# New Traceability Codes and Identification Algorithm for Tracing Pirates

Xin-Wen Wu, Paul Watters, and John Yearwood
Centre for Informatics and Applied Optimisation
University of Ballarat
Mt Helen, Ballarat, VIC 3353, Australia
{x.wu,p.watters,j.yearwood}@ballarat.edu.au

## Abstract

*With the increasing popularity of digital products, there is a strong desire to protect the rights of owners against illegal redistribution. Traditional encryption schemes alone do not provide a comprehensive solution to digital rights management, since they do not prevent users who are authorized to use a digital product for their own use from transferring the cleartext content to unauthorized users. However, traceability schemes can be used to trace the illegitimate redistributors effectively. Two types of traceability schemes have been proposed in the literature - traceability codes (TA codes), and codes with the identifiable parent properties (IPP codes). TA codes are special IPP codes, and many TA codes implement an efficient identification algorithm which can determine at least one redistributor. However, many IPP codes are not TA codes, in which case, no efficient identification algorithms are available.*

*In this paper, we generalize the definition of TA codes to derive a new family of traceability codes that is much larger than the family of traditional TA codes. By using existing decoding algorithms with respect to the Lee distance, an efficient identification algorithm is proposed for generalized TA codes. Furthermore, we show that the identification algorithm of generalized TA codes can find more redistributors than those of traditional TA codes.*

## 1. Introduction

Detecting copyright infringement is a major global business problem. Some approaches to detecting infringements include monitoring P2P networks and blocking the data transfer and/or identifying the end users [8]. However, since some data transferred using P2P networks is licensed to permit this, a heavy-handed approach to blocking all traffic is not appropriate. Also, while it is possible to monitor the content of (unencrypted) P2P traffic to search for matches on particular hashes of known copyrighted data, maintaining and distributing a list of all such files to all routers in real-time is not feasible. Thus, it makes more sense to encapsulate intellectual property rights within the digital product, and to ensure that access rights can be managed.

Traditional encryption schemes alone do not provide a complete solution to this problem, because they do not prevent authorized users from transferring the cleartext content to other (unauthorized) users. Also, once the transfer has been completed, then there is no means to trace the source of the leak by any encryption scheme [2, 5]. We call the authorized users who produce and redistribute illegal copies of the digital product to unauthorized parties *illegitimate redistributors* (or *traitors* in the literature).

Traceability schemes are an effective solution to this problem [1, 4, 9, 10, 11]. Two types of codes for traceability schemes which have been studied extensively in the literature are (1) codes with the identifiable parent property (IPP) and (2) traceability (TA) codes. The family of IPP codes includes the family of TA codes as a subset. However, IPP codes usually only guarantee that at least one traitor can be theoretically traced back from any pirate (i.e., an illegal copy of a digital product), but an efficient identification algorithm that traces the traitors is not always available. In contrast, TA codes not only guarantee that a traitor can be traced back, but also have efficient identification algorithms.

In the literature, TA codes are defined using the Hamming distance, which is used in the theory of error-control coding to measure the difference between two error-correcting codewords. It is well known that for many linear error-correcting codes (such as Reed-Solomon and algebraic-geometric codes), there are efficient decoding algorithms with high decoding capability in terms of the Hamming distance. These linear codes can be used as TA codes, and their decoding algorithms can be adapted to (or directly used as) identification algorithms (see [1, 10, 11]).

However, as shown in the literature, many IPP codes are not TA codes. For those IPP codes, no efficient identification algorithms are available, and this represents a barrier for those IPP codes to be used in digital rights management.

978-0-7695-3471-8/08 $25.00 © 2008 IEEE
DOI 10.1109/ISPA.2008.99

Also, the TA codes that are designed from error-control codes and have efficient decoding algorithms only form a small subset of the whole family of the IPP codes. So, a major theoretical challenge is to extend TA codes - based on the Hamming distance - to yield more efficient identification algorithms. Any solution to this theoretical challenge would clearly provide a solution for practical applications in DRM.

In this paper, we generalize the definition of TA codes to obtain new traceability codes. The family of generalized TA codes is much larger than the family of traditional TA codes. By using decoding algorithms with respect to the Lee distance in our previous work [12, 13], an efficient identification algorithm is given for generalized TA codes. We will show that the identification algorithm of generalized TA codes can find more redistributors than those of traditional TA codes.

## 2 IPP Codes and TA Codes

Let $\mathcal{A}$ be an alphabet with $|\mathcal{A}| = q$. A code $C$ of length $n$ is a subset of $Q^n$, the set of all $n$-tuples with components in $Q$. If $|C| = M$, we call $C$ a $q$-ary $(n, M)$ code. The elements of $C$ are called *codewords*. In practical applications, each codeword corresponds to a legitimate user of a digital product (or a legal copy of the digital product). A *pirate* (illegal copy) corresponds to an element of $Q^n$. We call any subset of the code, $D \subseteq C$, a group of users. If a group $D$ of users are suspected of colluding to produce a pirate, we call $D$ a *coalition*.

For a coalition $D \subseteq C$, a $n$-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is called a *descendant* of $D$, provided that for all $x_i$ ($i = 1, \ldots, n$),

$$x_i = a_i, \quad \text{for some } (a_1, \ldots, a_i, \ldots, a_n) \in D.$$

The set of all descendants of $D$ is denoted as $\mathrm{desc}(D)$.

Let $t$ be a positive integer. We define

$$\mathrm{desc}_t(C) = \bigcup_{D \subseteq C, \text{ and } |D| \leq t} \mathrm{desc}(D).$$

That is, $\mathrm{desc}_t(C)$ is the set of $n$-tuples that could be produced by a coalition of size at most $t$.

In the literature, traceability codes (TA codes) are defined based on the Hamming distance, which is used in the theory of error-control coding to measure the differences between two error-control codewords.

**Definition 2.1** For any two $n$-tuples $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n) \in Q^n$. The *Hamming distance* between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d_H(\mathbf{x}, \mathbf{y}) = |\{i \mid x_i \neq y_i\}|$$

that is, the number of coordinates where $\mathbf{x}$ and $\mathbf{y}$ differ.

To define a $t$-IPP code or a $t$-TA code $C$, we need to consider all the subsets of $C$ of size at most $t$. It is easy to calculate that for a code $C$ of length $n$, the number of subsets of size at most $t$ is $s = \sum_{i=1}^{t} \binom{n}{i}$.

**Definition 2.2** Suppose $C$ is a code of length $n$. Let $D_i \subseteq C$ ($i = 1, \ldots, s$) be all the subsets of $C$ such that $|D_i| \leq t$. For a positive integer $t \geq 2$,

(1) $C$ is a *$t$-IPP code*, provided that for all $\mathbf{x} \in \mathrm{desc}_t(C)$ it holds that

$$\bigcap_{\{i \mid \mathbf{x} \in \mathrm{desc}(D_i)\}} D_i \neq \emptyset.$$

(2) $C$ is a *$t$-TA code*, provided that for all $D_i$ and any $\mathbf{x} \in \mathrm{desc}(D_i)$ there exist at least one codeword $\mathbf{y} \in D_i$ such that

$$d_H(\mathbf{x}, \mathbf{y}) < d_H(\mathbf{x}, \mathbf{z}), \quad \text{for any } \mathbf{z} \in C - D_i.$$

Denote by $I(\mathbf{x}, \mathbf{y})$ the set of coordinates where $\mathbf{x}$ and $\mathbf{y}$ agree, that is, $I(\mathbf{x}, \mathbf{y}) = \{i \mid x_i = y_i\}$. Then, $|I(\mathbf{x}, \mathbf{y})| = n - d_H(\mathbf{x}, \mathbf{y})$. The condition in the above definition, $d_H(\mathbf{x}, \mathbf{y}) < d_H(\mathbf{x}, \mathbf{z})$, is equivalent to $|I(\mathbf{x}, \mathbf{y})| > |I(\mathbf{x}, \mathbf{z})|$.

**Remark 2.1:** We explain the meaning of the definitions of $t$-IPP codes and $t$-TA codes as follows.

- If $C$ is a $t$-IPP code, then for any pirate $\mathbf{x}$ and any coalition of size at most $t$ which can produce $\mathbf{x}$, we can (theoretically) trace back to at least one member of the coalition, because at least one member of the coalition also appears in all the other coalitions of size at most $t$ which can produce the pirate. However, the definition of $t$-IPP codes does not suggest any algorithm or approach to find out such a member of the coalition, because for a general code there is no known algorithm (except a brute-force search) to find all the coalitions of size at most $t$ which can produce the pirate.
- On the other hand, the definition of $t$-TA codes gives us more information on how to find a suspected traitor. By definition, an identification algorithm could be designed to make use of an algorithm which can find a codeword that is the nearest one (in terms of Hamming distance) to the pirate. It is well known that there exist error-control codes for which efficient decoding algorithms have been found. These decoding algorithms find the codeword which is the nearest one to a given $n$-tuple $\mathbf{x}$, or can even find all the codewords within a given distance to $\mathbf{x}$.

The following is an important result which shows that the family of $t$-IPP codes includes the family of $t$-TA codes as a subset. The proof of this result can be found in [11].

**Proposition 2.1** *A $t$-TA code is a $t$-IPP code.*

As discussed above, the definition of $t$-TA codes provides more information than that of $t$-IPP codes - that is, the definition of $t$-TA codes not only guarantees that we can trace back to at least one redistributor, but also provides information about designing an algorithm to find at least one redistributor. It implies that the family of $t$-TA codes is strictly smaller than the family of $t$-IPP codes. The following example supports this fact.

**Example 2.1** Let $\mathbf{F}_{11}$ be the finite field of 11 elements. Let $C$ be the following code of length 3 over the finite field $\mathbf{F}_{11}$.

$$C = \{(1,0,0),\ (4,1,1),\ (5,1,1)\}.$$

The code $C$ is a 2-IPP code; while it is not a 2-TA code.

In fact, the symbols in the first position of all the codewords are different. Thus, for any pirate $\mathbf{x} \in \mathbf{F}_{11}^3$, every coalition of size at most 2 which can produce $\mathbf{x}$ must contain the codeword which has the same first coordinate with the pirate as a common codeword.

Consider a pirate $\mathbf{x} = (1,1,1)$. Obviously, it is a descendant of the following coalition $D = \{(1,0,0),\ (4,1,1)\}$. Now,

$$d_H((1,1,1),\ (1,0,0)) = 2,$$

and

$$d_H((1,1,1),\ (4,1,1)) = 1.$$

$C - D = \{(5,1,1)\}$ and $d_H((1,1,1),\ (5,1,1)) = 1$. Thus, there is no codeword $\mathbf{y} \in D$ satisfying

$$d_H(\mathbf{x}, \mathbf{y}) < d_H(\mathbf{x}, \mathbf{z}), \quad \text{for any } \mathbf{z} \in C - D.$$

Therefore, $C$ is not a 2-TA code. ◇

Actually, $t$-TA codes only form a small subset of the whole family of $t$-IPP codes. There are a lot of $t$-IPP codes that are not $t$-TA codes (see [10, 11]).

## 3 Generalized TA Codes

In this section, we generalize the definition of $t$-TA codes to obtain new traceability codes.

**Definition 3.1** Suppose $C$ is a code of length $n$. Let $t \geq 2$ be an integer. Let $D_i \subseteq C$ ($i = 1, \ldots, s$) be all the subsets of $C$ such that $|D_i| \leq t$. We call $C$ a *generalized $t$-TA code* (denoted by *$t$-GTA code* for short), provided that there exist a well-defined distance $d(\cdot,\cdot)$, such that for all $D_i$ and any $\mathbf{x} \in \mathrm{desc}(D_i)$, there exist at least one codeword $\mathbf{y} \in D_i$ such that

$$d(\mathbf{x}, \mathbf{y}) < d(\mathbf{x}, \mathbf{z}), \quad \text{for any } \ \mathbf{z} \in C - D_i.$$

**Remark 3.1:** As the Hamming distance is a well-defined distance, the definition above is obviously a generalization of the definition of $t$-TA codes. Thus, the family of $t$-GTA codes includes all the $t$-TA codes as special members. We will give an example which shows that, on the other hand, $t$-GTA codes are not necessary $t$-TA codes.

First let us introduce another useful distance, namely, Lee distance. Let $q$ be a prime power. Let $\mathbf{Z}_q$ be the ring of integers modulo $q$. For any $a \in \mathbf{Z}_q$, the *Lee value* of $a$, denoted by $|a|$, is the nonnegative integer $\min\{a, q-a\}$. Consider the finite field $\mathbf{F}_q$ with $q$ elements. We define the Lee values of elements in $\mathbf{F}_q$ as follows. Let

$$\begin{array}{ccc} \mathbf{F}_q & \longrightarrow & \mathbf{Z}_q \\ \alpha & \longmapsto & \bar{\alpha} \end{array}$$

be an appropriate one-to-one mapping. Then, the Lee value $|\alpha|$ of $\alpha \in \mathbf{F}_q$ is defined as $|\alpha| = |\bar{\alpha}|$. For a $n$-tuple $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbf{F}_q^n$, the *Lee weight* is defined as $\|\mathbf{x}\|_L = \sum\limits_{i=1}^{n} |x_i|$. The *Lee distance* between two $n$-tuples $\mathbf{x}$ and $\mathbf{y}$ in $\mathbf{F}_q^n$, denoted by $d_L(\mathbf{x}, \mathbf{y})$, is defined as the Lee weight of $\mathbf{x} - \mathbf{y}$.

**Example 3.1** Consider the code over $\mathbf{F}_{11}$ given in Example 2.1,

$$C = \{(1,0,0),\ (4,1,1),\ (5,1,1)\}.$$

It has been proven that $C$ is not a 2-TA code. Now, we prove that with respect to the Lee distance, $C$ is a 2-GTA code.

Let us consider the Lee distance $d_L(\cdot,\cdot)$. The following are all the subsets of $C$ of size 2:

$$D_1 = \{(1,0,0),\ (4,1,1)\},$$

$$D_2 = \{(1,0,0),\ (5,1,1)\},$$

and

$$D_3 = \{(4,1,1),\ (5,1,1)\}.$$

We show that $C$ is a 2-GTA code with respect to the Lee distance, by proving that all $D_i$ satisfy the condition: For any $\mathbf{x} \in \mathrm{desc}(D_i)$, there is a codeword $\mathbf{y} \in D_i$ such that $d_L(\mathbf{x}, \mathbf{y}) < d_L(\mathbf{x}, \mathbf{z})$, for any $\mathbf{z} \in C - D_i$.

First, considering $D_1$ we have

$$\begin{aligned} \mathrm{desc}(D_1) = \{&(1,0,0),\ (4,1,1),\ (1,0,1),\ (1,1,0), \\ &(1,1,1),\ (4,0,0),\ (4,0,1),\ (4,1,0)\}. \end{aligned}$$

For $(1,0,0)$ and $(4,1,1)$, as they are in $D_1$, they have Lee distance 0 to themselves. Thus, the condition above is satisfied. Look at $(1,0,1)$,

$$d_L((1,0,1),\ (1,0,0)) = 1 < d_L((1,0,1),\ (4,1,1)) = 4$$

and

$$d_L((1,0,1),\ (1,0,0)) = 1 < d_L((1,0,1),\ (5,1,1)) = 5.$$

Thus, for $\mathbf{x} = (1,0,1) \in \operatorname{desc}(D_1)$, the codeword $(1,0,0) \in D_1$ is such a $\mathbf{y}$ satisfying the above condition. Now, for any of $(1,1,0)$, $(1,1,1)$, $(4,0,0)$, $(4,0,1)$, $(4,1,0)$, we can similarly find a $\mathbf{y} \in D_1$ such that $d_L(\mathbf{x},\mathbf{y}) < d_L(\mathbf{x},\mathbf{z})$, for any $\mathbf{z} \in C - D_1$. Therefore, $D_1$ satisfies the above condition.

Similarly, we can prove that $D_2$ and $D_3$ both satisfy the above condition. Therefore, $C$ is a 2-GTA code with respect to the Lee distance. ◇

We now present an interesting result.

**Theorem 3.1** *Any $t$-GTA code is a $t$-IPP code.*

*Proof:* Suppose $C$ is a $t$-GTA code with respect to a well-defined distance. Let $\mathbf{x} \in \operatorname{desc}_t(C)$. Then there is a subset $D_i \subseteq C$, where $|D_i| = t$, such that $\mathbf{x} \in \operatorname{desc}(D_i)$. Let $\mathbf{y} \in D_i$ such that $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z})$ for every $\mathbf{z} \in D_i$. Then $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z})$ for any $\mathbf{z} \in C$ by the definition of $t$-GTA code.

We shall prove that for any $D_j \subseteq C$ with $|D_j| \leq t$, if $\mathbf{x} \in \operatorname{desc}(D_j)$ then $\mathbf{y} \in D_j$. In fact, if $\mathbf{y} \notin D_j$, then there is a $\mathbf{w} \in D_j$ such that $d(\mathbf{x},\mathbf{w}) < d(\mathbf{x},\mathbf{y})$ by the definition of $t$-TA codes. This contradicts the fact that $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z})$ for any $\mathbf{z} \in C$. ◇

From Example 3.1 and Theorem 3.1, we see that the $t$-GTA codes have the following two advantages:

- The family of $t$-GTA codes includes the family of $t$-TA codes as a subset. There exist $t$-GTA codes that are new traceability codes (that is, they are not traditional $t$-TA codes). So for applications to digital rights management, using $t$-GTA codes, the digital industries have more choices than using $t$-TA codes.

- $t$-GTA codes are still members of the family of $t$-IPP codes. Inheriting the property of $t$-IPP codes, $t$-GTA codes guarantee that at least one traitor can be traced back.

In next section, we will see a third advantage of $t$-GTA codes, that is, there exist an identification algorithm for $t$-GTA codes which can reveal more traitors than those of $t$-TA codes.

# 4 Identification Algorithms of Generalized TA Codes

In this section, we propose an efficient identification algorithm for $t$-GTA codes with respect to the Lee distance. We also compare our identification algorithm with the identification algorithm for $t$-TA codes in [10], which was designed making use of the well-known list-decoding algorithm [7].

The main result on identification algorithm in [10] is as follows. (The identification algorithm in [10] is applicable for TA codes based on Reed-Solomon, algebraic-geometric, and concatenated codes. For simplicity, we only state the result for TA codes based Reed-Solomon codes.)

**Proposition 4.1** *Let $C$ be a $[n,k]$ Reed-Solomon code. Let $t$ be an integer and $t < \sqrt{\frac{n}{k-1}}$.*

*(1) $C$ is a $t$-TA code.*

*(2) Suppose $\mathbf{x} \in \operatorname{desc}_t(C)$. Then, there exists an codeword $\mathbf{y} \in C$ such that $d_H(\mathbf{x},\mathbf{y}) \leq n - n/t$. And for every $\mathbf{y} \in C$ satisfying $d_H(\mathbf{x},\mathbf{y}) \leq n - n/t$,*

$$\mathbf{y} \in \bigcap_{\mathbf{x} \in D_i,\ |D_i| \leq t} D_i.$$

*Thus, $\mathbf{y}$ is a traitor.*

*(3) The identification algorithm finds all the traitors $\mathbf{y} \in C$ that satisfy*

$$d_H(\mathbf{x},\mathbf{y}) \leq n - n/t.$$

**Example 4.1** Consider a $[n,k]$ Reed-Solomon code over the finite field $\mathbf{F}_{13}$. Suppose $n = 12$ and $k = 2$. Let $t = 3$. It is easy to verify the condition

$$t = 3 < \sqrt{\frac{n}{k-1}} = \sqrt{12}.$$

Thus, from the proposition above, $C$ is a 3-TA code. For any $\mathbf{x} \in \operatorname{desc}_3(C)$, the identification algorithm in [10] can finds the traitors $\mathbf{y} \in C$ satisfying $d_H(\mathbf{x},\mathbf{y}) \leq n - n/t = 8$.

Now, consider the coalition

$$\begin{aligned} D = \{&(0,0,0,0,0,0,0,0,0,0,0,0), \\ &(1,1,1,1,1,1,1,1,1,1,1,1)\} \subseteq C. \end{aligned}$$

$\mathbf{x} = (1,1,1,1,1,1,1,1,1,1,0,0)$ is a pirate, and $\mathbf{x} \in \operatorname{desc}(D)$. We have

$$d_H(\mathbf{x}, (0,0,0,0,0,0,0,0,0,0,0,0)) = 10,$$

and

$$d_H(\mathbf{x}, (1,1,1,1,1,1,1,1,1,1,1,1)) = 2 < 8.$$

Thus, the identification algorithm in [10] can find the traitor $(1,1,1,1,1,1,1,1,1,1,1,1)$, but can not find the traitor $(0,0,0,0,0,0,0,0,0,0,0,0)$. ◇

In [12, 13], generalizing the list-decoding algorithm [7], an efficient decoding algorithm is proposed for Reed-Solomon and algebraic-geometric codes with respect to the Lee distance. (See [12, 13] for the algorithm and complexity evaluation). The decoding capability of the algorithm is given as follows.

**Theorem 4.2** *Let $C$ be a $[n, k]$ Reed-Solomon code or algebraic-geometric code over $\mathbf{F}_q$. Then, for any word $\mathbf{x} \in C$, the decoding algorithm finds all the codewords $\mathbf{y} \in C$ that satisfy $d_L(\mathbf{x}, \mathbf{y}) \leq \tau$, where*

$$\tau = (u+1)(n - \left\lfloor \sqrt{(2u+1)n(k-1)} \right\rfloor - 1),$$

*where $u$ is any integer with $0 \leq u \leq (q-1)/2 - 1$.*

Now an efficient identification algorithm for $t$-GTA codes with respect to the Lee distance is given as follows.

**Identification Algorithm:** For a $t$-GA code $C$ based on a Reed-Solomon code or an algebraic-geometric code, the decoding algorithm in [12, 13] is used directly as an identification algorithm. For any pirate $\mathbf{x}$, the identification algorithm can find all the traitors $\mathbf{y} \in C$ satisfying

$$d_L(\mathbf{x}, \mathbf{y}) \leq (u+1)(n - \left\lfloor \sqrt{(2u+1)n(k-1)} \right\rfloor - 1).$$

In the following example, we show that the identification algorithm can find more traitors than the identification algorithm in [10].

**Example 4.2** Consider the same 3-TA code in Example 4.1. Let $u = 1$. Then

$$\tau = (u+1)(n - \left\lfloor \sqrt{(2u+1)n(k-1)} \right\rfloor - 1) = 10.$$

The identification algorithm can find all the traitors $\mathbf{y}$ satisfying

$$d_L(\mathbf{x}, \mathbf{y}) \leq 10.$$

Now,

$$d_L(\mathbf{x}, (0,0,0,0,0,0,0,0,0,0,0,0)) = 10,$$

and

$$d_L(\mathbf{x}, (1,1,1,1,1,1,1,1,1,1,1,1)) = 2.$$

Thus, the identification algorithm finds both the traitor $(1,1,1,1,1,1,1,1,1,1,1,1)$ and the traitor $(0,0,0,0,0,0,0,0,0,0,0,0)$. In this case, the identification algorithm outperforms the identification algorithm in [10]. ◇

# 5 Applications in DRM Systems and Malware Proliferation

In this section, we consider the applications of traceability codes to electronic data distribution systems, pay-TV systems, pay services on the Web, and other DRM systems.

There are two traitor-tracing models for various DRM systems. The first model was proposed by Chor et al. [4] to protect ownership rights in electronic data distribution systems. To do so, Chor et al. introduced a new form of cryptography that uses one key to encrypt the content. And there are multiple distinct decryption keys; each of them can decrypt the ciphertext.

Prior to being distributed, the digital product is divided into multiple segments. The $i$-th segment is encrypted by using an encryption key $e_i$. There are multiple distinct decryption keys, $d_i^{(1)}, d_i^{(2)}, d_i^{(3)} \ldots$; any of them can be used as a decryption key to recover the $i$-th segment. In this model, each legitimate subscriber, $u_k$, is given a sequence of $n$ decryption keys $d_1^{(k)}, d_2^{(k)}, \ldots, d_n^{(k)}$, where $n$ is the number of segments. The user $u_k$ uniquely corresponds to a key-set $(d_1^{(k)}, d_2^{(k)}, \ldots, d_n^{(k)})$, which is actually a codeword of a TA or generalized TA code $C$. A coalition $D = \{u_1, \ldots, u_k, \ldots\}$ can make a pirate key-set $(d_1, d_2, \ldots, d_n)$, where each $d_i$, $i = 1, \ldots, n$, is an decryption key belonging to some member in $D$. Once a pirate key-set has been observed, by using an identification algorithm, at least one traitor can be traced back and legal means can be taken.

However, this model works only under the assumption that traitors provide unauthorized users with decryption keys capable of decoding the original content. It would be ineffective if the traitors were simply to re-distribute the original content.

A second model was proposed by Fiat and Tassa [6] to overcome the shortcoming of the first model. Furthermore, the second model is a dynamic model in terms of that the feedback from the pirate broadcasting network is used, and the traitors are blocked as soon as they are detected. This model is effective for broadcast systems including pay-TV systems and pay services on the Web. Another application of the dynamic model to some conditional access schemes was also discussed by Fiat and Tassa (see [6] for the detail).

In the second model, similarly the content consists of multiple segments (e.g., a segment could be 1 minute's worth of video). Denote by $P_1, P_2, \ldots, P_n$ all the segments. A traceability code $C$ of length $n$ is used. For any codeword $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in C$, the $i$-th component $c_i$ is inserted (by using a watermarking scheme) into the $i$-th segment $P_i$ to generate a variant, denoted by $P_i\langle c_i \rangle$. Here the components of the codewords are short enough and watermarks are generated in a way such that all variants carry the same information to the extent that humans cannot distinguish between them easily. The legitimate subscriber who gets the copy $(P_1\langle c_1\rangle, P_2\langle c_2\rangle, \ldots, P_n\langle c_n\rangle)$ uniquely corresponds to the codeword $\mathbf{c} = (c_1, c_2, \ldots, c_n)$. As soon as a pirate copy is observed, the watermarks in the pirate copy would be retrieved, the identification algorithm would then find out at least one of the codewords that contributed to the pirate copy. Thus, a traitor is found; and the system would then disable his access to the content.

A detailed description of the dynamic model (including the issue of controlling what users get what variant of every segment as well as an analysis of the broadcast overhead for implementing such a traitor-tracing scheme) is given in [6]. It is easy to see that to implement the dynamic model for real-time traitor-tracing, the family of traceability codes should meet the following two requirements: (1) The family has many traceability codes; and (2) There are efficient identification algorithms available for these TA codes. Our results in previous sections extend the family of traceability codes and provides much more effective traceability codes. More importantly, our generalized TA codes with respect to the Lee distance have efficient identification algorithms.

We are currently working on applying our generalized TA codes to the problem of tracing changes in malware code, such as rootkits - generated by distinct, ”off the shelf” kits - to study the activity of different groups. Our investigation is focusing on tracking ”pirate” copies using a honeynet-style approach, such that when a kit is actively being used on the Internet, the identification algorithm can find out at least one of the codewords that contributed to the pirate copy. The forensic details of their activity can be passed onto law enforcement authorities. The practical problem is gaining access to malware sources to insert the codes, without breaking the law by participating in a crime.

## 6 Conclusion

By generalizing the definition of TA codes, we have been able to generate new traceability codes. The family of our generalized TA codes is much larger than the family of traditional TA codes. By using a decoding algorithm with respect to the Lee distance that we proposed in our previous work, an efficient identification algorithm is given for generalized TA codes. We also show that the identification algorithm for generalized TA codes can find more redistributors than those of traditional TA codes. Following this work, we are studying the problem of designing identification algorithms for generalized TA codes with respect to other metrics (see [3]), and practical applications.

## References

[1] A. Barg, G. Blakley and G. Kabatiansky, “Digital fingerprinting codes: problem statements, constructions, identification of traitors,” *IEEE Trans. Inform. Theory*, vol. 49, no.4, April 2003, pp. 852-865.

[2] D. Boneh and J. Shaw, “Collusion secure fingerprinting for digital data,” *IEEE Trans. Inform. Theory*, vol. 44, Sept. 1998, pp. 1897-1905.

[3] E. Candes and T. Tao, “Decoding by linear programming,” *IEEE Trans. Inform. Theory*, vol. 51, no. 12, Dec. 2005, pp. 4203-4215.

[4] B. Chor, A. Fiat, M. Naor and B. Pinkas, “Tracing traitors,” *IEEE Trans. Inform. Theory*, vol. 46, May 2000, pp. 893-910.

[5] M. Fernandez, M. Soriano and J. Cotrina, “Tracing illegal redistribution using error-and-erasures and side information decoding algorithms,” *IET Inf. Secur.*, vol. 2, no. 1, 2007, pp. 83-90.

[6] A. Fiat and Tassa, “Dynamic traitor tracing,” *J. Cryptology*, vol.14, 2001, pp. 211-223.

[7] V. Guruswami and M. Sudan, “Improved decoding of Reed-Solomon and algebraic-geometry codes,” *IEEE Trans. Inform. Theory*, vol. 45, Sept. 1999, pp. 1757-1767.

[8] J. Mee and P.A. Watters, “Detecting and tracing copyright infringements in P2P networks,” *Proceedings of the International Conference on Networking*, 2005, pp. 60-65.

[9] R. Safavi-Naini and Y. Wang, “Sequential traitor tracing,” *IEEE Trans. Inform. Theory*, vol. 49, no.5, May 2003, pp. 1319-1326.

[10] A. Silverberg, J. Staddon, J. Walker, “Applications of list decoding to tracing traitors,” *IEEE Trans. Inform. Theory*, vol. 49, no.5, May 2003, pp. 1312-1318.

[11] J. Staddon, D. Stinson, and R. Wei, “Combinatorial properties of frameproof and traceability code,” *IEEE Trans. Inform. Theory*, vol. 47, no.3, March 2001, pp. 1042-1049.

[12] X.-W. Wu, M. Kuijper, and P. Udaya, “Lee-Metric Decoding of BCH and Reed-Solomon Codes,” *IEE Electronics Letters*, Vol.39, No.21, 2003, pp.1522-1524.

[13] X.-W. Wu, M. Kuijper, and P. Udaya, “On the Decoding Radius of Lee-Metric Decoding of Algebraic-Geometric Codes,” *Proceedings of 2005 IEEE International Symposium on Information Theory*, Adelaide, Australia, 5 - 9 September, 2005, pp.1191-1195.