Random codes for Digital Fingerprinting

Jacob Löfvenberg
Niclas Wiberg

Report
Reg nr: LiTH-ISY-R-2059
ISSN 1400-3902

Avdelning, Institution
Division, department

Department of Electrical Engineering

Datum
Date

1998-09-22

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ _____

Rapporttyp
Report: category

☐ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☒ Övrig rapport
☐ _____

URL för elektronisk version

ISBN

ISRN

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

LiTH-ISY-R- 2059

Titel
Title

Random Codes for Digital Fingerprinting

Författare
Author

Jacob Löfvenberg, Niclas Wiberg

Sammanfattning
Abstract

Random codes for individual fingerprinting of digital documents are considered in the context of colluding pirates who want to create an untraceable copy. The performance of random fingerprinting in conjunction with a specific testing method is analysed.

Nyckelord
Keywords

Fingerprinting, watermarking, copyright protection, copyright enforcement

95-11-01/lli

Random Codes for Digital Fingerprinting[1]

Jacob Löfvenberg
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping, Sweden
jacob@isy.liu.se

Niclas Wiberg
Department of Electrical Engineering
Linköping University
SE-581 83 Linköping, Sweden
nicwi@isy.liu.se

Abstract

Random codes for individual fingerprinting of digital documents are considered in the context of colluding pirates who want to create an untraceable copy. The performance of random fingerprinting in conjunction with a specific testing method is analyzed.

Keywords: Fingerprinting, watermarking, copyright protection

1 Introduction

Fingerprinting of digital data is a possible solution to the problem of copyright violations. In this section, we describe what we mean by fingerprinting and attacks against it. We also define and describe some of the terminology used (by us and others) when discussing fingerprinting.

With fingerprinting we mean the act of embedding a unique identifier in a digital document of some kind, in such a way that it will be difficult for others to find, remove or destroy the identifier. The purpose of fingerprinting is to make a number of otherwise identical copies unique, so that if an illegally made copy is found, it is possible to trace the person or persons who made it. In order to be able to do this, the identifier used must be chosen carefully; we must use a code. For this to be of practical interest, the embedding technique must be such that the document is not degraded so much that the fingerprinted copies of the document are unusable.

We now outline the type of fingerprinting that we consider, which was also described in [1] and which we have mainly adopted from [3]. For a more general discussion on fingerprinting, see [2].

1.1 A Discrete Fingerprinting Model

We establish the following concepts:

- In the context of fingerprinting, a document is not a fixed piece of data, but rather a set of similar data pieces that *look the same to the users*. In other words, fingerprinting only works if some kind of alteration or distortion is allowed.
- These alterations can be managed by some *embedding technique*, such as choosing between synonyms in certain locations. Embedding techniques typically provide a collection of *variant locations* in the document, each of which provides a choice between a few *variants* that are equivalent to the user.
- Some *coding technique* is required in conjunction with the variants provided by the embedding technique. The coding should be resistant against active attacks by pirates.

1. A presentation of this material, under the same title, was made by the author at ISIT'98, MIT, Cambridge, Massachusetts, USA, in August 1998.

In the following, we will not go into the details on embedding techniques, but simply assume that each document copy is associated with a binary string $\bar{w} \in \{0, 1\}^n$ of length n, the *fingerprint pattern* (or just *fingerprint*) of the copy, where each component w_i corresponds to a variant location in the document. In each location the two variants should be equivalent and the nature of the variants and their locations should not be detectable by a user.

The users will be numbered $1, \ldots, M$, where M is the number of users. The binary string used to fingerprint the copy of user i, $i \in \{1, \ldots, M\}$, is called the *fingerprint pattern* of user i and is denoted by $\bar{w}^{(i)}$. The set of all such fingerprints forms a binary code $\Gamma = \{\bar{w}^{(1)}, \bar{w}^{(2)}, \ldots, \bar{w}^{(M)}\}$ of length n and size M.

A *pirate* is a user who illegally redistributes a copy, modified or unmodified. If several users with different copies of the document work together to create and redistribute an illegal copy, using some combination of their copies, they are called a *collusion* of pirates. If an illegal copy is discovered, the goal is to trace the pirate or pirates. This can be done by first recovering the fingerprint of the illegal copy and then comparing that fingerprint with a database of all users fingerprints. The recovered fingerprint, which we will denote by $\bar{z}$, may not necessarily correspond to any $\bar{w}^{(i)}$, unless a single pirate has distributed his unmodified copy.

Ideally, given a recovered fingerprint $\bar{z}$, we would like to identify *all* the involved pirates. In general this may be a too strong requirement, since a collusion could construct an illegal copy in such a way that only a subset of the pirates contributes significantly to the new copy, while the influence of the remaining pirates is small. We will consider a tracing to be successful as long as a non-empty subset of the colluding pirates is identified. Note also that there are two types of failures: the tracing fails if none of the pirates is identified, but also if an innocent user is inadvertently pointed out as a pirate.

1.2 The Code

How to choose the fingerprints, or codewords, to embed in the copies is an important question, but it is not addressed in this paper. We have chosen to consider binary random fingerprints with each bit drawn independently and with equal probability of zero and one. The main reason for this is that it is mathematically relatively easy to deal with.

1.3 The Marking Assumption

From [3] we adopt the following assumption:

> *The Marking Assumption:* Colluding pirates can detect a specific variant location if, and only if, the variant differs between their copies. They cannot change an undetected variant without destroying the fingerprinted object.

1.4 Collusion Attacks

If several pirates combine their copies, they can compare them location by location and find at least some of the variant locations. In these locations they can choose at will between the two alternatives and create a new copy by *mixing* the original copies. If this happens in sufficiently many locations, the pirates may be able to remove any trace of their identities from the new document copy.

Formally, for a collusion of c pirates, with fingerprints $W = \{\bar{w}^{(i_1)}, \bar{w}^{(i_2)}, \ldots, \bar{w}^{(i_c)}\}$, a location j is said to be *undetectable* if $w_j^{(i_1)} = w_j^{(i_2)} = \ldots = w_j^{(i_c)}$, that is, if all the pirates' fingerprints are equal in that location. Only in these variant locations are the pirates forced in their choice of which bit value to put in their illegally created copy, since they do not know the embedded appearance of the other bit value.

2 A Testing Method

The testing method we propose takes as input a group of users to test, and an illegal fingerprint which has been found somewhere. The method tests whether or not this group of users should be accused of being guilty of having created the illegal copy. The output from the method is thus a flag saying “guilty” or “not guilty”.

When using an illegal fingerprint to test if a group of pirates might be guilty, we do a hypothesis test. Doing this we can make two kinds of errors: Failing to accuse a group that is guilty, and accusing a group that is not guilty. We want the probability of both of these errors to be small, and this leads to a trade-off between conflicting goals. Of course, both of these probabilities are dependent on the fingerprint length and on the number of pirates.

2.1 Description of the Testing Method

When examining a group of possible pirates, we first count the number of locations such that $w_i^{(j)} = w_i^{(k)}$ for all j and k in the group (that is, the undetectable locations), and denote this number N. N is a binomially distributed random variable, $N \in \mathrm{Bi}(n, 2^{1-c})$, where c is the number of users in the group. The number of locations in which $z_i = w_i^{(j)}$ for all j in the group, is denoted r. If the group is innocent, r will be an outcome of a random variable taken from a binomial distribution, $r \in \mathrm{Bi}(N, 2^{-1})$. The probability that r will be close to N will be small, and it will tend to 0 as N grows. Both of these binomially distributed, random variables have their origin in the fact that all users’ fingerprints are random (see subsection 1.2).

We will accuse the group of being guilty if $r = N$ and $N > N_t$, where N_t is a threshold. The reasons for these conditions are, firstly, that if r is not equal to N, then by the marking assumption the group cannot be the guilty one, and secondly, that the larger N is, the smaller the probability that $r = N$ if the pirates are innocent. The probability of falsely accusing a given group of non-pirates is thus $\Pr(N > N_t, r = N)$.

In the method described, we only take into account the undetectable locations. The reason for this is that under the marking assumption the pirates have no freedom of choice in these locations. The output has to be equal to the bits in their fingerprints, and hence, the output in these locations is totally deterministic (for a given group of pirates). In all other locations the pirates can choose which value to output, and they can thereby possibly remove information used in the tracing.

2.2 Analysis of the Tracing Method

When performing the testing we can make two kinds of errors, corresponding to the first and second kinds of errors in hypothesis testing. In this case this means accusing a group that is not guilty and failing to accuse a group that is guilty. The probabilities of making these two kinds of errors are of course interesting and we will find expressions for them in the following.

The method will falsely accuse an innocent group C if $r = N$ and $N > N_t$. The probability of this happening when the group is innocent is

$$\sum_{i=N_t+1}^{n} \Pr(N = i)\Pr(r = N \mid N = i) = \sum_{i=N_t+1}^{n} \Pr(N = i)2^{-i}$$

$$= \sum_{i=N_t+1}^{n} 2^{-i}\binom{n}{i}2^{i(1-c)}(1-2^{1-c})^{n-i}$$

$$= \sum_{i = N_t + 1}^{n} \binom{n}{i} 2^{-ic} (1 - 2^{1-c})^{n-i} \tag{2:1}$$

The method will fail to accuse a guilty group C if $N \leq N_t$, that is, if the number of undetectable locations for C is too small. The probability of this happening is

$$\sum_{i = 0}^{N_t} \Pr(N = i) = \sum_{i = 0}^{N_t} \binom{n}{i} 2^{i(1-c)} (1 - 2^{1-c})^{n-i} \tag{2:2}$$

The expressions (2:1) and (2:2) are easy to compute numerically. In figure 1 below we see the possible trade-offs for different numbers of pirates and different fingerprint lengths. By choosing the value of N_t it is possible to choose the combination of error probabilities from the points on the curve corresponding to the value of c and n for the group under consideration. (Since N_t is an integer, $0 \leq N_t < n$, the number of possible choices on each curve is finite, but the points are quite dense.) The right end-points of the curves correspond to the lowest possible probabilities of failing to accuse a guilty group, for that specific combination of c and n. No other points exist below these.

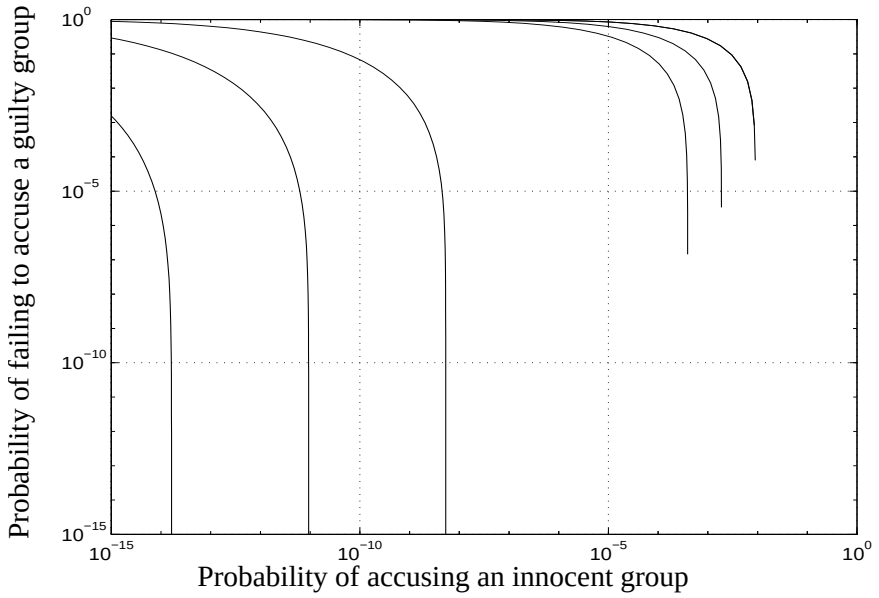

Figure 1. *Possible trade-offs between probability of accusing an innocent group and failing to a accuse a guilty group. The curves are, from the left: first three with number of pirates c = 5 and fingerprint length n = 1000, 800 and 600, and then three with c = 7 and n = 1000, 800 and 600.*

If we instead, by choosing appropriate fingerprint lengths, try to fix the error probabilities we can see how the fingerprint length needed for a certain error probability, varies with the number of pirates. In figure 2 we have chosen the fingerprint lengths according to figure 3, and as can be seen, the error probabilities varies very little. This is true also outside the plotted range.

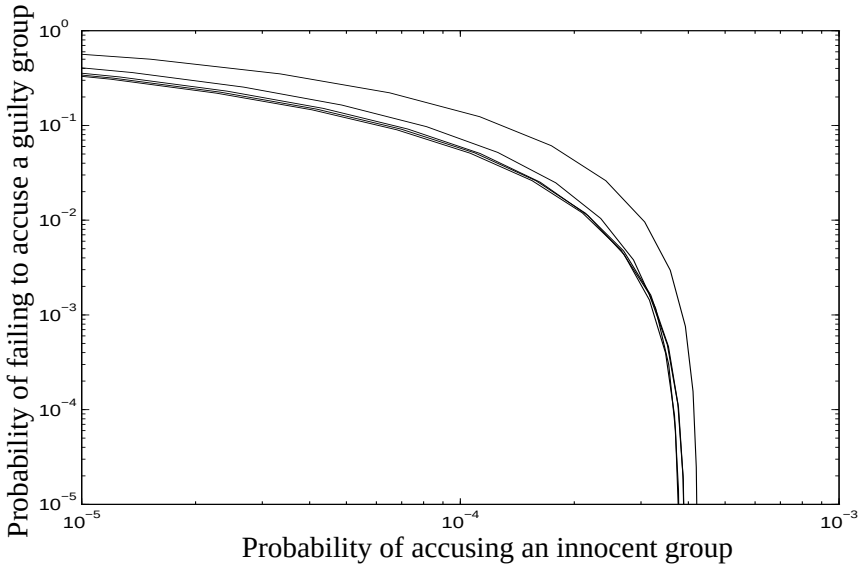

Figure 2. *Possible trade-offs between the different kinds of errors when the fingerprint lengths have been chosen such that the error probabilities should be as similar as possible. The number of pirates are 2, 3, ..., 7.*

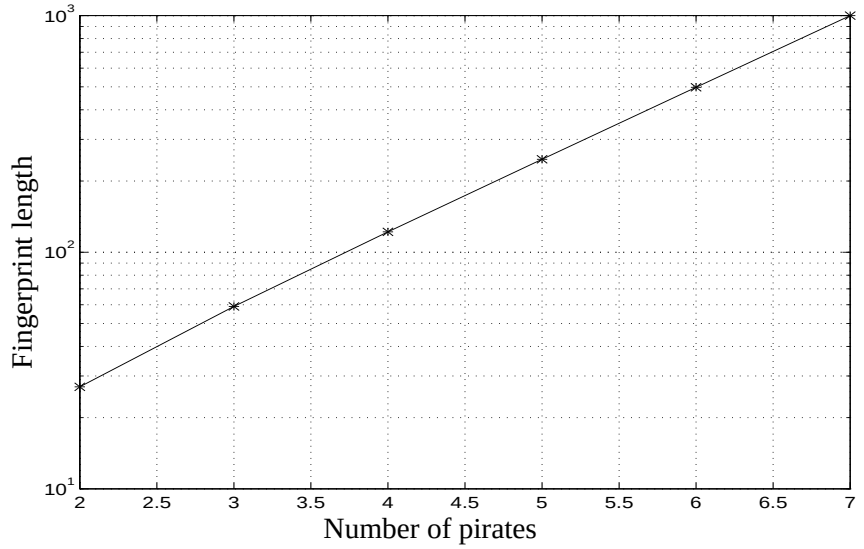

Figure 3. *The fingerprint lengths for the error probabilities of the pirates in figure 2.*

3 Discussion and Conclusions

In figure 1 we see that the number of pirates have a strong influence on the performance of the testing method. If we instead fix the error probabilities (figure 2), we can see in figure 3 that the fingerprint length needed grows exponentially with the number of pirates, which of course is not satisfactory. On the other hand, if we extrapolate figure 3, we see that we can deal with up to 10 pirates with approximately 8000 variant locations and a reasonably low error probability. It is

important to remember though, that this is a *testing* method, not a tracing method. It can only be used for testing whether or not a certain, proposed group of users should be considered guilty of having created a certain illegal copy.

Also, this paper does not consider the probability of accusing a group consisting in part of pirates and in part of innocent users. This probability ought to be greater than the probability of accusing a group consisting solely of innocent users.

References

[1] T. Lindkvist, J. Löfvenberg and N. Wiberg, *Fingerprinting of Digital Information—Introduction and some Preliminary Results*, Report LiTH-ISY-R-1985, ISSN 1400-3902. Available at: http://www.it.isy.liu.se/research

[2] N. Wagner, "Fingerprinting", *Proceedings of the 1983 Symposium on Security and Privacy*, pp. 18–22, April 1983.

[3] D. Boneh, J. Shaw, "Collusion-secure fingerprinting for digital data", *Advances in Cryptology: Proceedings of Crypto '95*, Springer-Verlag, pp. 452–465, 1995.