Two-level Fingerprinting: Stronger Definitions and Code Constructions

N. Prasanth Anthapadmanabhan
Wireless Networking and Communications Group
University of Texas at Austin
Austin, TX 78712
Email: anprasanth@mail.utexas.edu

Alexander Barg
Dept. of ECE and Inst. for Systems Research
University of Maryland, College Park, MD 20742
and IPPI RAS, Moscow, Russia
Email: abarg@umd.edu

***Abstract*—We develop the concept of hierarchical fingerprinting introduced in our recent work (ISIT2009). The object of two-level fingerprinting is content protection against coalitions of t pirates in such a manner that one of the pirates can be identified exactly if $t \leq t_2$ or localized to within a small group if $t_2 < t \leq t_1$, where t_1 and t_2 are parameters of the system. In this work, we require the additional property that in the latter case no innocent users are accused of belonging to the pirate coalition. We construct two-level fingerprinting codes with polynomial complexity of identification that satisfy the strong definition of hierarchical fingerprinting described above.**

I. Introduction

In [1] we introduced the concept of hierarchical fingerprinting under which the content distributor is capable of identifying pirates exactly if the pirate coalition is of a small size $t \leq t_2$ or localizing them within a small group of users for coalitions of larger size, namely $t_2 < t \leq t_1$, where t_1 and t_2 are parameters of the system. However, the construction in [1] gave no guarantee with respect to accusing individual users for large coalitions (i.e., together with the correct group, the decoder could in principle accuse an innocent user). This shortcoming is addressed in the present work where we consider a stronger definition of hierarchical fingerprinting which provides an additional guarantee for the case of large coalitions: if $t_2 < t \leq t_1$, then the identifying algorithm locates the correct group of users and either outputs no individual users, or otherwise they are (with high probability) members of the pirate coalition. The main purpose of this paper is to introduce the strong version of two-level fingerprinting that accounts for this requirement. To justify the new definition, we develop ideas from our earlier work to characterize achievable rates for two-level fingerprinting codes. In addition, we also construct two-level fingerprinting codes with polynomial complexity of identification.

II. Problem Statement

Consider a distributor in possession of copyrighted content. Suppose a copy of the content needs to be distributed to M_1 groups of licensed users with each group containing M_2 users. A user $\mathbf{u}$ is identified by a pair of indices $\mathbf{u} \equiv (u_1, u_2) \in [M_1] \times [M_2]$, where the notation $[N]$ stands for $\{1, \ldots, N\}$.

The distributor's objective is to ensure that the content is protected against unauthorized distribution. In order to differentiate the distributed copies, in each copy the distributor embeds a string of symbols called a fingerprint. The fingerprint symbols are hidden over the entire content in locations unknown to the users. The fingerprint locations, however, are the same for all users. We are assuming that the fingerprints are represented by strings of length n over a finite alphabet $\mathcal{Q} = \{0, \ldots, q-1\}$. The distributor assigns fingerprints according to an encoding mapping

$$C : [M_1] \times [M_2] \to \mathcal{Q}^n. \tag{1}$$

A collusion attack occurs when a subset, or *coalition* U of users attempt to create an unregistered fingerprint $\mathbf{y}$ from their assigned fingerprints $C(U) = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ as well as from public information about the system. We assume that the set of all possible forgeries $\mathbf{y}$ generated by the coalition in this manner is given by

$$\mathcal{E}(\mathbf{x}_1, \ldots, \mathbf{x}_t) = \left\{ \mathbf{y} \in \mathcal{Q}^n \middle| y_i \in \{x_{1i}, \ldots, x_{ti}\}, \forall i \in [n] \right\}. \tag{2}$$

Given a forged fingerprint $\mathbf{y}$ from an illegal copy, the distributor attempts to trace one of the pirates. In general, the distributor's decoding objective is as follows: If the coalition size is less than a certain threshold t_2, then one of the guilty users must be recovered. If the coalition size is larger than t_2 but at most t_1 (where $t_1 > t_2$), then the group index of a guilty user must be recovered, while also ensuring that an innocent user is not accused. Note however that the distributor has no information about the actual size of the coalition that is in effect. The decoding (tracing) algorithm is defined by the following mapping:

$$D : \mathcal{Q}^n \to ([M_1] \cup \{\perp\}) \times ([M_2] \cup \{\perp\}). \tag{3}$$

A $\perp$ output from the decoding algorithm will signify a decoding failure for the corresponding index. For convenience, we write $D_1(\mathbf{y})$ and $D_2(\mathbf{y})$ to represent the first and second components respectively of $D(\mathbf{y})$. For a user $\mathbf{u} \equiv (u_1, u_2)$, we also write $\mathcal{G}(\mathbf{u}) := u_1$ to denote the group index of $\mathbf{u}$.

The pair of mappings (C, D), defined by (1) and (3), is called an $(n, M_1, M_2)_q$ *two-level code*. The *rate pair* of an $(n, M_1, M_2)_q$ two-level code is defined as

$$(R_1, R_2) := \left((1/n) \log_q M_1, (1/n) \log_q M_2 \right).$$

We assume that the code (C, D) is available publicly and that this knowledge can be exploited by the coalition in designing forgeries.

A. Traceability codes

Let $d_H(\mathbf{x},\mathbf{y})$ denote the Hamming distance between the vectors $\mathbf{x},\mathbf{y} \in \mathcal{Q}^n$ and let $s_H(\mathbf{x},\mathbf{y}) := n - d_H(\mathbf{x},\mathbf{y})$. We denote the Hamming weight of $\mathbf{x}$ by $w_H(\mathbf{x})$. Recall that for one-level codes, the traceability property [4] assumes decoding of $\mathbf{y}$ to the user whose fingerprint is the closest to $\mathbf{y}$ by the Hamming distance. In this section we extend the notion of traceability to two-level codes. For any code C of length n, any coalition U of size at most t and any $\mathbf{y} \in \mathcal{E}(C(U))$, clearly there exists some user $\mathbf{u} \in U$ such that $s_H(C(\mathbf{u}),\mathbf{y}) \geq n/t$. Motivated by this fact, we extend the tracing algorithm as follows: Given a forgery $\mathbf{y}$, let

$$d^\star = \min_{\mathbf{u}\in[M_1]\times[M_2]} d_H(C(\mathbf{u}),\mathbf{y})$$

and $\mathbf{u}^\star$ be a user for which the minimum is achieved. Let

$$D(\mathbf{y}) = \begin{cases} \mathbf{u}^\star & \text{if } d^\star \leq (1-1/t_2)n \\ (\mathcal{G}(\mathbf{u}^\star),\perp) & \text{otherwise.} \end{cases}$$

Definition 2.1: A two-level code C has (t_1,t_2)-*traceability* property (or is (t_1,t_2)-TA), where $t_1 > t_2$, if

(a) For any coalition U of size at most t_2 and any $\mathbf{y} \in \mathcal{E}(C(U))$, the decoding result $D(\mathbf{y}) \in U$.
(b) For any coalition U of size at most t_1 and any $\mathbf{y} \in \mathcal{E}(C(U))$, the decoding result $D_1(\mathbf{y}) \in \mathcal{G}(U)$. In addition, if $D_2(\mathbf{y}) \neq \perp$, then $D(\mathbf{y}) \in U$.

We next derive a simple sufficient condition for a two-level code C to be (t_1,t_2)-TA. For a given two-level code C, as in [1], we define two minimum distances:

$$d_1(C) := \min_{\substack{\mathbf{u},\mathbf{v}\in[M_1]\times[M_2]\\ u_1\neq v_1}} d_H(C(\mathbf{u}),C(\mathbf{v})), \tag{4}$$

$$d_2(C) := \min_{\substack{\mathbf{u},\mathbf{v}\in[M_1]\times[M_2]\\ u_2\neq v_2}} d_H(C(\mathbf{u}),C(\mathbf{v})). \tag{5}$$

Proposition 2.2: Suppose $t_1 > t_2$ and C is a two-level code of length n with

$$d_1(C) > n\left(1-\frac{1}{t_1^2}\right) \quad \text{and} \quad d_2(C) > n\left(1-\frac{1}{t_1 t_2}\right). \tag{6}$$

Then C is (t_1,t_2)-TA.

Proof: We check only the second claim in Part (b) of Def. 2.1 because otherwise the proof is the same as in [1]. Let U be a coalition of size at most t_1 and $\mathbf{y} \in \mathcal{E}(C(U))$. Observe that a decoding error occurs with $D_2(\mathbf{y}) \neq \perp$ only if there exists some innocent user $\mathbf{u}' \notin U$ such that $d_H(C(\mathbf{u}'),\mathbf{y}) \leq n(1-1/t_2)$. However, for any user $\mathbf{u}' \notin U$, we find that $s_H(C(\mathbf{u}'),\mathbf{y}) \leq t_1(n-d_2(C)) < n/t_2$, thus ruling out the possibility of this error event. ∎

B. Randomized fingerprinting codes

Randomization of the encoding/decoding maps can enable the system to support a larger number of users in exchange for a small error probability of identification.

Suppose that the distributor chooses the encoding and decoding mappings $(\mathcal{C},\mathcal{D})$ randomly from a family $\{(C_k,D_k), k \in \mathcal{K}\}$ of $(n,M_1,M_2)_q$ codes according to some probability distribution on the set $\mathcal{K}$ of keys. We assume that the users have complete knowledge of the family of codes and the probability distribution over $\mathcal{K}$, but the exact choice of key is a secret known only to the distributor.

Let U be a coalition of size t. In order to create a forged fingerprint, the coalition may use an arbitrary randomized strategy $V(\cdot|\cdot,\ldots,\cdot)$, where $V(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_t)$ gives the probability that the forgery $\mathbf{y}$ is generated on observing the fingerprints $\mathbf{x}_1,\ldots,\mathbf{x}_t$. Due to the assumed restrictions (2) on creating forgeries, we say a strategy V is *admissible* if

$$V(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_t) = 0 \text{ for all } \mathbf{y} \notin \mathcal{E}(\mathbf{x}_1,\ldots,\mathbf{x}_t).$$

The class of admissible strategies is denoted by $\mathcal{V}_t$. Let $\mathbf{Y}$ be a random vector denoting the forgery created by the coalition in this manner. We now define the probabilities for the error events during decoding:

$$\begin{aligned} e_1(\mathcal{C},\mathcal{D},U,V) &= \mathbf{P}\left[\mathcal{D}_2(\mathbf{Y}) = \perp, \mathcal{D}_1(\mathbf{Y}) \notin \mathcal{G}(U)\right] \\ &= \mathbf{E}_K \sum_{\mathbf{y}:D_{2K}(\mathbf{y})=\perp, D_{1K}(\mathbf{y})\notin\mathcal{G}(U)} V(\mathbf{y}|C_K(U)), \\ e_2(\mathcal{C},\mathcal{D},U,V) &= \mathbf{P}\left[\mathcal{D}_2(\mathbf{Y}) \neq \perp, \mathcal{D}(\mathbf{Y}) \notin U\right] \\ &= \mathbf{E}_K \sum_{\mathbf{y}:D_{2K}(\mathbf{y})\neq\perp, D_K(\mathbf{y})\notin U} V(\mathbf{y}|C_K(U)), \\ e_3(\mathcal{C},\mathcal{D},U,V) &= \mathbf{P}\left[\mathcal{D}(\mathbf{Y}) \notin U\right] \\ &= \mathbf{E}_K \sum_{\mathbf{y}:D_K(\mathbf{y})\notin U} V(\mathbf{y}|C_K(U)). \end{aligned}$$

Definition 2.3: A randomized code $(\mathcal{C},\mathcal{D})$ is said to be a (t_1,t_2)-*fingerprinting code with* ϵ-*error* where $t_1 > t_2$ if:

(a) For any coalition U of size at most t_2 and any admissible strategy V, the error probability $e_3(\mathcal{C},\mathcal{D},U,V) \leq \epsilon$.
(b) For any coalition U of size at most t_1 and any admissible strategy V, the error probability $e_1(\mathcal{C},\mathcal{D},U,V) + e_2(\mathcal{C},\mathcal{D},U,V) \leq \epsilon$.

In this paper, we investigate the set of rate pairs (R_1,R_2) that are achievable for the above two-level fingerprinting codes with the error probability decaying to 0 (as the code length increases). We also construct two-level fingerprinting codes that have a tracing algorithm operating with complexity polynomial in the fingerprint length n.

III. Achievable Rates

We use the idea of [1] for code construction but use a new decoding procedure in order to address the stronger identification guarantees of Definition 2.3. We note that the construction below is inspired by error-correcting codes with unequal error protection [2] used in communications problems.

We construct an $(n,M_1,M_2)_q$ randomized code $(\mathcal{C},\mathcal{D})$ as follows. For $w \in [n]$ and $R_1,R_2 \in [0,1]$ define $\mathcal{S}_{w,n} = \{\mathbf{x} \in \mathcal{Q}^n : w_H(\mathbf{x}) = w\}$, $M_1 = \lfloor q^{nR_1} \rfloor$, and $M_2 = \lfloor q^{nR_2} \rfloor$. Fix $\omega \in [0,1]$ and take n such that $w = \omega n$ is an integer. For $i \in [M_1]$, we first pick the "centers" $\mathbf{R}_i$ independently and uniformly at random from $\mathcal{Q}^n$. Next, we choose $\mathbf{S}_{ij}$, $(i,j) \in [M_1] \times [M_2]$ independently and uniformly at random from

$\mathcal{S}_{w,n}$. Finally, using modulo q operations in $\mathcal{Q}$, the fingerprint $\mathbf{X_u}$ for each user $\mathbf{u} \equiv (u_1, u_2) \in [M_1] \times [M_2]$ is generated as

$$\mathbf{X_u} = \mathbf{R}_{u_1} + \mathbf{S_u}$$

Let us recall from [1] the following simple facts which will be useful in our error probability analysis.

Lemma 3.1: If $\mathbf{S}$ is uniformly distributed over $\mathcal{S}_{w,n}$, then $\mathbf{P}[S_l = a] = \omega/(q-1)$ for $l \in [n]$, $a \in \mathcal{Q}\backslash\{0\}$. Moreover, the r.v.'s $\{S_l, l \in [n]\}$ are pairwise independent asymptotically.

Lemma 3.2: Suppose $p \in [0,1]$ and $\epsilon > 0$. For $l \in [n]$, let Z_l be a Bernoulli r.v. with success probability p such that $\{Z_l, l \in [n]\}$ are pairwise independent. Then, with $Z := \sum_{l\in[n]} Z_l$, we have

$$\mathbf{P}[Z \notin [n(p-\epsilon), n(p+\epsilon)]] \leq \frac{p(1-p)}{\epsilon^2 n}.$$

Below, for two functions $f(n), g(n)$, we write $f(n) \doteq g(n)$ if $\lim_{n\to\infty} n^{-1}\log(f(n)/g(n)) = 0$. We denote the q-ary entropy function by $h(x)$. Let $q \geq 3$. For $\omega, \gamma, \alpha, \beta \in [0,1]$, with $\alpha \leq 1-\gamma$, $\beta \leq \gamma$, $\alpha+\beta \leq \omega$, $\omega - \alpha \leq \gamma$, let us define

$$\begin{aligned}&\phi(\omega,\gamma,\alpha,\beta)\\ &= (1-\gamma)h\left(\frac{\alpha}{1-\gamma}\right) + (\gamma-\beta)h\left(\frac{\omega-\alpha-\beta}{\gamma-\beta}\right)\\ &\quad + \gamma h\left(\frac{\beta}{\gamma}\right) + (\omega-\alpha)\log_q\left(\frac{q-2}{q-1}\right) - \beta\log_q(q-2).\end{aligned}$$

Lemma 3.3: Let $\mathbf{S}$ have a uniform distribution over $\mathcal{S}_{w,n}$ and $\delta \in [0,1]$. Suppose $\mathbf{y} \in \mathcal{Q}^n$ such that $w_H(\mathbf{y}) = \gamma n$, where $\gamma \in [0,1]$. Then

$$\mathbf{P}[d_H(\mathbf{S},\mathbf{y}) \leq \delta n] \doteq q^{-nE(\omega,\gamma,\delta)},$$

where

$$E(\omega,\gamma,\delta) = h(\omega) - \max_{\substack{\alpha,\beta:\\ \gamma-\beta+\alpha\leq\delta}} \phi(\omega,\gamma,\alpha,\beta).$$

The following notation is used below. For a coalition $U = \{\mathbf{u}^1, \ldots, \mathbf{u}^t\}$, we denote the realizations of $\mathbf{X}_{\mathbf{u}^i}, \mathbf{R}_{u_1^i}, \mathbf{S}_{\mathbf{u}^i}$ by $\mathbf{x}_i, \mathbf{r}_i, \mathbf{s}_i$ respectively, with $\mathbf{x}_i = \mathbf{r}_i + \mathbf{s}_i, i \in [t]$. We use $\mathbf{R}$ and $\mathbf{S}$ as representative r.v.'s which have the same distribution as $\mathbf{R}_i$ and $\mathbf{S}_{ij}$ respectively in our construction.

A. $(t,1)$-fingerprinting

Clearly, a single user (size-1 coalition) cannot generate any forgery except its own fingerprint. Taking advantage of this fact, we *define the decoder as follows:* if there is a user $\mathbf{u}$ such that $C(\mathbf{u}) = \mathbf{y}$, then $D(\mathbf{y}) = \mathbf{u}$, otherwise $D(\mathbf{y}) = \mathcal{G}(\mathbf{u}^\star, \perp)$, where $\mathbf{u}^\star$ is the user whose fingerprint is the closest to $\mathbf{y}$.

Theorem 3.4: For any ω such that $\frac{t-1}{t}\left(1-\frac{1}{q^{t-1}}\right)+\omega \leq \frac{q-1}{q}$, the randomized code $(\mathcal{C},\mathcal{D})$ is $(t,1)$-fingerprinting with error probability decaying to 0 if

$$R_1 < 1 - h\Big(\frac{t-1}{t}\Big(1-\frac{1}{q^{t-1}}\Big)+\omega\Big), \tag{7}$$

$$R_2 < h(\omega). \tag{8}$$

The proof, which will be omitted, consists of three cases that account for the three error events defined above. The analysis of two of them is rather similar to the proof of Theorems 4.3 and 4.4 from [1]. At the same time, we need to analyze the probability of an additional error event which accounts for the stronger notion of two-level fingerprinting used in this paper.

B. $(t,2)$-fingerprinting

Let $\epsilon > 0$ be arbitrarily small and define $\delta = (q-1)/2q$. Given a forgery $\mathbf{y}$, let $\mathbf{u}^\star$ and $d^\star$ be as above. *The decoder is defined as follows:*

$$\mathcal{D}(\mathbf{y}) = \begin{cases}\mathbf{u}^\star & \text{if } d^\star \leq n(\delta+\epsilon)\\ (\mathcal{G}(\mathbf{u}^\star), \perp) & \text{otherwise.}\end{cases}$$

For $m = 1, \ldots, t$, let

$$\tilde{f}_m(\omega) = \max_{\substack{\gamma,\alpha,\beta:\\ \omega^m(\frac{q-1}{q})^{t-m}\leq\gamma\leq 1-\frac{(1-\omega)^m}{q^{t-m}}, \gamma-\beta+\alpha\leq\delta}} \phi(\omega,\gamma,\alpha,\beta).$$

Theorem 3.5: Let $q \geq 3$. For any ω such that $\frac{t-1}{t}\big(1-\frac{1}{q^{t-1}}\big)+\omega \leq \frac{q-1}{q}$, the randomized code $(\mathcal{C},\mathcal{D})$ is $(t,2)$-fingerprinting with error probability decaying to 0 if

$$R_1 < 1 - h\Big(\frac{t-1}{t}\Big(1-\frac{1}{q^{t-1}}\Big)+\omega\Big), \tag{9}$$

$$R_2 < h(\omega) - \max(\tilde{f}_m(\omega), m = 1, \ldots, t). \tag{10}$$

Proof: For any coalition of size 2, we can show that with probability approaching 1 a guilty user is within distance $n(\delta+\epsilon)$ from the forgery. Once we make this observation, the analysis for size-2 coalitions proceeds in exactly the same manner as in Theorem 4.6 in [1]. In the case of size-t coalitions, we should take into account an additional error event which occurs if $\mathcal{D}_2(\mathbf{Y}) \neq \perp$ and $\mathcal{D}(\mathbf{Y}) \notin U$.

Consider a coalition U of size t. Suppose that the coalition's members belong to L different groups with each group containing $m_i, i = 1, \ldots, L$ members of U. Thus, we have $\sum_{i=1}^L m_i = t$ and

$$\begin{aligned}&e_2(\mathcal{C},\mathcal{D},U,V)\\ &= \sum_{\substack{\mathbf{r}_1,\ldots,\mathbf{r}_L\\ \mathbf{s}_1,\ldots,\mathbf{s}_t}} \mathbf{P}[\mathbf{r}_1,\ldots,\mathbf{r}_L,\mathbf{s}_1,\ldots,\mathbf{s}_t] \sum_{\mathbf{y}} V(\mathbf{y}|\mathbf{x}_1,\ldots,\mathbf{x}_t)\\ &\qquad \times \mathbf{P}\Big[\mathcal{D}_2(\mathbf{y}) \neq \perp, \mathcal{D}(\mathbf{y}) \notin U \Big| \mathbf{r}_1,\ldots,\mathbf{r}_L,\mathbf{s}_1,\ldots,\mathbf{s}_t\Big].\end{aligned}$$

We now focus on the inner conditional probability. Observe that $\mathcal{D}_2(\mathbf{y}) \neq \perp$ only if there exists a user $\mathbf{u}' \notin U$ such that $d_H(\mathbf{X}_{\mathbf{u}'}, \mathbf{y}) \leq n(\delta+\epsilon)$. Since $\mathbf{u}'$ can either belong to one of the L groups in $\mathcal{G}(U)$ or be in a different group, we have

$$\begin{aligned}&\mathbf{P}\Big[\mathcal{D}_2(\mathbf{y}) \neq \perp, \mathcal{D}(\mathbf{y}) \notin U \Big| \mathbf{r}_1,\ldots,\mathbf{r}_L,\mathbf{s}_1,\ldots,\mathbf{s}_t\Big]\\ &\leq \sum_{i=1}^{L} q^{nR_2}\mathbf{P}[d_H(\mathbf{S},\mathbf{y}-\mathbf{r}_i) \leq n(\delta+\epsilon)]\\ &\qquad + q^{nR_1}\mathbf{P}[d_H(\mathbf{R},\mathbf{y}) \leq n(\delta+\epsilon)+w].\end{aligned}$$

Observe that the last term behaves as $q^{-n(1-h(\delta+\omega)-R_1)}$ for ϵ arbitrarily small and thus approaches 0 for R_1 satisfying (9).

W.l.o.g. let us consider the term $\mathbf{P}[d_H(\mathbf{S},\mathbf{y}') \leq n(\delta+\epsilon)]$, where $\mathbf{y}' := \mathbf{y} - \mathbf{r}_1$. Define $\tilde{U} = \{\mathbf{s}_1, \ldots, \mathbf{s}_{m_1}, \mathbf{x}_{m_1+1} -$

$\mathbf{r}_1, \ldots, \mathbf{x}_t - \mathbf{r}_1\}$. We note that $\mathbf{y}' \in \mathcal{E}(\tilde{U})$. Let $s_0(\tilde{U})$ denote the proportion of the columns which are all-zero in the matrix formed using the vectors in $\tilde{U}$ as the rows. Similarly, for the same matrix, $s_1(\tilde{U})$ denotes the proportion of columns where every element is non-zero. Define $\mathcal{T}_n$ to be the set of all $(\mathbf{r}_1, \ldots, \mathbf{r}_L, \mathbf{s}_1, \ldots, \mathbf{s}_t)$ such that

$$s_0(\tilde{U}) \simeq (1-\omega)^{m_1} \prod_{i=2}^{L} \Big(\frac{(1-\omega)^{m_i}}{q} + \frac{q-1}{q} \Big(\frac{\omega}{q-1} \Big)^{m_i} \Big),$$

$$s_1(\tilde{U}) \simeq \omega^{m_1} \prod_{i=2}^{L} \Big(\frac{\omega^{m_i}}{q} + \frac{q-1}{q} \Big(1 - \frac{\omega}{q-1} \Big)^{m_i} \Big).$$

For simplicity, we use the approximate relations $\simeq$ and $\lesssim$ when the respective exact relations hold within some arbitrarily small ϵ. Using Lemma 3.1 and Lemma 3.2, it can be shown that $(\mathbf{R}_1, \ldots, \mathbf{R}_L, \mathbf{S}_1, \ldots, \mathbf{S}_t)$ is in $\mathcal{T}_n$ with probability approaching 1. Now, assuming that the vectors indeed belong to $\mathcal{T}_n$, we can show using Jensen's inequality that

$$\omega^{m_1} \Big(\frac{q-1}{q} \Big)^{t-m_1} \lesssim \frac{w_H(\mathbf{y}')}{n} \lesssim 1 - \frac{(1-\omega)^{m_1}}{q^{t-m_1}}.$$

Finally, using Lemma 3.3 and taking $\epsilon \to 0$, we conclude that the error probability approaches 0 if R_2 satisfies (10). ∎

IV. Constructions with Polynomial-time Tracing

Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ denote finite alphabets of size Q_1 and Q_2 respectively. We introduce the operation $*$ which is defined in the following way: For $\mathbf{x} \in \mathcal{Q}_1^N$, $\mathbf{y} \in \mathcal{Q}_2^N$, $\mathbf{x} * \mathbf{y} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{Q}_1 \times \mathcal{Q}_2)^N$.

Let C_1 be an $[N, K_1, \Delta_1]_{Q_1}$ linear code and C_2 be an $[N, K_2, \Delta_2]_{Q_2}$ linear code. The $*$ product extends to the codes C_1 and C_2 as follows: $C_1 * C_2 = \{\mathbf{x}_1 * \mathbf{x}_2 : \mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2\}$. To obtain a two-level code, we associate C_1 with groups and C_2 with users within the group. Thus $C_1 * C_2$ can be viewed as an $(N, Q_1^{K_1}, Q_2^{K_2})_{Q_1 Q_2}$ two-level code over the alphabet $\mathcal{Q}_1 \times \mathcal{Q}_2$. Obviously, it is true that $d_1(C_1 * C_2) \geq \Delta_1$ and $d_2(C_1 * C_2) \geq \Delta_2$, where the quantities d_1 and d_2 are defined in (4), (5). Therefore, choosing $\Delta_1 > N(1 - 1/t_1^2)$ and $\Delta_2 > N(1 - 1/t_1 t_2)$ makes the resulting code $C_1 * C_2$ into a (t_1, t_2)-TA code by Proposition 2.2. This observation forms the motivation for the code choices in our concatenated scheme described below.

Let C_1 and C_2 both be RS (or one-point AG) codes with parameters $[N, K_1, \Delta_1]_{Q_1}$ and $[N, K_2, \Delta_2]_{Q_2}$ respectively, where $K_i = R_{i,\text{out}} N$ and $\Delta_i = \delta_i N$ for $i = 1, 2$. Each codeword from C_1 corresponds to a particular group, while the codewords of C_2 are associated with the user indices within a group. Then the outer code $C_{\text{out}} = C_1 * C_2$ is an $(N, Q_1^{K_1}, Q_2^{K_2})_{Q_1 Q_2}$ deterministic two-level code. Let $(\mathcal{C}_{\text{in}}, \mathcal{D}_{\text{in}})$ denote an $(m, Q_1, Q_2)_q$ randomized code which is (t_1, t_2)-fingerprinting with ϵ-error under exhaustive search decoding. For every outer coordinate $i \in [N]$, we generate an *independent* instance of $(\mathcal{C}_{\text{in}}, \mathcal{D}_{\text{in}})$ for the inner level.

For a given user $\mathbf{u} \equiv (u_1, u_2)$, the fingerprint is assigned as follows. At the outer level, pick $\mathbf{x}_1 \in C_1$ and $\mathbf{x}_2 \in C_2$ corresponding to u_1 and u_2 respectively, and construct $\mathbf{x} = \mathbf{x}_1 * \mathbf{x}_2$. Next, for each $i = 1, \ldots, N$, encode $(x_{1i}, x_{2i}) \in \mathcal{Q}_1 \times \mathcal{Q}_2$ using the realization of the two-level code $\mathcal{C}_{\text{in}}$. This procedure results in a concatenated code $\mathcal{C}$ which is a randomized $(n, Q_1^{K_1}, Q_2^{K_2})_q$ two-level code with $n = Nm$. In the subsequent text, the users are identified with the codewords of C_{out}. For $\mathbf{x} = \mathbf{x}_1 * \mathbf{x}_2 \in C_{\text{out}}$, with some abuse of notation we write $\mathcal{G}(\mathbf{x}) = \mathbf{x}_1$.

W.l.o.g. we suppose that the tracing strategy $\mathcal{D}_{\text{in}}$ always outputs exactly one user (if there is no decoding failure). In practice, if the decoder outputs multiple candidates, the decoding output is chosen randomly from the candidate list. In addition, we assume that the inner fingerprinting code is "symmetric" across the users, meaning that the fingerprints of different users are identically distributed random variables. We also assume that this applies to different groups as a whole.

Our tracing algorithm makes use of Guruswami-Sudan (GS) list decoding [5] for errors and erasures. Let C be an $[N, K, \delta N]_q$ RS code (or one-point AG code) over the alphabet Q. Then for any given $\mathbf{y} \in (\mathcal{Q} \cup \{\perp\})^N$, where $\perp$ denotes an erasure, the number of codewords $\mathbf{x} \in C$ such that $s_H(\mathbf{x}, \mathbf{y}) \geq N\sqrt{1-\delta}$ is polynomial in N. Moreover, there exists an algorithm with complexity polynomial in N which outputs the list of all such codewords.

Given a forged fingerprint $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, where each $\mathbf{y}_i$ is a q-ary length-m vector, our decoding algorithm operates as follows.

1) For every $i \in [N]$, apply the tracing strategy $\mathcal{D}_{\text{in}}$ to $\mathbf{y}_i$ to obtain $(\hat{y}_{1i}, \hat{y}_{2i}) \in (\mathcal{Q}_1 \cup \{\perp\}) \times (\mathcal{Q}_2 \cup \{\perp\})$. Completing this procedure for all N outer coordinates produces a vector $\hat{\mathbf{y}} = \hat{\mathbf{y}}_1 * \hat{\mathbf{y}}_2$.
2) Let $\xi > \epsilon$. Run the GS list decoding algorithm for $C_j, j = 1, 2$, with $\perp$ treated as an erasure, to compute the lists
$$L_j(\hat{\mathbf{y}}_j) = \{\mathbf{x}_j \in C_j : s_H(\mathbf{x}_j, \hat{\mathbf{y}}_j) \geq N(1-\xi)/t_j\} \quad (11)$$
3) Finally, the decoder outputs the list $L(\hat{\mathbf{y}})$ computed as follows. For every $\mathbf{x}_1 \in L_1(\hat{\mathbf{y}}_1)$, $\mathbf{x}_2 \in L_2(\hat{\mathbf{y}}_2)$, if $s_H(\mathbf{x}_1 * \mathbf{x}_2, \hat{\mathbf{y}}) \geq N(1-\xi)/t_2$, then $(\mathbf{x}_1, \mathbf{x}_2) \in L(\hat{\mathbf{y}})$. If there is no such $\mathbf{x}_2$ for a given $\mathbf{x}_1$, then $(\mathbf{x}_1, \perp) \in L(\hat{\mathbf{y}})$.

The concatenated code thus defined together with the decoding algorithm described is denoted by $(\mathcal{C}, \mathcal{D})$ below.

Theorem 4.1: Let $0 < \epsilon < \xi$ and $\sigma_i = \frac{1-\xi}{t_i} - t_1(1-\delta_i), i = 1, 2$. Suppose that the relative minimum distances of C_1 and C_2 satisfy

$$\delta_i \geq 1 - \frac{(1-\xi)^2}{t_1 t_i} + \frac{\epsilon}{t_i(Q_i - t_1)}, \quad i = 1, 2, \qquad (12)$$

and the inner code $(\mathcal{C}_{\text{in}}, \mathcal{D}_{\text{in}})$ is (t_1, t_2)-fingerprinting with ϵ-error. Then the concatenated code $(\mathcal{C}, \mathcal{D})$ is (t_1, t_2)-fingerprinting with decoding complexity $\text{poly}(N)$ and error probability at most

$$q^{-ND(\xi \| \epsilon)} + Q_1^{K_1} q^{-ND(\sigma_1 \| \frac{\epsilon}{Q_1 - t_1})} + t_1 Q_2^{K_2} q^{-ND(\sigma_2 \| \frac{\epsilon}{Q_2 - t_1})}. \qquad (13)$$

Proof outline: We write a coalition U of size t as a subset $\{\mathbf{x}^1, \ldots, \mathbf{x}^t\} \subseteq C_{\text{out}}$ where $\mathbf{x}^i = \mathbf{x}_1^i * \mathbf{x}_2^i$, $i = 1, \ldots, t$. Let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, where $\mathbf{Y}_i$ are q-ary vectors of

length m, be a random forgery generated by U using any admissible strategy. In any outer coordinate $i \in [N]$, the coalition observes at most t distinct $\mathcal{Q}_1 \times \mathcal{Q}_2$ symbols among $\{(x_{1i}^1, x_{2i}^1), \ldots, (x_{1i}^t, x_{2i}^t)\}$. At the inner level this is equivalent to the action of a virtual coalition of size at most t, and correspondingly $\mathbf{Y}_i$ is generated by an admissible strategy from these symbols. This enables us to utilize the fingerprinting property of the inner code to derive some properties of the vector $\hat{\mathbf{Y}} = \hat{\mathbf{Y}}_1 * \hat{\mathbf{Y}}_2$ which is the result of inner level decoding performed in Step 1 of the algorithm. We use the above argument to analyze the probability of missed detection and of identifying an innocent user for the separate cases of t_1- and t_2-sized coalitions.

Consider a coalition $U = \{\mathbf{x}^1, \ldots, \mathbf{x}^{t_1}\}$ of size t_1 and a random forgery $\mathbf{Y}$ generated by U. We analyze the 3 cases that correspond to the three error events described before Definition 2.3. As mentioned above, this amounts to a virtual coalition of at most t_1 symbols acting in every outer coordinate. Therefore, for every $i \in [N]$, the probability that the group index $\hat{Y}_{1i}$ output by the inner decoder $\mathcal{D}_{\text{in}}$ does not match one of the symbols $\{x_{1i}^1, \ldots, x_{1i}^{t_1}\}$ is at most ϵ. Let Z_1 be a binomial r.v. denoting the number of coordinates where this error event occurs. Then there exists an $l \in [t_1]$ such that $s_H(\mathbf{x}_1^l, \hat{\mathbf{Y}}_1) \geq (N - Z_1)/t_1$, and so the probability that none of the guilty groups are output

$$\begin{aligned}\mathbf{P}\left[L_1(\hat{\mathbf{Y}}_1) \cap \mathcal{G}(U) = \emptyset\right] &\leq \mathbf{P}\left[\frac{N - Z_1}{t_1} < \frac{N(1-\xi)}{t_1}\right] \\ &= \mathbf{P}\left[Z_1 > N\xi\right] \leq q^{-ND(\xi\|\epsilon)},\end{aligned}$$

where the last inequality holds since $\xi > \epsilon$. This concludes the analysis of the missed group detection case.

Next, consider an innocent group $\mathbf{x}_1' \notin \mathcal{G}(U)$. The decoder makes an error if $\mathbf{x}_1' \in L_1(\hat{\mathbf{Y}}_1)$. In any coordinate $i \in [N]$, there are two possible ways for the group index output by $\mathcal{D}_{\text{in}}$ to be x_{1i}'. The first possibility is that $x_{1i}' \in \{x_{1i}^1, \ldots, x_{1i}^{t_1}\}$, and the number of such positions is at most $t_1(1-\delta_1)N$. Otherwise, if x_{1i}' is different from the above symbols, it can be output when the inner decoder makes an error. Due to the assumed symmetry of $(\mathcal{C}_{\text{in}}, \mathcal{D}_{\text{in}})$ the probability of this event is at most $\epsilon/(Q_1 - t_1)$. Let $\tilde{Z}_1$ be a binomial r.v. counting the number of coordinates where the latter error event occurs. Then $s_H(\mathbf{x}_1', \hat{\mathbf{Y}}_1) \leq \tilde{Z}_1 + t_1(1-\delta_1)N$ and we obtain

$$\begin{aligned}\mathbf{P}\left[\mathbf{x}_1' \in L_1(\hat{\mathbf{Y}}_1)\right] &= \mathbf{P}\left[s_H(\mathbf{x}_1', \hat{\mathbf{Y}}_1) \geq \frac{N(1-\xi)}{t_1}\right] \\ &\leq \mathbf{P}\left[\tilde{Z}_1 \geq N\sigma_1\right] \leq q^{-ND(\sigma_1\|\frac{\epsilon}{Q_1 - t_1})},\end{aligned}$$

because by (12), $\sigma_1 > \epsilon/(Q_1 - t_1)$.

Finally, consider an innocent user $\mathbf{x}' \notin U$ such that $\mathbf{x}_1' \in \mathcal{G}(U)$. In any coordinate $i \in [N]$, there are two possible ways for $\mathcal{D}_{\text{in}}$ to output the symbol (x_{1i}', x_{2i}'). One possibility is that $(x_{1i}', x_{2i}') \in \{(x_{1i}^1, x_{2i}^1), \ldots, (x_{1i}^{t_1}, x_{2i}^{t_1})\}$, and there are at most $t_1(1-\delta_2)N$ such positions. Secondly, if (x_{1i}', x_{2i}') is different from the actual coalition's symbols, it may be output when $\mathcal{D}_{\text{in}}$ makes an error. The probability of this event is at most $\epsilon/(Q_2 - t_1)$ due to the assumed symmetry of the inner code. Let $\tilde{Z}_2$ be a binomial r.v. denoting the number of coordinates where the second error event occurs. Then $s_H(\mathbf{x}', \hat{\mathbf{Y}}) \leq \tilde{Z}_2 + t_1(1-\delta_2)N$, and we get

$$\begin{aligned}\mathbf{P}\left[\mathbf{x}' \in L(\hat{\mathbf{Y}})\right] &= \mathbf{P}\left[s_H(\mathbf{x}', \hat{\mathbf{Y}}) \geq \frac{N(1-\xi)}{t_2}\right] \\ &\leq \mathbf{P}\left[\tilde{Z}_2 \geq N\sigma_2\right] \leq q^{-ND(\sigma_2\|\frac{\epsilon}{Q_2 - t_1})},\end{aligned}$$

since $\sigma_2 > \epsilon/(Q_2 - t_1)$ from (12). Applying the union bound, we conclude that the error probability is less than the estimate (13). The analysis for size-t_2 coalitions is carried out in a similar manner.

It is not difficult to show that the complexity of the identification procedure is a polynomial function of the code's length N. ■

Let us analyze the rates attained by Theorem 4.1 by fixing our code choices. Let C_1 and C_2 be extended RS codes with parameters $[Q, K_1]$ and $[Q, K_2]$, respectively, satisfying the condition (12). At the inner level, consider a sequence of q-ary (t_1, t_2)-fingerprinting codes with error probability $\epsilon = o(1)$ and rate $(R_{\text{in}}, R_{\text{in}})$. The tracing procedure of the inner code will be performed by exhaustive search; for instance, the codes in Section III can be used at the inner level. We have $m \approx O(\log_q n)$ since $n = mq^{mR_{\text{in}}}$. Hence, the tracing for the inner code has only polynomial complexity in the code length n. With ξ, t_1, t_2 fixed and m growing, we have

$$D\Big(\sigma_i \Big\| \frac{\epsilon}{Q - t_1}\Big) \sim N\sigma_i \log_q \frac{Q}{\epsilon} \geq n\sigma_i R_{\text{in}}, \quad i = 1, 2.$$

Let $R_i = R_{i,\text{out}} R_{\text{in}}$, $i = 1, 2$ denote the rate pair of the concatenated code. Since for RS codes we have $1 - \delta_i \sim R_{i,\text{out}}$, the error probability (13) approaches 0 if for $i = 1, 2$

$$R_i < \left(\frac{1-\xi}{t_i} - t_1 R_{i,\text{out}}\right) R_{\text{in}}, \quad \text{i.e.,} \quad R_i < \frac{1-\xi}{t_i(t_1+1)} R_{\text{in}}.$$

Finally, taking ξ arbitrarily small and m sufficiently large to satisfy $\epsilon < \xi$ we obtain the following result.

Corollary 4.2: There exists a sequence of q-ary (t_1, t_2)-fingerprinting codes of length n with error probability decaying with n, having decoding complexity $\text{poly}(n)$ and rate pair $R_1 = \Omega(R_{\text{in}}/t_1^2)$, $R_2 = \Omega(R_{\text{in}}/t_1 t_2)$.

ACKNOWLEDGMENT

A. Barg was supported in part by NSF grants CCF0635271, CCF0830699, CCF0916919, and DMS0807411.

REFERENCES

[1] N. P. Anthapadmanabhan and A. Barg, "Two-level fingerprinting codes," *Proc. IEEE Internat. Sympos. Information Theory* (ISIT 2009), Seoul, Korea, Jun. 28 - Jul. 3, 2009, pp. 2261–2265.

[2] L. A. Bassalygo, V. A. Zinov'ev, V. V. Zyablov, M. S. Pinsker and G. Sh. Poltyrev, "Bounds for codes with unequal error protection of two sets of messages," *Probl. of Inform. Trans.*, Vol. 15, No. 3, pp. 190–197, Jul.–Sep. 1979.

[3] D. Boneh and J. Shaw, "Collusion-secure fingerprinting for digital data," *IEEE Trans. Inform. Theory*, Vol. 44, No. 5, pp. 1897–1905, Sep. 1998.

[4] B. Chor, A. Fiat, M. Naor and B. Pinkas, "Tracing traitors," *IEEE Trans. Inform. Theory*, Vol. 46, No. 3, pp. 893-910, May 2000.

[5] V. Guruswami, *List decoding of error-correcting codes*, Lecture Notes in Computer Science, Vol. 3282, Springer, 2005.