UPGRADE
The European Online Magazine for the IT Professional
http://www.upgrade-cepis.org
Vol. III, No. 6, December 2002

UPGRADE is the European Online Magazine for the Information Technology Professionals, published bimonthly at <http://www.upgrade-cepis.org/>.

Publisher
UPGRADE is published on behalf of CEPIS (Council of European Professional Informatics Societies, <http://www.cepis.org/>) by NOVÁTICA <http://www.ati.es/novatica/>, journal of the Spanish CEPIS society ATI (Asociación de Técnicos de Informática <http://www.ati.es/>).
UPGRADE is also published in Spanish (full issue printed, some articles online) by NOVÁTICA, and in Italian (abstracts and some articles online) by the Italian CEPIS society ALSI <http://www.alsi.it> and the Italian IT portal Tecnoteca <http://www.tecnoteca.it/>.
UPGRADE was created in October 2000 by CEPIS and was first published by NOVÁTICA and INFORMATIK/INFORMATIQUE, bimonthly journal of SVI/FSI (Swiss Federation of Professional Informatics Societies, <http://www.svifsi.ch/>).

E-mail addresses for editorial correspondence: <nicolet@acm.org> and <rfcalvo@ati.es>

E-mail address for advertising correspondence: <novatica@ati.es>

ISSN 1684-5285

Coming issue:
"Human-Computer Interaction"

Security in e-Commerce

Guest Editors: Javier Areitio-Bertolín, Javier López-Muñoz, José A. Mañas-Argemí, and Stephanie Teufel

Joint issue with NOVÁTICA

Fingerprinting Schemes for the Protection of Multimedia Distribution Rights

Marcel Fernández-Muñoz, Miquel Soriano-Ibañez, Josep Domingo-Ferrer, and Francesc Sebé-Feixas

It is a broadly accepted opinion that selling digital contents through computer networks is one of the most natural applications of electronic commerce, e-Commerce, because even distribution can be done in real time over the network at the moment of purchase. Nonetheless, current turnover in e-Commerce is lower than had been anticipated a few years ago. One of the problems that is hampering market development is the difficulty of adequately protecting the distribution rights of contents being sold. Fingerprinting is the most usual solution, and consists of uniquely marking the object before its distribution. One possible attack against fingerprinting schemes is through collusion of several dishonest users, who try to fabricate a copy that does not reveal their identities when redistributed. Several kinds of codes exist that provide redistributor tracing in case of collusion attacks. This paper provides a classification and a description of some of them. It also discusses coding and decoding for each scheme.

Keywords: Collusion, Copy Detection, Distribution of Multimedia Content, Error Correction Codes, Fingerprinting, Multimedia Copyright Protection.

1 Introduction

One of the advantages the Internet provides is convenient distribution of digital products; evidence that this is so is that a great number of economic sectors are using the network as a natural medium for information interchange. Using documents in digital form not only results in easier distribution and management, but it also facilitates honest or dishonest transformation of documents by the user community. Therefore, new problems related to copyright and distribution protection of digital products arise.

Cryptographic techniques are not adequate for solving such problems, as there are doubts about how the receiver will behave once he/she has received the product. The sentence of a Californian court of appeal against Napster, published in February 2001, showed that distribution of multimedia contents

Marcel Fernández-Muñoz received his degree in Telecommunications Engineering from the Universitat Politècnica de Catalunya (Barcelona, Spain) in 1998. Since 2002 he has been an Assistant Professor at the same university, where he is currently working toward his Ph.D. degree in Telematics. His interests are digital content protection and error correcting codes. He has authored more than 10 publications in these fields. <marcelf@entel.upc.es>

Miquel Soriano-Ibañez received his degree in Telecommunication Engineering and his Ph.D., both with honours, from the Universitat Politècnica de Catalunya, (Barcelona, Spain) in 1992 and 1996, respectively. Since 1997 he has been an associate professor at the same university, where he is also Vicedean of the School of Telecommunications Engineering. His current research interests include data protection, network security and electronic commerce. He has published over 50 papers in these areas in international journals and conferences. He also frequently serves as a consultant to government and industry. <soriano@entel.upc.es>

Josep Domingo-Ferrer received a B.Sc. in Computer Engineering from the Universitat de Lleida (Spain) in 1999 and an M.Sc. in Computer Engineering from the Universitat Rovira i Virgili (Tarragona, Spain) in 2001. He is currently a predoctoral researcher at the University Rovira i Virgili and is working toward his Ph.D. degree in Telematics at the Universitat Politècnica de Catalunya (Barcelona, Spain). His interests are computer security, secure e-commerce, cryptography and digital content protection. He has authored more than 10 publications in these fields.
<jdomingo@etse.urv.es>

Francesc Sebé-Feixas received his M.Sc. and Ph.D. degrees with honours in Computer Science from the Universitat Autònoma de Barcelona, Spain, in 1988 and 1991, respectively. He also holds an M.Sc. in Mathematics. He is currently an Associate Professor at the Universitat Rovira i Virgili (Tarragona, Spain) where he is also Dean of the School of Engineering. His fields of expertise are data security, cryptography and inference control. In these fields, he has authored two patents and over 80 publications, and has won US Government, European Commission and Spanish research contracts. He has chaired the program committees of Statistical Data Protection'98 (sponsored by Eurostat) and IFIP CARDIS'2000 and has served on more than 15 program committees of IT security conferences. He has been a guest editor of the Computer Networks and Intl. Journal of Uncertainty, Fuzziness and Knowledge Based Systems and has edited two international books. He currently is a Senior Member of IEEE and serves as Chairman of the IEEE Information Theory Spanish Chapter.
<fsebe@etse.urv.es>

through the Internet without respecting intellectual property rights is an illegal activity.

Copyright protection can be achieved in two ways: *a priori* protection (copy prevention) and *a posteriori* protection (copy detection). The recent failure of the DVD copy prevention system [3] shows the lack of effectiveness of *a priori* protection. In the scientific community, the belief exists that *a posteriori* protection is the only viable mechanism.

This fact has fostered the development of new techniques allowing copyright protection in the new digital economy. Watermarks are based on embedding a mark in the copy sold. The same mark is embedded in all copies, so that intellectual property can be protected but not distribution rights. Fingerprinting (introduced by Wagner in 1983 [20]) consists of using marks that identify not only the content owner but the buyer as well, in such a way that each copy is personalized and its distribution can be traced. Unlike watermarking, collusion attacks are feasible for fingerprinting, i.e. several buyers can come together and pool their copies so as to remove the mark or generate a copy with a new mark different from those they were assigned. Consequently, an adequate fingerprinting scheme must allow identification of dishonest buyers who took part in the collusion. Detection of such dishonest buyers can be achieved by using schemes with tracing properties, as described in [6].

Fingerprinting schemes can be classified as symmetric, asymmetric and anonymous.

a) In *symmetric schemes*, only the merchant participates in the marking procedure. This leaves the buyer unprotected against fraudulent merchants in that buyer A can be falsely accused of illegal distribution if the merchant provides another buyer B with the copy bearing A's mark.
b) *Asymmetric fingerprinting* aims at solving such a problem by preventing the merchant from seeing the marked copy delivered to the buyer, but it does so in such a way that the merchant can identify the buyer if the latter unlawfully redistributes his or her copy. To do this, both the merchant and the buyer take part in the marking procedure. The merchant will be able to identify the buyer's mark, but cannot generate it without the buyer's presence. Asymmetric schemes pose the problem of lack of anonymity: the merchant knows the buyer's identity.
c) *Anonymous fingerprinting* solves the aforementioned inconvenience. In it, the merchant does not know the buyer's marked copy nor his/her identity. This fact does not prevent buyer identification in case of illegal redistribution. The system is based on a trusted third party (register authority) who knows the buyer's real identity.

Fingerprinting techniques involve the following issues:

1. What kind of mark to use.
2. Mark location and embedding technique being used.
3. Algorithms that allow colluder identification from the obtained mark in a reasonable time.

The second issue has been treated in several papers. The best option depends on the type of file to be marked (audio, video, etc.). Among others, the work by Sebé et al. [19] should be mentioned here.

In this paper, we will focus on the kind of mark to be used and the algorithms, which are largely based on error correcting codes, to be applied. In fact, codewords can be used as *fingerprinting codewords*. Every distributed copy is assigned a different codeword that will be embedded into the document by the marking algorithm.

Using the current techniques for error control, if the number of symbols e in which the received word differs from the nearest codeword is lower than the half the minimum distance of the code, the decoder will generate a single codeword as output, whose distance to the received word is e. However, if the number of differing symbols is greater than the previous bound, uniqueness is not guaranteed. In any case, there is no assurance that the nearest codeword is related to a dishonest distributor. So, other coding techniques are needed.

Different proposals by the authors will be presented in the rest of this paper whose goal is to satisfy the requirements of different scenarios. In Section 2, a taxonomy of codes based on their traceability properties is presented. Section 3 presents an IPP, Identifiable Parent Property, scheme. Section 4 analyses coding and decoding of a TA, Traceability, scheme with soft-decision techniques. Binary fingerprinting schemes robust against collusions of up to 2 and 3 buyers are presented in Sections 5 and 6, respectively. Finally, conclusions are detailed in Section 7.

2 Classification of Codes with Tracing Properties

The ideal code would be one which allowed all participants in an illegal collusion to be identified, regardless of the collusion size. It can be proven that no such ideal code exists.

From now on, we assume that words that can be generated by a collusion of w users constitute the descendant of all their codewords $a^1, a^2, a^3, ..., a^w, \in F_q^n$, where the descendant is defined as follows:

$$desc(a^1, a^2, a^3, ..., a^w) := \{x \in F_q^n : x_i \in \{a_i^1, a_i^2, a_i^3, ..., a_i^w\}, 1 \le i \le n\}$$

In other words, if several users collude and pool their corresponding marked files, it is assumed that they take certain symbols from each copy whenever there is discrepancy (mark detection).

Given a code C, the descendant code C^* is defined as:

$$C^* := \bigcup_{a^1 \in C, ..., a^w \in C} desc(a^1, ..., a^w)$$

Depending on the features offered, the following classification can be established:

- FC, Frameproof Codes: a code is w-FC if no collusion with a number of participants $\le w$ can frame an innocent user who did not participate in the collusion, by generating the mark corresponding to that innocent user.
- SFC, Secure Frameproof Codes: a code is w-SFC if no collusion with a number of participants £w can frame another disjoint collusion of size also $\le w$ by generating a mark which could also have been generated by the second collusion.
- IPP, Identifiable Parent Property: a code is w-IPP if no collusion with a number of participants $\le w$ can generate a n-tuple which prevents identification of the colluders.

- TA:, Traceability: a code is w-TA if, for any n-tuple generated by a collusion of at most w users, the nearest codeword corresponds to one member who has taken part in the collusion. In these schemes, it is enough to use common channel decoding techniques in order to find a dishonest user.

It can be shown that the above classification presents different schemes from the least to the most restrictive, and each class is included in the previous one, *i.e.*, a code w-TA is also w-IPP; w-IPP implies w-SFC and w-SFC implies w-FC.

In [4], Boneh and Shaw introduced the concept of secure fingerprinting against buyers collusions. The authors proved that no binary code completely IPP exists for $w \geq 2$ and proposed a general construction to obtain codes secure against collusions of up to *w* buyers that allows identification of at least one colluder with probability 1–ε. Such codes are called w-secure with error ε. For a community of N possible buyers and given $\varepsilon>0$, $L=2w\log(2N/\varepsilon)$ and $d=8w^2\log(8wL/\varepsilon)$ a code with N codewords of length

$$l=2Ldw=32w^4\log(2N/\varepsilon)\log(8wL/\varepsilon)$$

is constructed which allows identification of one of the colluders with probability 1–ε.

3 Schemes 2-IPP q-ary

When an altered mark (a word that does not belong to the code) is received, it can usually have several possible parent couples. If the code is IPP, the intersection of all these couples is not empty, so that one of the dishonest users can be accused with absolute certainty.

3.1 Coding

Hollmann et al. [12] proved that, given an integer q, power of a prime p, a Reed-Solomon code (shortened or extended) exists over F_q with $(n, \lceil n/4\rceil, n-\lceil n/4\rceil+1)$ that is 2–IPP $(q \geq n-1)$.

3.2 Decoding

IPP decoding algorithms are precisely based on searching all possible couples that can be parents of the received word, and finding the intersection between all of them.

In [16] Silverberg et al. presented algorithms that return a list of all possible parents of a received word using list decoding [11] techniques. In [8] Fernández and Soriano presented a new proposal for IPP codes and a methodology that includes its advantages [16].

When two dishonest users build the altered mark, they choose between two possible symbols in the positions in which their codewords differ. In the case of a Reed-Solomon code, given a descendant, the two error patterns (one for each parent) can have a Hamming weight of at most n/2, which is far beyond the error-correcting capacity of the code. When IPP codes are used, given a descendant, at least one codeword exists which agrees with it in a number of symbols $>2(n-d)$ because, if there are only two possible parents, one of them must contribute a number of symbols greater than or equal to n/2. Using the value of d for which the code is IPP, the proof is immediate. Moreover, it can be proven that such words are parents of the descendant.

As Berlekamp indicates in [2], re-encoding is a possible way to find the error pattern. This model considers that all information symbols in the received codeword are correct, *i.e.* that all errors have occurred in redundancy symbols. This idea is interesting for decoding Reed-Solomon IPP codes because, if the descendant word has at least k symbols from one of its parents, it can be recoded using such symbols as though they were information. In this way, all possible parents contributing at least k symbols to the descendant can be obtained.

The algorithm proposed in [8] seeks the positive_parent (the one agreeing with the descendant in $\geq 2(n-d)$ positions) and constructs all possible couples of parents. The generic idea is based on re-encoding the symbols of the received word that do not agree with those of the positive_parent.

4 w-TA Schemes Based on Soft Decision Decoding

4.1 Coding

In [16], it is proven that, given a code C[n,k,d], if $d>n(1-1/w2)$, C is a w-TA code.

4.2 Decoding

The objective of the decoding algorithm is to find a list containing all parents whose combination yields the received word. However, since such parents can contribute only very few positions, identification of all of them is not possible. In fact, as has been stated in Section 2, IPP codes (and consequently also w-TA) guarantee detection of one of the parents (for whom participation in the collusion is positive). We will call this parent a positive_parent.

When soft-decision decoding techniques are used, the decoder takes advantage of side information generated by the receiver and, instead of managing received symbols directly, it uses probability values about reliability of such symbols. Such information is usually given to the algorithm as a reliability matrix. The soft-decision algorithm used is Koetter-Vardy's, KV [13].

The idea of the soft-decision based tracing algorithm is to obtain information about positive_parents and prepare it so that the decoder sees it as if coming from a channel. From this information, we want to find a greater number of positive_parents. This is achieved by performing several iterations of the KV algorithm, in which every iteration builds a new reliability matrix that takes into account positive_parents found in previous iterations.

In the case of w colluders, there will be at most w parents, so that one of them must contribute at least $\lceil n/w\rceil$ descendant symbols. Since $d>n-n/w^2$, we can guarantee that there always exists a codeword that agrees with the descendant in $\geq w(k-1)+1$ symbols. Moreover, in [9] it is proven that such a codeword is a positive_parent. Once a positive_parent is known, new bounds to identify other possible positive_parents are iteratively established, using the known parent(s).

An informal description of the tracing algorithm is presented next; the complete description can be found in [9].

Algorithm:

Input: w: positive integer; C: Reed-Solomon code with length n and minimum distance $d>n-n/w^2$; descendant (received word) $\mathbf{p}\in \mathbf{des}_C(C_t)$, $|Ct|\leq w$.

Output: A list L of all positive_parents of **p**.

1. Given a received word **p**, a reliability matrix is constructed assuming that all received symbols are correct and that all parents contribute with the same amount of symbols.
2. With this matrix R, run the KV algorithm. From the obtained list, codewords $u_{i1}, ..., u_{ij}$, are taken whose distance to **p** is $\leq n-(w(k-1)+1)$ and they are added to the list L.
3. If all parents have been found ($|L|=w$), or if the number of positions of **p** to be "covered" is such that there are no more positive_parents, then output L and quit the algorithm.
4. State the inputs to a new reliability matrix, in which:
   a) Symbols of **p** covered by positive_parents found up to now are assumed to be "erasures".
   b) Non covered symbols are assumed to be correct.
   It is also assumed that all positive_parents not found yet contribute with the same number of symbols to the construction of **p**.
5. Run the KV algorithm on the matrix from the previous step. From the obtained list, take those codewords $u_{i1}, ..., u_{ij}$ that are positive_parents and add them to L.
6. If all parents have been found($|L|=w$), or if the number of positions of **p** to be "covered" is such that there are no more positive_parents, then output L and exit the algorithm; otherwise go to Step 4.

5 Construction of a Binary Fingerprinting Scheme Secure against Collusions of Size 2

One of the parameters taken into account when designing a code is redundancy. In order to minimize code redundancy, one possible technique to use is code concatenation.

A concatenated code or *supercode* is the combination of a $[n_i, k_i, d_i]$ q_i-ary code ($q_i \geq 2$), called internal code, with an $[n_o,k_o,d_o]$ q_i^{ki}-ary code, called external code. Combination consists of a mapping among internal code codewords and the elements from Fq_i^{ki}, which yields a q_i-ary code of length n_in_o and dimension k_ik_o. Note that the size of the resulting concatenated code corresponds to that of the external code.

In order to construct a binary fingerprinting code C we will use:

- As internal code, a dual binary Hamming code, S_r, with parameters $[2^r-1, r, 2^{r-1}]$.
- As external code, an IPP Reed-Solomon code defined over $F_2{}^r$ with parameters $(n, \lceil n/4\rceil, n-\lceil n/4\rceil +1)$.
- A function ϕ: $F_2{}^r \rightarrow S_r$

A codeword from C is built by concatenating the y_i's obtained from applying to every symbol of a codeword $x= (x_1, ... ,x_n)$, belonging to the Reed-Solomon code, the function $\mathbf{y}_i=\phi(x_i)$, $1\leq i \leq n$. Therefore, $\mathbf{y} \in C$, is defined as,

$$\mathbf{y}=(\mathbf{y}_i ,..., \mathbf{y}_n) \,/\, \mathbf{y}_i=\phi(x_i)$$

Codes obtained through the previous procedure are similar to those of [1], but the particular choice of the internal and external codes allows more efficient decoding methods to be used.

Identification of a colluder in a fingerprinting scheme that uses the C code consists of decoding a concatenated code where both the internal and the external code should be decoded beyond their correcting capacity.

For the internal code, if $v \in S^*_r$, there are three possibilities for the sets of parent couples.

- All parent couples have a common element, **x**, such that $dist(\mathbf{x}, \mathbf{v}) \leq 2^{r-2}-1$
- There is a unique parent couple {**x**, **y**} such that $dist(\mathbf{x}, \mathbf{v}) = dist(\mathbf{y}, \mathbf{v}) = 2^{r-2}$
- There are three possible parent couples {**x**, **y**}, {**x**, **z**} and {**y**, **z**}, such that $dist(\mathbf{x},\mathbf{v}) = dist(\mathbf{y}, \mathbf{v}) = dist(\mathbf{z}, \mathbf{v}) = = 2^{r-2}$

Therefore, an algorithm is needed which finds all codewords of S_r at a distance 2^{r-2} of **v**. For that purpose, we can make a small modification to the algorithms of Chase [5] as shown in [10]. This results in the Simplified Chase algorithm (SC), whose output is one, two or three codewords.

The number of "errors" that can be introduced by colluders requires the decoding algorithm of the external code to be capable of correcting about n/2 "errors". The algorithm used is KV, because the reliability matrix allows the external decoding procedure to take maximum advantage of the information coming from internal decoding.

Thus, decoding the concatenated code consists of three steps

1. Decoding the internal code.
2. Building the reliability matrix from the result of the previous step and decoding of the external code.
3. Identifying positive_parents from the output of the second step.

A complete description of the algorithm can be found in [10]. It is also proven that the probability of success of an attack against this code decreases exponentially with the code length.

6 Short Binary Fingerprinting Codes Secure Against Collusions of Size 3

In [7], it is proven that the error correcting capacity of dual binary Hamming codes can provide security against collusions of size 2. In this way, 2-secure fingerprinting codes are obtained with shorter codewords than those 2-secure codes obtained through the general construction by Boneh and Shaw [4]. The advantage of shorter codewords is that they introduce less distortion when embedded in the digital content to be protected.

In [14], a construction to obtain codes secure against collusions of size 3 is presented. Codewords of the resulting 3-secure code are shorter than those of 3-secure codes obtained with the general construction by Boneh and Shaw. The basic idea is to compose a new kind of code, called scattering code [15], with a dual binary Hamming code.

7 Conclusions

Collusion attacks are the main threat to fingerprinting systems. Therefore, codes that allow identification of dishonest users should be used. In this paper, the close connection between channel coding and codes with tracing capabilities has been shown. A taxonomy of different schemes as a function of their properties has been presented.

Additionally, an overview of several contributions made by the authors which aim at identifying fraudulent users by using a broad range of channel coding techniques (code concatenation, list decoding, soft decision, scattering codes, etc.) has been given.

Acknowledgements

The first and second authors are partially supported by the Spanish Ministry of Science and Technology through project no. TIC2000-1120-C03-03 "ACIMUT". The third and fourth authors are partially supported by the European Commission under project IST-2001-32012 "Co-Orthogonal Codes" and by the Spanish Ministry of Science and Technology and the European FEDER fund through project no. TIC2001-0633-C03-01 "STREAMOBILE".

References

[1] A. Barg, G. R. Blakley, G. Kabatiansky. "Digital fingerprinting codes: problems statements, constructions, identification of traitors". Technical report, DIMACS 2001-52, 2001.

[2] E. Berlekamp. "Bounded distance +1 soft-decision Reed-Solomon Decoding", IEEE Trans. Inform. Theory, 42(3), 704–720, 1996.

[3] <http://www.lemuria.org/DeCSS>

[4] D. Boneh, J. Shaw. "Collusion-secure fingerprinting for digital data", Crypto'95, LNCS 963, 452–465, 1995.

[5] D. Chase. "A class of algorithms for decoding block codes with channel measurement information". IEEE Trans. Inform. Theory, 18:170–182, 1972.

[6] B. Chor, A. Fiat, M. Naor. "Tracing traitors", Crypto'94, LNCS 839, 480–491, 1994.

[7] J. Domingo-Ferrer, J. Herrera-Joancomartí. "Short collusion-secure fingerprinting based on dual binary Hamming codes", Electronics Letters, vol. 36, no. 20, pp. 1697–1699, 2000.

[8] M. Fernandez, M. Soriano. "Algorithm to decode Identifiable Parent Property codes" Electronics Letters. Vol. 6, number 12. pp 552–553, 2002.

[9] M. Fernandez, M. Soriano. "Soft-decision decoding of traceability codes", IEEE International Conference on Multimedia and Expo ICME 2002.

[10] M. Fernández, M. Soriano. "Fingerprinting concatenated codes with efficient decoding", Information Security Conference 2002, LNCS 2433, 2002.

[11] V. Guruswami, M. Sudan. "Improved decoding of Reed-Solomon and algebraic-geometry codes", IEEE Trans. Inform. Theory, 45(6), 1757–1767, 1999.

[12] H. D. L. Hollmann, J. H. van Lint, J.-P. Linnartz, L. M. G. M. Tolhuizen. "On codes with the Identifiable Parent Property", J. Combinatorial Theory, 82(2), 121–133, 1998.

[13] R. Koetter and A. Vardy. "Algebraic soft-decision decoding of Reed-Solomon codes", ISIT'00, 2000.

[14] F. Sebé, J. Domingo-Ferrer. "Short 3-secure fingerprinting codes for copyright protection", ACISP 2002, LNCS, 2384, pp. 316–327, 2002.

[15] F. Sebé, J. Domingo-Ferrer. "Scattering codes to implement short 3-secure fingerprinting for copyright protection", Electronics Letters. Vol. 38, number 17, pp 958–959, 2002.

[16] A. Silverberg, J. Staddon, J. Walker. "Efficient traitor tracing algorithms using list decoding", LNCS, 2248 (2001), 175 ff, 2001.

[17] J. N. Staddon, D. R. Stinson and R. Wei, "Combinatorial properties of frameproof and traceability codes", CACR Technical Report CORR 2000–16, 2000.

[18] J.H. van Lint. Introduction to Coding Theory, Springer-Verlag, 2nd. Edition, 1992.

[19] F. Sebé, J. Domingo-Ferrer, Jordi Herrera-Joancomartí. "Spatial-Domain Image Watermarking Robust Against Compression, Filtering, Cropping and Scaling" Proceedings of the Information Security - ISW'00 (2000), LNCS 1975 pp. 44–53, 2000.

[20] N. Wagner. "Fingerprinting", Proceedings of the 1983 IEEE Symposium on Security and Privacy, 18–22, 1983.

