Theorem 1: Binary random vectors (fingerprints) of length n are assigned to the $M = 2^{nR}$ users where each coordinate is chosen independently with equal probability of being 0, 1. For a small ε , we say the assigned fingerprints $\mathbf{x}_1, \mathbf{x}_2$ are close if $d_H(\mathbf{x}_1, \mathbf{x}_2) \leq n(\frac{1}{2} + \varepsilon)$. If the pair $(\mathbf{x}_1, \mathbf{x}_2)$ is close we denote it by $\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2$, otherwise for a non-close pair we write: $\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2$. Given a forged fingerprint \mathbf{y} , the average probability of misidentification over this ensemble can be upper bounded by:

$$P_m(\mathbf{y}|\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2) + P(\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2),$$

where $P_m(\mathbf{y}|\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2)$ is the misidentification probability when \mathbf{y} is produced by a close pair $(\mathbf{x}_1, \mathbf{x}_2)$ and $P(\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2)$ is the probability that the pirates did not constitute a close pair. Both probability are averaged over the random coding ensemble. By the following argument, we will show that these probabilities goes exponentially to zero as n goes to infinity hence the proof.

In Appendix I-A we have proved that $P(\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2)$ goes to zero as n goes to infinity. Now we turn to $P_m(\mathbf{y}|\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2)$. Since $d_H(\mathbf{x}_1, \mathbf{x}_2) < n(\frac{1}{2} + \varepsilon)$, the Hamming distance of the forged copy \mathbf{y} with at least one of the pirates must be less than $h(n) := n(\frac{1}{4} + \frac{\varepsilon}{2})$ due to the marking assumption. Without loss of generality, we assume this pirate to be \mathbf{x}_1 . Using minimum Hamming distance decoding, misidentification occurs if there is another binary vector \mathbf{z} of length n in the codebook such that $d_H(\mathbf{y}, \mathbf{z}) \leq d_H(\mathbf{y}, \mathbf{x}_1)$. The total probability of this event in the random ensemble is upper-bounded by

$$\frac{M \sum_{i=1}^{h(n)} \binom{n}{i}}{2^n} \doteq M * 2^{-n(1-H(0.25))},$$

where the union bound is used. The probability of misidentification in a random code of size $M = 2^{nR}$ is at most

$$2^{-n(1-H(0.25)-R)}.$$

The above probability goes exponentially to zero as $n \rightarrow \infty$ for all rates $R < 1 - H(0.25)$.

□

Intuitively, with a high probability, the forged copy will be produced by a pair of *close* pirates. Therefore, the minimum Hamming distance between the pirates \mathbf{x}_1 and the forged copy \mathbf{y} is approximately $n/4$ implying that we can treat the "channel" between them as a binary symmetric channel (BSC) with crossover probability $1/4$ (whose capacity is $1 - H(0.25)$ [9]). Next, we extend our result to binary linear codes

Theorem 5: For all rates less than $1 - H(0.25)$, there exists a *linear* MD-achievable fingerprinting code, when $t = 2$.

Proof: Consider the ensemble of binary linear codes with binary parity generator matrix G where elements of G are chosen equally and independently from $\{0, 1\}$ similar to Theorem 2. The size of matrix G is $(n - l) \times n$, with rate $R = (n - l)/n$ and the codeword length n . It should also be noted that in the following all matrix multiplications and additions are done in module-2 unless otherwise stated. In order to randomize the codebook, the distributor employs the following strategy: Generating the secret key vectors as independent binary random vectors of length n , whose coordinates are chosen to be 0, 1 independently with probability $1/2$. We denote the vector indexed by secret key k as \mathbf{k} . The vector \mathbf{k} is added in the binary domain to the codeword, and the resulting vector is assigned to the corresponding user. Note that this operation will not change the detectable positions, where the codewords are the different. With forged copy \mathbf{y} , the decoder subtracts \mathbf{k} and performs MD decoding. As we mentioned earlier, the secret key is unknown to the users and is only known to the distributor.

Similar to the proof of Theorem 4, we can upper-bound the probability of misidentification as

$$P_m(\mathbf{y}|\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2) + P(\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2). \quad (18)$$

In Appendix I-B we have established that over the ensemble of linear random codes described above, $P(\mathbf{x}_1 \stackrel{N}{\leftrightarrow} \mathbf{x}_2)$ also goes to zero as the code length goes to infinity. Now let us consider $P_m(\mathbf{y}|\mathbf{x}_1 \stackrel{C}{\leftrightarrow} \mathbf{x}_2)$. The codes assigned to the users which are the result of the addition of a secret key to a linear code can be written as:

$$\mathbf{u}G + \mathbf{k} \quad (19)$$

where \mathbf{u} is an information message vector. Notice that the ensemble defined by (19) is the same as ensemble of *coset codes* introduced in [11]. In our proof, we need the following lemmas for the coset codes ensemble that are proved in [11].

Lemma 1: The probability of any binary vector \mathbf{v} being a codeword in the ensemble defined by (19) is equal to 2^{-n} .

Lemma 2: Let $\mathbf{v}_1, \mathbf{v}_2$ be the codewords corresponding to two different information sequences $\mathbf{u}_1, \mathbf{u}_2$. Then over the ensemble of codes, $\mathbf{v}_1, \mathbf{v}_2$ are statistically independent.

Similar to the proof of Theorem 4, again due to the marking assumption we can assume $d_H(\mathbf{y}, \mathbf{x}_1) < h(n)$. Using MD decoding, misidentification occurs if there is another binary vector \mathbf{z} of length n in the codebook such that $d_H(\mathbf{y}, \mathbf{z}) \leq d_H(\mathbf{y}, \mathbf{x}_1)$. The total number of binary vectors for which $d_H(\mathbf{y}, \mathbf{z}) \leq d_H(\mathbf{y}, \mathbf{x}_1)$ can be upper bounded by: $\sum_{i=1}^{h(n)} \binom{n}{i} \doteq 2^{nH(0.25)}$. By Lemma 1 and Lemma 2 over the ensemble each of such vectors \mathbf{z} is independent of \mathbf{x}_1 with probability 2^{-n} . Therefore, the total probability of this event in the ensemble is upper-bounded by:

$$M * 2^{-n(1-H(0.25))},$$

where again the union bound is used. The probability of misidentification in a random coset code of size $M = 2^{nR}$ is at most

$$2^{-n(1-H(0.25)-R)}.$$

The above probability goes exponentially to zero as $n \rightarrow \infty$ for all rates $R < 1 - H(0.25)$. □

When the coalition size, t is larger than two, the minimum distance decoding will fail due to the following argument. Let $t = 3$ and assume that the forged copy is produced by

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3,$$

where the additions are modulo-2. It is easy to check that this attack satisfies the marking assumption. For $t > 3$ the coalition can consider only three of the pirates, ignore the rest and apply this attack. Following the footsteps in the proof of Theorem 3, it is easy to see that the MD-achievable rate is zero. Indeed, it can also be shown that the resulting ‘‘BSC channel’’ has crossover probability $1/2$, and this negative result is obtained [9].

V. BELIEF PROPAGATION FOR FINGERPRINTING

Implementing the exact minimum distance decoder may require prohibitive complexity (especially for large codeword lengths). This motivates our approach of using the BP framework to approximate the MD decoder. More specifically, in this section, we present explicit constructing of graph-based codes, along with the corresponding BP decoders, which are tailored for the fingerprinting application.

A. Averaging attack

As remarked earlier, the two-pirate averaging attack will produce a “channel” *almost equivalent* to the classical BEC. This inspires the use of graphical codes based on the Repeat Accumulate (RA) framework [12], such as the nonsystematic irregular RA code of [13] and the irregular ARA code of [14], which were shown to be capacity achieving for the BEC. In our simulations, we use the original regular RA codes of [12] due to their simplicity and good performance for low rate scenarios. It is worth noting that all the techniques discussed in the sequel can be applied directly to the irregular codes presented in [13], [14]. For the sake of completeness we review briefly the encoding procedure for regular RA codes: first, the information bits are repeated a constant number of times (by a regular repetition code) and interleaved. The interleaved bits are then accumulated to generate the code symbols. Similarly, one can employ the standard BP iterative decoding approach [15] to identify the pirates. However, we argue next that significant performance improvement can be obtained via a key modification to the iterative decoder*.

It is well known that the standard iterative algorithm will fail if a stopping set exists in the erased positions [10]. Unfortunately, a stopping set always will exist in the erased positions produced by averaging attack. To see this, it is more convenient to represent the RA code using the appropriate bipartite Tanner graph containing a set of variables $\mathcal{V} = \{v_1, v_2, \dots\}$ and a set of check nodes. The reader are referred to [12], [13], [14] for more details on the graphical representation of RA codes. A stopping set \mathcal{S} is, therefore, a subset of \mathcal{V} , such that all neighbors of \mathcal{S} are connected to \mathcal{S} at least twice. The standard BP algorithm can now be stated as the follows.

[Standard BP]:

*in the following, the fingerprinting codeword alphabets are $\{0, 1\}$ after decoder transformation and the addition is module-2.

- 1) Find a check node that satisfies the following
 - This check node is not labelled as “finished”.
 - The values of all but one of the variable nodes connected to the check node are known. Set the value of the unknown erased one to be the module-2 addition of the other variable nodes. And label that check node as “finished”.
- 2) Repeat step 1 until all check nodes are labeled as “finished” or the decoding cannot continue further. If the latter happens, declare the decoding fail.

It is now easy to see that, in the stopping set, every check node is connected to at least two erased variable nodes and the decoder will halt at this point. The following result establishes the limitation of the standard BP decoder in our fingerprinting scenario

Proposition 1: Let \mathcal{V}_{B1} and \mathcal{V}_{B2} be the set of values of the variable node set \mathcal{V} corresponding to pirate fingerprints \mathbf{x}_1 and \mathbf{x}_2 , respectively. And let \mathcal{V}_d be the set of variable nodes where the corresponding values in \mathcal{V}_{B1} and \mathcal{V}_{B2} are different. Then \mathcal{V}_d is a stopping set.

Proof: This statement is proved by contradiction. First we assume that \mathcal{V}_d is not a stopping set. It means that $\exists j \in \bigcup_{i \in \mathcal{V}_d} N(i)$ where the check node j has only one neighbor i' in \mathcal{V}_d . Here we denote the neighbor of node i in the graph as $N(i)$. For the neighboring variable nodes of this check node, we have

$$\begin{cases} \mathcal{V}_{B1}(i) = \mathcal{V}_{B2}(i) & \forall i \neq i', i \in N(j) \\ \mathcal{V}_{B1}(i') = \mathcal{V}_{B2}(i') + 1 & i' \in N(j). \end{cases} \quad (20)$$

However, from the check equation of this check node

$$\sum_{i \in N(j)} \mathcal{V}_{B1}(i) = \sum_{i \in N(j)} \mathcal{V}_{B2}(i) = 0, \quad (21)$$

where the addition is module-2. It is obvious that (20) contradicts with (21) since the total number of variable nodes such that $\mathcal{V}_{B1}(i) \neq \mathcal{V}_{B2}(i), i \in N(j)$ should be even. Thus \mathcal{V}_d is a stopping set.

□

Since, under averaging attack, the bits of the forged fingerprint will be erased whenever the pirate fingerprints are different in the Tanner graph, then \mathcal{V}_d will be always contained in the erased

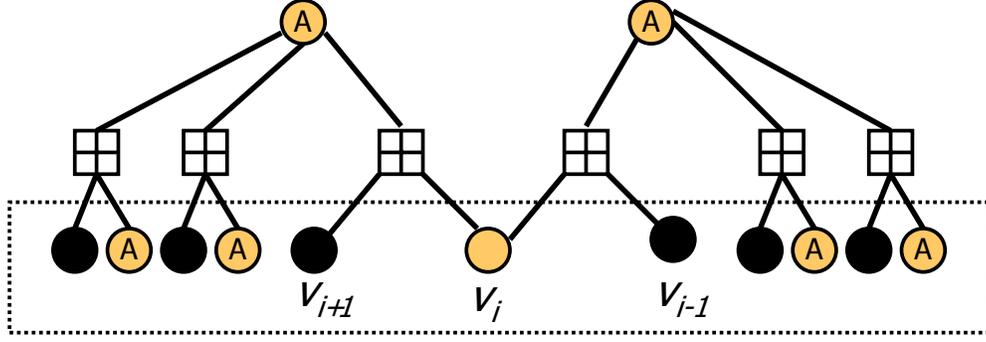


Fig. 1. Proper variable node to be chosen in step 2 of proposed modified BP algorithm for two-pirate averaging attack.

positions and the iterative decoder will fail. The modification, presented next, will break the stopping set \mathcal{V}_d , and hence, allow the iterative decoder to proceed forward. The key observation is that for every erased position in \mathcal{V}_d , the pirate fingerprints can only be represented by only two combinations $\{0, 1\}$ or $\{1, 0\}$. It allows us to choose one variable node in this stopping set, and set its value to 1. The modified forged fingerprint will then be “closer” to one of the pirate fingerprints. In summary, the decoder becomes

[Modified BP for fingerprinting]:

- 1) Perform the standard BP algorithm, remove all the “finished” labels and *Go to* step 2
- 2) Choose a proper variable node in \mathcal{V}_d (different from previous choices), and set its value to 1. If the decoder has executed this step more than N_{max} times, declare a decoding failure and exit.
- 3) Run the standard BP on the new graph. If the decoder fails, reset the variable nodes to their original values and *Go to* step 2.

In step 2, we must make sure that the chosen variable node breaks the stopping set \mathcal{V}_d . The neighboring variables nodes of a degree-3 check node in RA code are good choices. From the check equations in (21), the erased variables nodes will appear in pair. If we set the value of one of the two erased neighbor variable node v_i as 1, this degree-3 check node is connected to $\mathcal{V}_d \setminus v_i$ with only one edge. Then $\mathcal{V}_d \setminus v_i$ is not a stopping set. We also need to choose the variable node which will affect as much other variable nodes in $\mathcal{V}_d \setminus v_i$ as possible by setting its

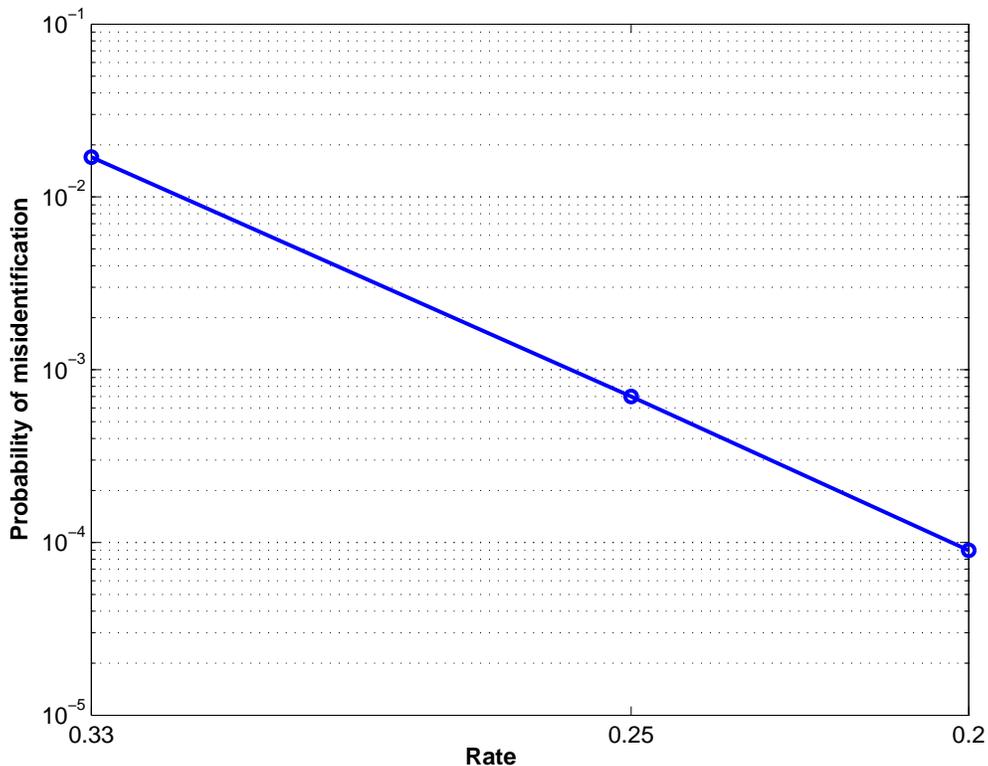


Fig. 2. Probability of misidentification under two-pirate averaging attack using RA codes with different rates and modified BP algorithm without variable node selection.

value. Since all check nodes of RA code are degree-3, we choose such variable node v_i in the degree-2 variable-node-chain of RA code, as shown in Fig. 1. The check node is depicted as \boxplus , the unerased variable nodes as black circle and the erased ones as hollow circle. Furthermore, each variable node which will benefit from guessing v_i is shown as hollow circle with the letter “A” in the figure. The key observation is that, for node v_i , the two neighboring accumulator output nodes, i.e., v_{i-1} and v_{i+1} , correspond to non-erased bits. This implies that that setting the value of v_i will at least affect 6 other variable nodes of rate $1/3$ RA code.

Now, we are ready to report our simulation results. First, we show the performance of proposed algorithm with different rate RA codes without variable node selection in Fig 2 (i.e., we select the first unerased variable node in the RA degree-2 variable-node-chain and set $N_{max} = 1$). Here, the number of information bits $n/R = 16384$ is fixed for all rates, to make the number of users M the same. We observe that, without selecting the variable node as shown in Fig 1, the

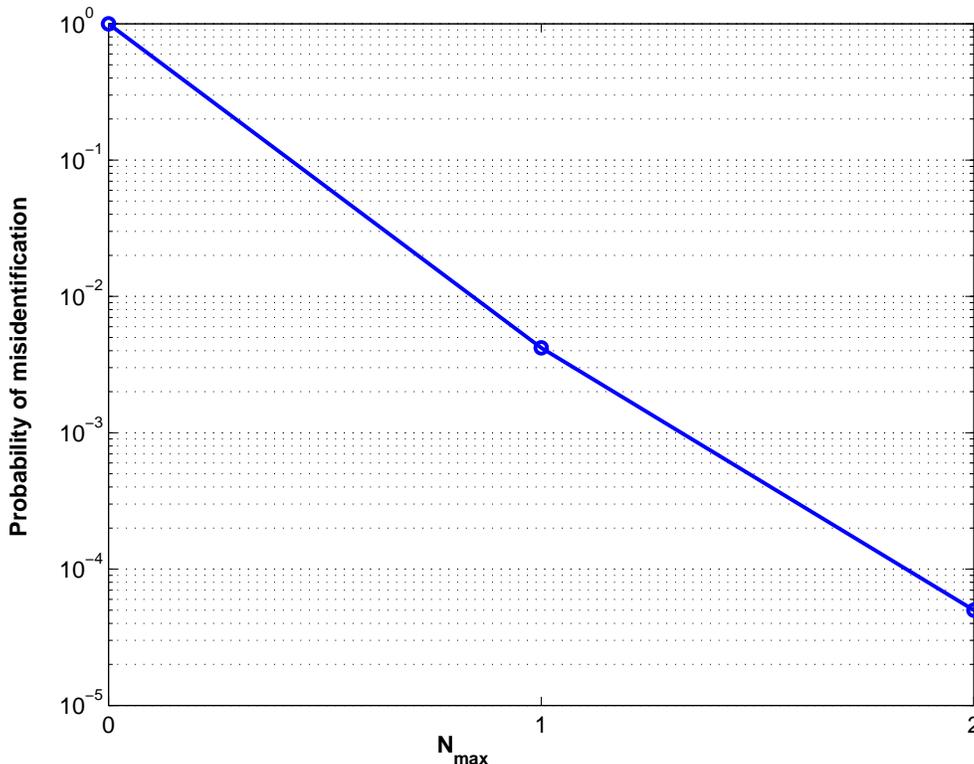


Fig. 3. Probability of misidentification under two-pirate averaging attack using rate 1/3 RA code and modified BP algorithm with different N_{max} .

probability of misidentification \bar{P}_m^a is high for rate 1/3. This performance can be improved by the proposed algorithm for variable node selection and increasing N_{max} as depicted in Fig. 3. Finally, in Fig. 4 we report \bar{P}_m^a with different code length n and $N_{max} = 2$.

Finally, we note that our algorithm is similar, in spirit, to the proposed guessing algorithm in [7]. The critical difference is that the structure of our problem ensures that the guessed bit always corresponds to one of the pirates, and hence, we do not need to worry about the possibility of contradictions as the iteration proceeds.

B. The Marking Assumption: The Memoryless Attack

In this subsection, we report our simulation results for the two-pirate memoryless attack. In this attack, when the pirates encounter a detectable position, they choose 0, 1 independently and with equal probability to form the forged copy. We use rates 1/8, 1/9 and 1/10 ARA codes based on the low rate protographs presented in [8]. The protographs of the codes are depicted

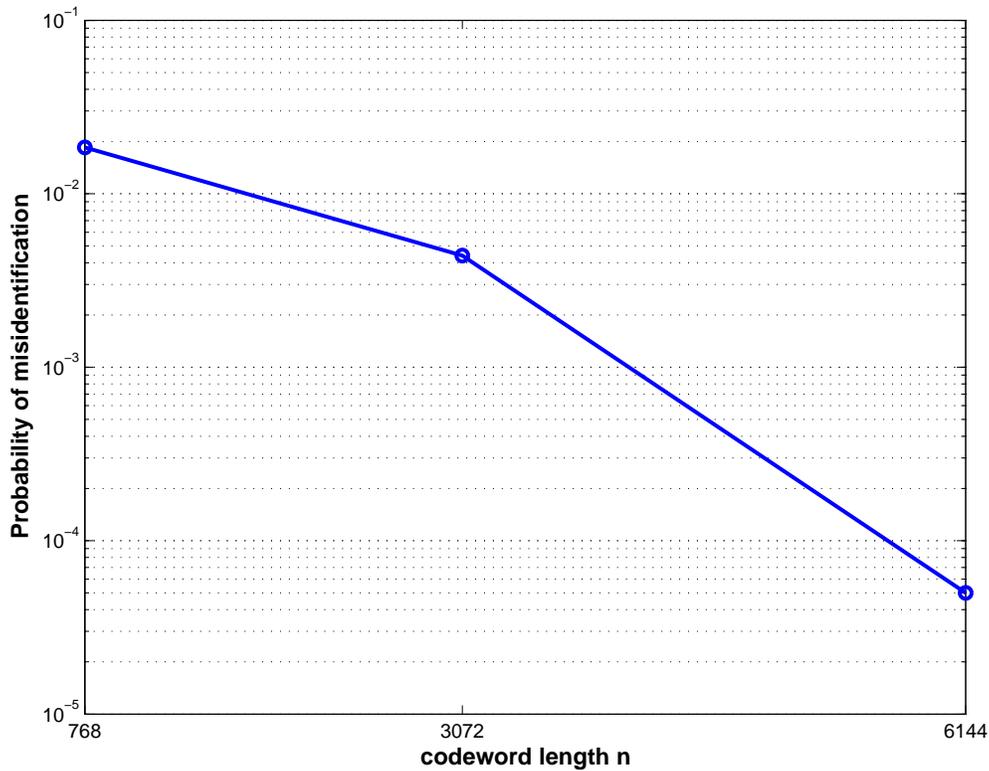


Fig. 4. Probability of misidentification under two-pirate averaging attack using rate 1/3 RA code and modified BP algorithm with different code lengths n .

in Fig 5. For a formal description of the ARA codes, we refer the interested readers to [8], [14] and references therein. Decoding is done iteratively using the BP framework with a maximum number of iterations equal to 60. Here, the decoder treats the forged fingerprint as the output of a BSC with crossover probability equal to 0.25. In Fig 6, the probability of misidentification \bar{P}_m is depicted versus different code lengths for different rates. As shown in the figure, it is clear a vanishing small misidentification probability is achievable for rate 1/9 which is about an order of magnitude higher than the best result available in the literature for explicit fingerprinting codes.

VI. CONCLUSION

This paper developed an information theoretic framework for the design of low complexity coding/decoding techniques for fingerprinting. More specifically, we established the superior performance of the minimum distance decoder and validated our theoretical claims via explicit

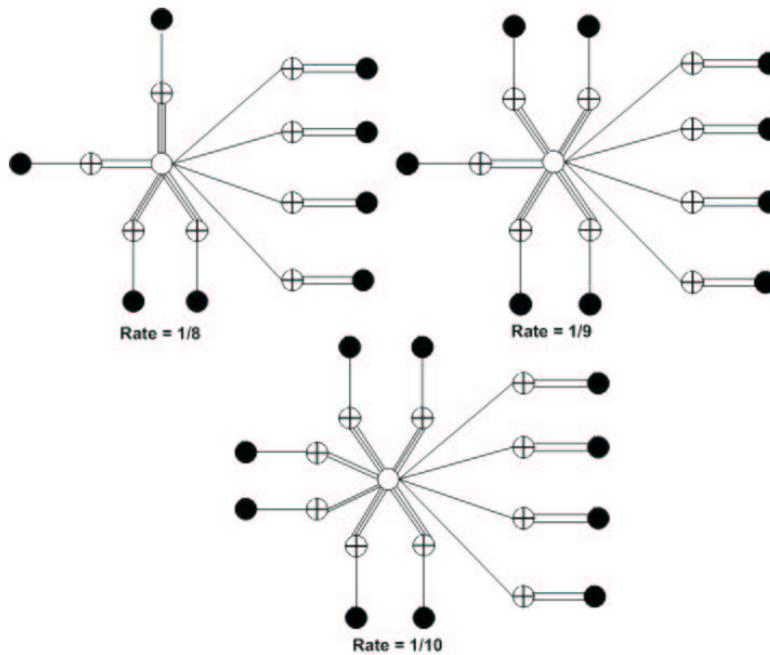


Fig. 5. Protographs of rate 1/8, 1/9, 1/10 ARA codes.

construction of BP encoding/decoding schemes. In the averaging attack scenario, our framework was inspired by the equivalence between our problem and the BEC. We also showed that the worst case attack, under the marking assumption, is equivalent to a BSC with a cross-over probability equal to $1/4$. Our approach for the averaging attack can handle arbitrary coalition sizes, whereas it was shown that the MD decoder recover from marking assumption attacks only with coalitions composed of two pirates. This negative result motives our current investigations on more sophisticated approaches for pirate tracing using the intimate connection between collusion in digital fingerprinting and multiple access channels.

APPENDIX I

ON NON-CLOSE PAIRS IN RANDOM ENSEMBLE

We will examine the probability of non-close pairs for random i.i.d and linear codebook ensembles, and show that these events will not happen with high probability.

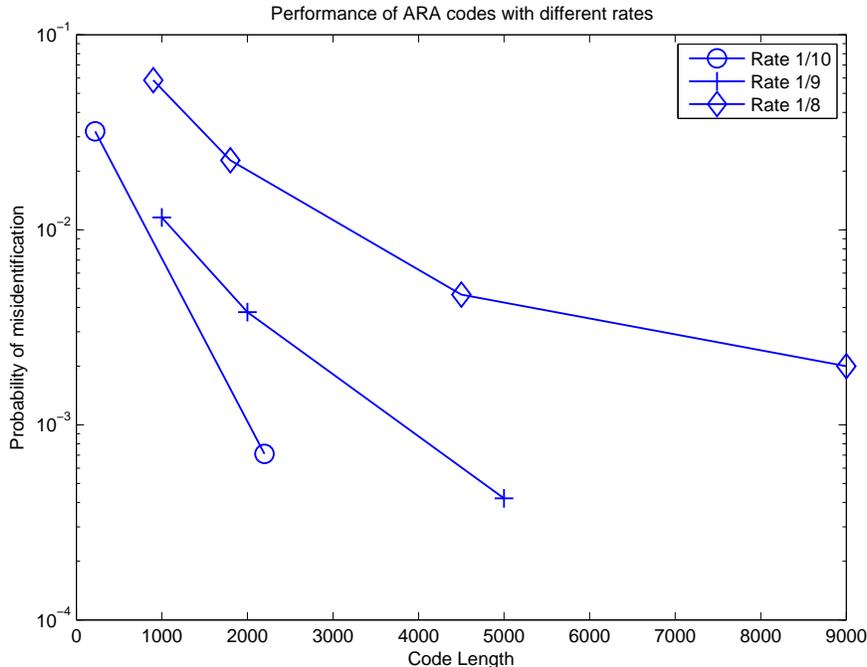


Fig. 6. Probability of misidentification for ARA codes with different rates and code lengths, under two-pirate memoryless attack.

A. i.i.d codebook ensemble

For a codebook C in the i.i.d ensemble and $1 \leq d \leq n$, define the number of unordered pairs of codewords $(\mathbf{x}_i, \mathbf{x}_j)$ with $i \neq j$ in C at distance d apart as

$$S_c(d) := \sum_{i=1}^M \sum_{j=1}^{i-1} \Phi\{d_H(\mathbf{x}_i, \mathbf{x}_j) = d\}, \quad (22)$$

where $\Phi(\cdot)$ is the indicator function. In [16], it is established that with probability going to one as $n \rightarrow \infty$

$$S_c(d) \doteq \begin{cases} 2^{n(2R+H(\frac{d}{n})-1)} & n\delta_{GV}(2R) < d < n(1 - \delta_{GV}(2R)) \\ 0 & \text{otherwise,} \end{cases} \quad (23)$$

where $\delta_{GV}(\cdot)$ is the Gilbert-Varshamov distance which for $0 < R < 1$, $\delta_{GV}(R)$ is defined as the root $\delta < 0.5$ of the equation $H(\delta) = 1 - R$. And $\delta_{GV}(R)$ is zero for $R \geq 1$. Using (23), we can write the probability of non-close pairs in the codes of the random ensemble as

$$\frac{\sum_{d > n(1/2+\epsilon)} 2^{n(2R+H(d/n)-1)}}{2^{2nR}} < \frac{n2^{n(2R-1+H(\frac{1}{2}+\epsilon))}}{2^{2nR}}. \quad (24)$$

which goes exponentially to zero as $n \rightarrow \infty$.

B. Random binary linear codebook ensemble

For a code C in the linear ensemble and $1 \leq d \leq n$ by the symmetry of linear codes we can write

$$S_c(d) = \sum_{i=1}^M \sum_{j=1}^{i-1} \Phi\{d_H(\mathbf{x}_i, \mathbf{x}_j) = d\} = \frac{1}{2} \sum_{i=1}^M \sum_{j \neq i} \Phi\{d_H(\mathbf{x}_i, \mathbf{x}_j) = d\} = \frac{M}{2} N_c(d) \doteq 2^{nR} N_c(d), \quad (25)$$

where $N_c(d) := \sum_{j \neq i} \Phi\{d_H(\mathbf{x}_i, \mathbf{x}_j) = d\}$. In [16], it is shown that with probability going to one as $n \rightarrow \infty$

$$N_c(d) \doteq \begin{cases} 2^{n(R+H(d/n)-1)}, & n\delta_{GV}(R) < d < n(1 - \delta_{GV}(R)) \\ 0, & \text{otherwise.} \end{cases} \quad (26)$$

Therefore, the average probability of a pair being non-close can be written as

$$\frac{\sum_{d > n(1/2+\epsilon)}^{n(1-\delta_{GV}(R))} 2^{n(2R+H(d/n)-1)}}{2^{2nR}} < \frac{n 2^{n(2R-1+H(\frac{1}{2}+\epsilon))}}{2^{2nR}}, \quad (27)$$

which again goes exponentially to zero as $n \rightarrow \infty$.

APPENDIX II

COMPUTATION OF $M_b(l, |\mathcal{E}|, |\mathcal{E}| - 1)$

We will show that for $l \geq |\mathcal{E}|$

$$M_b(l, |\mathcal{E}|, |\mathcal{E}| - 1) = M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 1)(2^{|\mathcal{E}|} - 1). \quad (28)$$

To this end, by symmetry,

$$M_b(l, |\mathcal{E}|, |\mathcal{E}| - 1) = M_b(|\mathcal{E}|, l, |\mathcal{E}| - 1).$$

And from Appendix A of [10] and $|\mathcal{E}| \leq l$, the RHS equals to

$$\begin{aligned} M_b(|\mathcal{E}|, l, |\mathcal{E}| - 1) &= M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 1) 2^{|\mathcal{E}|-1} \\ &+ M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 2)(2^l - 2^{|\mathcal{E}|-2}). \end{aligned}$$

From Appendix A of [10], we also have the following recursive formula for $j = 1 \dots |\mathcal{E}| - 2$

$$\begin{aligned} M_b(|\mathcal{E}| - j, l, |\mathcal{E}| - 1 - j) &= M_b(|\mathcal{E}| - 1 - j, l, |\mathcal{E}| - 1 - j) 2^{|\mathcal{E}|-1-j} \\ &+ M_b(|\mathcal{E}| - 1 - j, l, |\mathcal{E}| - 2 - j)(2^l - 2^{|\mathcal{E}|-2-j}). \end{aligned}$$

And $M_b(|\mathcal{E}|, l, |\mathcal{E}| - 1)$ equals to

$$\sum_{j=1}^{|\mathcal{E}|-1} \left\{ M_b(|\mathcal{E}| - j, l, |\mathcal{E}| - j) 2^{|\mathcal{E}|-j} \prod_{p=1}^{j-1} (2^l - 2^{|\mathcal{E}|-1-p}) \right\} \quad (29)$$

$$+ M_b(1, l, 0) * (2^l - 1) \prod_{p=1}^{|\mathcal{E}|-2} (2^l - 2^{|\mathcal{E}|-1-p}),$$

where $M_b(1, l, 0) = 1$.

Finally, using (15) in (29),

$$M_b(|\mathcal{E}|, l, |\mathcal{E}| - 1) = \sum_{j=1}^{|\mathcal{E}|} M_b(|\mathcal{E}| - 1, l, |\mathcal{E}| - 1) 2^{|\mathcal{E}|-j},$$

And it is easy to check that the above formula equals to (28).

REFERENCES

- [1] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1897–1905, Sep. 1998.
- [2] W. Trappe, M. Wu, Z. J. Wang, and K. J. R. Liu, "Anti-collusion fingerprinting for multimedia," *IEEE Trans. Signal. Proc.*, vol. 51, no. 4, pp. 1069–1087, Apr. 2003.
- [3] A. Barg, G. R. Blakley, and G. Kabatiansky, "Digital fingerprinting codes: Problem statements, constructions, identification of traitors," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 825–865, Apr. 2003.
- [4] A. D. Friedman, R. L. Graham, and J. D. Ullman, "Universal single transition time asynchronous state assignments," *IEEE Trans. Comput.*, vol. 18, no. 6, pp. 541–547, Jun. 1969.
- [5] N. Anthapadmanabhan, A. Barg, and I. Dumer. (2007) On the Fingerprinting Capacity Under the Marking Assumption. [Online]. Available: <http://arxiv.org/pdf/cs.IT/0612073/>
- [6] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 599–618, 2001.
- [7] H. Pishro-Nik and F. Fekri, "On decoding of low-density parity-check codes over the binary erasure channel," *IEEE Trans. Inform. Theory*, vol. 50, no. 3, pp. 439–454, 2004.
- [8] D. Divsalar, S. Dolinar, and C. Jones, "Low-rate LDPC codes with simple protograph structure," in *Proceedings of International Symposium on Information Theory, (ISIT)*, 2005, pp. 1622–1626.
- [9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [10] C. Di, D. Proietti, I. Telatar, T. J. Richardson, and R. L. Urbanke, "Finite-length analysis of low-density parity-check codes on the binary erasure channel," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1570–1579, 2002.
- [11] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.
- [12] D. Divsalar, H. Jin, and R. McEliece, "Coding theorems for turbo-like codes," in *Proc. 1998 Allerton Conf*, 1998, pp. 201–210.
- [13] H. D. Pfister, I. Sason, and R. Urbanke, "Capacity-achieving ensembles for the binary erasure channel with bounded complexity," *IEEE Trans. Inform. Theory*, vol. 51, no. 7, pp. 2352–2379, 2005.

- [14] H. D. Pfister and I. Sason, "Accumulate-repeat-accumulate codes: Capacity-achieving ensembles of systematic codes for the erasure channel with bounded complexity," *IEEE Trans. Inform. Theory*, vol. 53, no. 6, pp. 2088–2115, 2007.
- [15] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Efficient erasure correcting codes," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 569–584, 2001.
- [16] A. Barg and G. D. Forney, "Random codes: minimum distances and error exponents," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2568–2573, Sep. 2003.