DNA-Inspired Anonymous Fingerprinting for Efficient Peer-To-Peer Content Distribution

David Megías
Universitat Oberta de Catalunya,
Internet Interdisciplinary Institute (IN3),
Estudis d'Informàtica, Multimèdia i Telecomunicació,
Rambla del Poblenou, 156,
E-08018 Barcelona, Catalonia, Spain
Email: dmegias@uoc.edu

Josep Domingo-Ferrer
Universitat Rovira i Virgili,
UNESCO Chair in Data Privacy,
Department of Computer Engineering and Mathematics,
Av. Països Catalans 26,
E-43007 Tarragona, Catalonia
Email: josep.domingo@urv.cat

***Abstract*—When selling electronic content, the merchant would like each buyer to receive a different copy of the content fingerprinted with a serial number, in order to be able to trace redistributors should illegal redistribution happen. On the other hand, the merchant would like content distribution to be as scalable as possible, in order for mass transactions to be possible. Multicast content distribution fails to satisfy the first requirement: all receivers get exactly the same copy of the content, which makes it difficult to trace illegal redistributors. Unicast distribution of fingerprinted content, on the other hand, fails to satisfy the second requirement: for each buyer, the merchant needs to compute a fingerprint and establish a connection. P2P content distribution is a third option combining the strengths of multicast and unicast: the merchant needs to establish unicast connections only with a few seed buyers; on the other hand, with a suitable fingerprinting mechanism, illegal redistributors can still be identified and honest buyers can stay anonymous. We present a P2P content distribution scheme with such an anonymous fingerprinting mechanism, which is inspired in the way DNA sequences combine and spread from ancestors to descendants.**

***Index Terms*—P2P content distribution; anonymous fingerprinting; redistribution tracing; DNA-based fingerprints; bio-inspired computing.**

I. Introduction

Fingerprinting digital contents [2] is an attractive option to protect the rights of content authors and owners when contents are sold or otherwise distributed over the Internet. Basically, fingerprinting consists of embedding an imperceptible mark in the distributed content (which may be audio, pictures or video) to identify the content buyer. The embedded mark is different for each buyer, but the content should stay *perceptually* identical for all buyers. In case of illegal redistribution, the embedded mark will allow identifying the redistributor (against whom subsequent action might be taken).

Most fingerprinting schemes can be classified in three types [3]: symmetric, asymmetric and anonymous. In the first type, the merchant is the one who embeds the mark into the content; hence, the buyer cannot be formally accused of illegal redistribution, since the merchant also had access to the fingerprinted content and could be himself the redistributor. In asymmetric fingerprinting, the merchant does not have access to the fingerprinted copy, but he can recover the mark in case of illegal redistribution and thereby identify the malicious buyer. In anonymous fingerprinting, in addition to asymmetry, the buyer preserves her anonymity and hence she cannot be linked to the purchase of a specific content, except if she participates in an illegal redistribution. Fingerprinting schemes in the literature share the common feature of being centralized: in a way or another, the content owner/merchant has to be involved in the fingerprinting every time the content is sold to a certain buyer. Hence, distribution is basically *unicast*, which has the shortcoming that scalability is limited by the computing resources and bandwidth available at the content owner/merchant. This problem is further aggravated if one uses asymmetric or anonymous fingerprinting, which require more computation, communication and storage than the usual symmetric fingerprinting. *Multicast* transmission of content, in which a content is simultaneously transmitted to a group of receivers, is much more effective and bottleneck-free. Yet, multicast transmission does not allow sending different copies to each user, as required by fingerprinting schemes. So the question is: can we distribute fingerprinted content in a way that is more scalable than unicast transmission?

P2P distribution systems allow answering the above question in the positive. In these systems, content receivers become senders to other users. This model can be viewed as an intermediate option between unicast and multicast. P2P distribution of all kinds of contents has become popular in recent years with the bandwidth increase of home communications. BitTorrent [5], Kademlia [11] or eDonkey2000 [10] are some example P2P protocols for private file exchange. It must be noted that P2P distribution is not limited to private users in home environments; some content providers are starting to facilitate P2P download of their products (*e.g.* [17]). Using a P2P system allows the merchant to establish only a small number of unicast connections with a set of M "seed" buyers who become new sources of content for other buyers. The content can eventually reach a set of N buyers with $M \ll N$.

A. Contribution and plan of this paper

We propose a P2P content distribution scheme (based on a specific P2P software) in which the merchant creates only

a set of M seed copies of the content and sends them to M seed buyers. All subsequent copies are generated from the M seed copies. The copy obtained by a buyer is a combination of the copies supplied by her "parents" (sources). The fingerprint of each buyer is constructed as a binary sequence combining the sequences of her parents, in a way parallel to how DNA sequences of living beings are formed by combining the DNA sequences of their parents. The proposed scheme, which saves bandwidth and computation at the merchant, still allows tracking illegal redistributors but preserves the anonymity of honest buyers. The proposed method is thus inherently scalable compared to other systems in the literature [15], [3], [1], [7], which require (non-scalable) unicast transmissions and rely on complex CPU-intensive and/or bandwidth-consuming cryptographic protocols. The cryptographic protocols used in our approach reduce to the transmission of a few encrypted hashes with low computation and communication costs. In fact, the method proposed in this paper even avoids running the embedding algorithm for non-seed buyers and thus it outperforms the abovementioned methods.

The rest of this paper is organized as follows. Section II introduces DNA-inspired fingerprints, which are the foundation of the proposed scheme. Section III describes the method for P2P distribution of DNA-inspired fingerprinted contents, including the algorithm for tracing illegal redistributors. A security and privacy analysis is given in Section IV. Simulation results are reported in Section V. Section VI is a conclusion.

II. DNA-inspired fingerprints

In this section, we introduce a novel concept of automatic DNA-inspired binary fingerprints. The terms used in this paper are derived from those used in genetics to refer to DNA and heredity. The definitions of these terms in the context of this paper are introduced below.

DNA sequence: in nature, DNA is a molecule consisting of an ordered set of nucleotides, where each nucleotide is one of the following four (smaller) molecules: adenine, cytosine, guanine and thymine, usually represented by their initial letters ("A", "C", "G" and "T"). Although the DNA molecule consists of two strings, the nucleotides are always paired A-T and C-G in the two different strings, meaning that the DNA structure is redundant. The DNA molecule is a double helix string in which each string is the complementary of the other one. Hence, one of the strings contains all the information required to build the other one.

In this paper, a DNA-inspired fingerprint is constructed as a binary sequence and each bit can be considered as the counterpart of the nucleotides in real DNA sequences. Although each of the real DNA sequences' nucleotides can be thought of as a two-bit symbol (since there are four different nucleotides), the analogy can still be established using 1-bit nucleotides. This is similar to what is done in genetic algorithms [9].

Gene: a segment of the DNA sequence which encodes a given protein –and thus has some impact in heredity and in the biological chemistry of the living being– is called a gene. Similarly, a segment of the DNA-inspired fingerprint sequence is called a "gene" in this paper. Although real life genes have different sizes, the sizes of the genes in this paper are taken to be equal for simplicity and without loss of generality (variable sized genes might be used with no additional complexity in the proposed system). In addition, in nature, not all segments of the DNA sequence encode genes. Nevertheless, in this paper, all "nucleotides" (bits) do belong to one of the genes of the DNA-inspired fingerprint.

Mating and heredity: in nature, the genes of an offspring are basically a combination of the genes of its parents (although some other processes such as mutation and crossover may produce fragments of DNA which are different in the offspring with respect to both its parents).

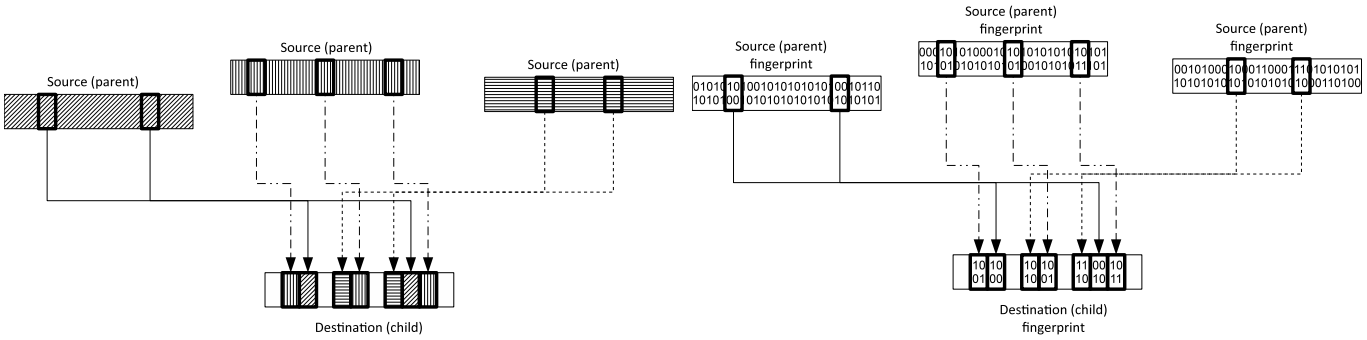
Similarly, in this paper, when a buyer obtains a copy of a P2P-distributed content using some specific software, the DNA-inspired fingerprint of her copy will be a combination of the genes of the sources of the content (referred to as "parents" from the biological analogy). In this case, the number of parents for a buyer does not have to be exactly two as in the natural world. Hence, the mating process in the suggested fingerprinting scenario must be understood in a generalized sense, not limited to two parents.

In this proposal, fingerprints can be considered as being "automatically generated" from the fingerprints of the parents. Despite this "automatic generation" of fingerprints, the constructed sequences are still valid for identification purposes, just like DNA traces can be used in criminal investigations to identify the suspect of an offence.

Mutation and crossover: different types of changes may occur in DNA molecules resulting in the modification of some of the nucleotides in the sequence. These changes may affect a single nucleotide or a full segment of the DNA sequence. Basically, crossover occurs when the two complementary DNA strings are recombined during DNA replication and mutation refers to different random-like errors during DNA replication.

Mutation and crossover provide mechanisms which allow the DNA sequence of an offspring to include genes which are different from those of its parents. If it is allowed that a buyer can obtain her copy of the contents from only one parent (source), mutation (and/or crossover) shall be used to produce a different version of the fingerprint, as required in fingerprinting applications (since two different buyers must have different fingerprints). Note that, although the DNA-inspired fingerprints are defined as a single bitstream, it is still possible to consider that a complementary sequence exists by using its negation. Crossovers can thus be simulated between a binary fingerprint and its negation.

Although mutation and crossover provide a hypothetical mechanism to obtain a different fingerprint for parent and child in the single-parent case, still, the best and easiest strategy for a practical implementation of the scheme is to avoid this solution by enforcing at least two parents for each buyer. The implementation presented in this paper uses neither mutation nor crossover since the system compels each buyer to obtain the content's fragments from at least two buyers.



(a) Mating process (b) Heredity

Fig. 1: Upload/download of the content (mating process) and automatic DNA-inspired fingerprint construction (heredity)

DNA relationship test: in real-world investigations, DNA relationship testing is often conducted to prove or disprove a blood relation between two or more individuals. Taking into account the mating and heredity processes and properties, blood relatives are known to share longer segments of their DNA sequences compared to those of non-relatives.

The equivalent of this DNA relationship test in the proposed fingerprinting scheme is a function to compute the correlation between two binary strings. Since ancestors and descendants in the distributed fingerprinting scenario share several of their genes, a measurable correlation exists between their DNA-inspired fingerprints.

III. P2P DISTRIBUTION OF DNA-INSPIRED FINGERPRINTED CONTENTS

In P2P content distribution, receiving users become sources of the content for others as soon as fragments are received. When a file is downloaded using this kind of P2P applications, fragments of several sources are joined together. In most of these systems contents are indexed using hash values and two files with the same hash value are considered identical.

The upload and download processes for obtaining a file from different sources in a P2P fashion are shown in Fig. 1a. In this figure, the destination (or child) is downloading fragments of the file from three different sources (or parents). When all fragments are downloaded they are joined together to construct a copy of the content.

A. Requirements on fingerprint embedding

To use the DNA-inspired fingerprints described in Section II in a P2P distribution scenario it is required to begin with a (reduced) number M of seed buyers that provide the first few copies of the content to other buyers. These seed M copies of the content may have randomly generated fingerprints such that their pairwise correlation is low.

The main requirements of the embedding scheme are as follows:

1) The DNA-inspired fingerprint must be a binary sequence that is spread along the whole file. The fingerprint must be formed as the concatenation of separated pieces (genes) that are embedded in different fragments of the file. These fragments will be distributed by the P2P software as a "atomic" components of the content. Hence, each of the fragments carries a full gene of the fingerprint. If the P2P software works with fragments of say 16-KB (kilobyte), each gene will be embedded into one of these fragments. In this case, the fingerprint extraction method must be robust against fragmentation in 16-KB pieces if the beginning and the end of the fragments are not altered. This idea is illustrated in Fig. 1b.

   Note that not all embedding systems allow this kind of fragmentation. In some schemes, the embedded fingerprint must be extracted from the whole file and not fragment by fragment. An example of block-based audio watermarking system which may be used for fingerprinting in this scenario is presented in [13].
2) Obviously, the versions downloaded by different buyers will not be bitwise identical. The P2P-distributed download of the contents will produce different copies for different buyers, but all the copies of the contents should be equivalent from a perceptual point of view (all buyers require the same high quality for the purchased content). In this case, a standard hash function which produces different hash values even after a single bit change would not be useful for indexing, since the copies obtained by different buyers would not be associated to the same index. A way to overcome this difficulty is to use perceptual hash functions [6], for which the same hash value is obtained for different versions of the same content if they are perceptually identical.

Provided that the previous two requirements are met, the fingerprinting process is "automatic" as contents are downloaded by buyers from different sources. Embedding is required only for the seed buyers. Afterwards, no additional overhead for embedding is required since the distribution process ensures different fingerprints for different buyers. The only constraint for this process to work properly is that each buyer must choose at least two parents (sources) of the file. Otherwise, in a single-parent distribution, the copies and DNA-inspired

fingerprints for both parent and child would be identical. If the single-parent case does not want to be excluded, mutation and/or crossover may be used to modify the child's fingerprint at least in some genes. In this way, parent and child would have different identifiers, but fingerprinting would no longer be automatic, since the modified genes would have to be embedded at their correct positions. The simplest way to ensure that fingerprints are different for different buyers is to enforce at least two parents for each buyers, and this is the solution adopted in this paper.

B. The P2P distribution protocol

To bootstrap the system, a few copies of the content with different fingerprints must be produced. The merchant can produce a small number M of instances of the content and embed different pseudo-random binary sequences (DNA-inspired fingerprints) into them. These copies are then transferred to the M seed buyers who may be genuine buyers of the content or dummy buyers created by the merchant to facilitate the distribution. In either case, the seed buyers will be contacted by second-generation buyers to download further copies of the content. The association of the first M fingerprints with the pseudonyms of the first M buyers must be recorded either by the merchant or some trusted authority.

Once the system is bootstrapped, all further transactions occur without any need to run the embedding algorithm as far as at least two parents are chosen for each buyer (as discussed above). Note also that all fingerprints from buyer $M+1$ to the final one (N) will remain completely anonymous (only known by the buyers themselves) and cannot be related to real identities. Thus, anonymous fingerprinting is obtained in a much simpler way than with any of the existing proposals in the literature [15], [3], [1], [7], which require running complex cryptographic protocols for *every* transaction. As detailed below, in our proposal only the transaction monitor keeps a record of the transactions between buyers in case they are required in future DNA relationship tests. In any case, real identities are not known by the transaction monitor and, hence, privacy is fully preserved.

The P2P distribution protocol is summarized below.

Protocol 1 (P2P distribution):

1) For $i := 1$ to M, the merchant generates the i-th pseudo-random binary fingerprint and embeds it into the original content to produce the first M fingerprinted files. These M fingerprints must have low pairwise correlations.
2) For $i := 1$ to M, the merchant forwards the i-th seed copy to the i-th seed buyer. If the seed buyers are genuine rather than dummy buyers, this step can be anonymized using cryptographic protocols.
3) For $i := M+1$ to N, the i-th buyer obtains her copy of the content by joining fragments obtained from a set S_i of parent nodes (sources) such that $S_i \subseteq \{B_1, \cdots, B_{i-1}\}$ and $|S_i| > 1$, where $|\cdot|$ is the cardinality operator and B_j refers to the j-th buyer. This transaction is performed via a proxy (or a set of proxies) and with an anonymous protocol such as [4]. The proxy registers each transaction at the transaction monitor. When all the fragments have been transferred for a buyer, the whole fingerprint's hash is also stored in the transaction monitor.

The hash of the fingerprint is stored as a ciphertext in the transaction monitor: it is encrypted using the public key of the transaction monitor and the public key of each parent. There will be as many records for each buyer as parents she has, and the fingerprint's hash is stored encrypted with the parent's public key and the transaction monitor's public key. In this way, in case of a traitor tracing investigation, the transaction monitor will need the cooperation of at least one parent to decrypt the hash. This provides additional anonymity and protection to buyers.

C. The traitor tracing protocol

We now show that the proposed fingerprinting method allows identification of illegal redistributors (traitors) of fingerprinted contents. Assuming that the embedding scheme is secure and robust enough so that malicious users cannot easily erase their fingerprints without making the content unusable (this is the standard marking assumption, [2]), the following method can be used by a tracing authority to identify the source of an illegally distributed copy.

Protocol 2 (Traitor tracing):

1) The fingerprint f of the illegally distributed content X_f is extracted using the extraction method.
2) The initial test set T_0 is built with the M buyers of the seed versions of the file.
3) Let $i := 0$.
4) The tracing authority contacts the buyers in the current set T_i and retrieves the hashes of their fingerprints from the transaction monitor. This step requires the private key of one parent of these buyers (the merchant in case $i = 0$ and the selected ancestor in the set T_{i-1} otherwise) and the private key of the transaction monitor. Then, the fingerprints of the buyers of T_i are extracted from their copies of the content and the hash function h is applied to each segment to obtain the fingerprints' hashes. If any of the buyers' fingerprints produces a hash which does not match the corresponding record in the transaction monitor, the associated buyer is accused of forgery (contract breach).
5) If no forgery has been detected in the previous step, the DNA relationship test is carried out with the fingerprints of the buyers in the current set T_i. This step is conducted as a simple bitstream correlation. Given the fingerprint f to be traced and the test fingerprint f' extracted from the copy $X_{f'}$ corresponding to a buyer in T_i, both fingerprints with length L, the correlation $C(f, f')$ between f and f' is computed as follows:

$$C(f, f') = \frac{1}{L} \sum_{j=1}^{L} (-1)^{f_j \oplus f'_j}, \tag{1}$$

where f_j and f'_j are, respectively, the j-th bits of f and f', and $\oplus$ refers to the exclusive-or operation. In case

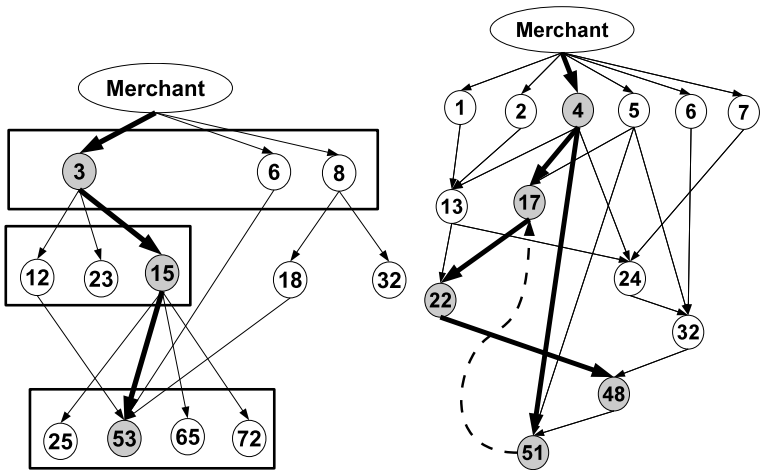
of forgery, this step can be computed with the hashes of the fingerprints instead of the fingerprints themselves. If the correlation of the hashes is one, the corresponding buyer is charged of illegal redistribution and the traitor tracing protocol halts.
A guilty redistributor may try to randomly alter $X_{f'}$, but this would make the content unusable according to the marking assumption [2]. In addition, such an alteration would be detected as a hash mismatch. If a buyer is afraid of being framed by the authority, she might prefer not to reveal $X_{f'}$; in this case, buyer and tracing authority may engage in a secure multiparty computation to compute Expression (1).

6) If no culprit has been identified so far, there may be three outcomes of the previous step:
   a) One or more buyers in T_i refuse to collaborate with the tracing authority in computing their correlations with f. In such a case, depending on the correlation between the hash h_f and the hash(es) of the refusing buyer(s) (recorded in the transaction monitor), the refusing buyer(s) is(are) accused either of redistribution (if hashes are identical) or contract breach (otherwise). If the correlation between hashes is less than 1, this correlation can be used as a replacement for the correlation between the fingerprints.
   b) One buyer in T_i has $C(f, f') = 1$; in this case, this buyer is identified as the culprit.
   c) Otherwise, the buyer in T_i who has the maximum correlation with f is taken as the most likely ancestor of the buyer of the illegally distributed copy and a new set T_{i+1} of buyers is built with all the children of this ancestor buyer, excluding any children buyers who have been already analyzed (remember that each buyer will have several parents). These children can be obtained from the transaction records of the transaction monitor. Once the new set T_{i+1} is available, set $i := i + 1$ and go to Step 4.

The maximum correlation criterion will work with a high probability, but a a higher correlation might accidentally be obtained for a non-ancestor of the buyer of the illegally distributed copy. For example, a descendant D of the illegal redistributor I may have, as another ancestor A, a node of the graph which is also an ancestor of I. This would produce a high correlation but the chain from A to D skips the illegal redistributor I. In this situation, *backtracking* is required in the tracing algorithm described above. A complete subnetwork should be exhausted until all nodes of the subgraph having no children are considered. When a complete subnetwork is exhausted, the element of T_i with the second maximum correlation would be chosen as the candidate ancestor of the traitor to be identified. When all elements of T_i have been considered without success (*i.e.* without being able to accuse anyone), the procedure would backtrack to the set T_{i-1}. Backtracking has been needed in a very small number of the simulations presented in Section V.

We now give some examples of the operation of Algorithm 2. Figure 2a shows an example of a subgraph of the P2P distribution network. This can be modeled as a directed graph from content sources (parents) to content destinations (children). The figure illustrates how Algorithm 2 discovers that the fingerprint in the content (traced fingerprint) is the one of buyer B_{53}. The system begins testing the DNA relationship between this traced fingerprint and a set T formed by all the children of the merchant. If $M = 10$, then $T_0 = \{B_1, B_2, \ldots, B_{10}\}$. In this case, three buyers among those in T_0, namely B_3, B_6 and B_8, have DNA-inspired fingerprints with the top three correlations with the traced fingerprint (no wonder if one knows that B_{53} is the traced buyer, because B_3, B_6 and B_8 are ancestors of B_{53}). It turns out that B_3 is the one with the highest correlation (again, no surprise if one knows that B_{53} is the traced buyer; B_3 and B_6 are the most likely to have fingerprints with the highest correlation with B_{53}, since two of the parents of B_{53} are children of B_3 and B_6 is also a parent of B_{53}). The next iteration is performed with all the children of B_3, namely $T_1 = \{B_{12}, B_{15}, B_{23}\}$. The DNA relationship yields the highest correlation with B_{15} (if one knows that B_{53} is the traced buyer, since B_{53} has four parents including B_{12} and B_{15}, the correlation of the fingerprints of the latter two buyers with the fingerprint of B_{53} must be around 0.25). Now, the set T_2 is formed with buyer B_{15}'s children: $T_2 = \{B_{25}, B_{53}, B_{65}, B_{72}\}$. In this situation, B_{53} will be found to have a fingerprint with correlation 1 with the traced fingerprint unless she refuses to take the DNA relationship test. In any case, she will be accused of illegal redistribution because of the perfect match between her fingerprint's hash and the traced one. In Figure 2a, the nodes highlighted in grey are the ones yielding the highest correlations and rectangles are used to enclose the nodes that are involved in DNA relationship tests.

Figure 2b shows an example of a situation which requires backtracking, where B_{48} is the traitor. The curved dotted arrow in the figure does not represent an edge of the graph, but the backtracking process. In this situation, the set $T_0 = \{B_1, B_2, \ldots, B_{10}\}$ is formed as in the previous example and the maximum correlation is obtained for B_4. Note that this is an ancestor of B_{48} (as expected) but it shares a common child (B_{51}) with the traitor. The new set of candidates is constructed with B_4's children as $T_1 = \{B_{13}, B_{17}, B_{24}, B_{51}\}$. In this case, B_{51} is very likely to produce maximum correlation with the traced fingerprint (the one of B_{48}), because B_{48} is a parent of B_{51} and the other parent (B_4) is an ancestor of B_{48}. Once B_{51} is selected, her children (if any) and all her descendance subgraph would be examined without finding a correlation $C = 1$. Finally, after analyzing all the descendance subgraph, backtracking occurs, going back to the set T_1 and picking the second highest correlation in the set, which is found for B_{17}, who is a true ancestor of the traitor (B_{48}). After that, two more iterations are required to find the illegal redistributor (descendants of B_{17} and descendants of B_{22}).



(a) Typical situation (b) Backtracking case

Fig. 2: Traitor tracing examples for the P2P content distribution scheme (typical situation and backtracking case)

IV. Security and privacy analysis

This section analyzes the security and privacy properties for the proposed system: the security assumptions (including buyer frameproofness) are summarized and a collusion-resistant version of the scheme is discussed.

A. Security assumptions

The security assumptions of the proposed system are the following:

1) The proxies and the transaction monitor do not know real identities, only pseudonyms. Hence, neither the proxies nor the transaction monitor can break the privacy of buyers by themselves. A specific purchase by a buyer can only be leaked by a coalition of the merchant and at least one of the proxies or the transaction monitor. The merchant is the only party having access to the real identities.
2) The only threat to buyer security (resulting in an innocent buyer being framed) is a coalition of all proxies chosen by a buyer. Proxies have access to the cleartext of the content's fragments and they have to share the fragments of the child buyers' fingerprint hash. This means that proxies need to have contact between them during the process. If all the proxies chosen by a buyer collude, they can replicate the content transferred to the buyer by joining the different pieces together. Thus, they can redistribute the content illegally to frame an innocent buyer. This paper assumes that proxies are honest.
3) The traitor tracing protocol requires that at least one of the parents of a buyer provide her secret key to obtain the fingerprint's hash stored at the transaction monitor. If a buyer refuses to co-operate by providing her fingerprint's correlation with the traced fingerprint, at least one of the parents of the buyer must co-operate to decrypt her child's fingerprint hash. If all the parents of a non co-operative child refused to do so, the system would not be able to trace the child if she were the traitor. However, cheating parents would have to pay some punishment (fine) due to contract breach, and they are unlikely to incur such a risk to favor an unknown child.
4) Parents are also expected to provide the fingerprint's hash bit of each fragment. A simple solution to avoid cheating about it consists in having the fingerprint's hash bits and fragments signed at the origin by the merchant. The signature could thus be verified by the proxies (who are assumed honest).
5) As already discussed, the merchant has access to the seed buyers' fingerprints and, hence, if they are genuine buyers, they can be accused if the merchant behaved maliciously and redistributed their copy of the content. To avert such a dishonest behavior, the simplest alternative is for the M seed buyers not to be real buyers, but dummy buyers created by the merchant to bootstrap the P2P distribution protocol. Hence, the first real buyer would be the $M+1$-th one.

B. Collusion resistance

Fingerprinting schemes are not only required to trace authentic fingerprints, but also forged instances created by advanced attackers. When several buyers collude, they create a new copy of the content whose fingerprint is formed by a mixture of the bits of the fingerprints of the colluders (standard marking assumption) and, thus, a new forged fingerprint results. In this section, we show how the existing anti-collusion fingerprinting codes (*e.g.* [2], [8], [16], [14]) can be used also in the proposed DNA-inspired fingerprinting system. This means that the suggested scheme can be made as resistant against collusion as any of the existing anti-collusion techniques of the literature.

The following method is suggested to endow the DNA-inspired fingerprinting system with collusion resistance:

- Each gene shall be encoded with a (gene-level) anti-

TABLE I: Average number and percentage of DNA relationship tests on non-seed buyers in an exponentially growing population

Gener.	Popul.	Average DNA tests		Backtrack.
		1 simul.	100 simul.	(100 sim.)
$k = 2$	$N = 20$	3.40 (34.0%)	3.71 (37.1%)	0%
$k = 3$	$N = 40$	6.93 (23.1%)	7.29 (24.3%)	0%
$k = 4$	$N = 80$	12.26 (17.5%)	11.69 (16.7%)	0.6%
$k = 5$	$N = 160$	18.99 (12.7%)	17.05 (11.4%)	1.2%
$k = 6$	$N = 320$	24.31 (7.8%)	23.76 (7.7%)	2.7%

collusion code which can be used to reconstruct the gene of one of the colluders. Since the merchant embeds the fingerprint of the seed buyers, an honest merchant suffices to guarantee that all the genes are encoded using this specific codebook. Hence, if a set of colluders fabricate a copy of the content and redistribute it, each gene may be decoded to recover the corresponding gene of one of the colluders.

- The fingerprints must be constructed in such a way that their hashes are also codewords of a (hash-level) collusion-resistant code. In this way, after a collusion, when the genes have already been reconstructed by the tracing authority, the hash of at least one of the colluders will be obtained. In this case, cooperation by the proxies is required to construct a valid codeword for each hash. For example, if using an error-correcting code as an anti-collusion code (*e.g.* [8]) and assuming the code is in systematic form, the "data" bits of the hash can be chosen randomly, whereas specific parents having the required hash bit shall be selected for the redundancy bits of the hash. The proxy can contact potential parents subsequently, requiring a specific hash bit for a given gene, and only the ones having the specific hash bit for that gene would be accepted as the source for that specific fragment of the content. In this case, the traitor tracing algorithm will stop when a correlation equal to one is found with respect to the fingerprint's hash instead of the whole fingerprint, since only the hash will be perfectly reconstructed in case of collusion.

V. Simulation results

This section presents a set of simulated experiments to illustrate the properties of the proposed system. In particular, we focus on the number of buyers which will be required to cooperate with the tracing authority in case of a traitor tracing investigation. All simulations presented below use DNA-inspired fingerprints formed by 4096 bits, divided into 128 genes of 32 bits each. A more detailed analysis and empirical results on the method proposed in this paper, including examples which are closer to real-world scenarios, can be found in [12].

The first simulation consists of producing different generations of buyers using an exponential growth approach and checking the average number of required DNA relationship tests. The following assumptions are made:

1) The first generation is formed by $M = 10$ seed buyers.
2) At each generation, the population increases by 100%. This means that, on average, each buyer sends the whole content allowing to feed a new buyer (a new copy of the entire content). Hence, the second generation would be formed by M new buyers. The third generation would be formed by $2M$ buyers, and so on. With this assumption, the population increases exponentially after each generation. For example, after six generations, the population would be $M + M + 2M + 4M + 8M + 16M = 32M$. If k is the number of generations, the total population is $N = 2^{k-1}M$.
3) For each buyer, between two and four parents are chosen at random from the previous generations. Hence, the average number of parents per non-seed nodes is three.

After simulation, the results shown in Table I have been obtained. The results show a single simulation and the average of 100 simulations with 100 different seeds in the pseudo-random number generator in order to reduce the bias of the results. It can be seen that no significant differences appear between 1 and 100 simulations. The last column represents the average percentage of buyers requiring backtracking in the 100 simulations. Not surprisingly, as the network (graph) becomes larger, more buyers will require backtracking, but the percentage is always small. In any case, the fraction of non-seed buyers affected by *one* DNA relationship test decreases to zero as the number of generations grows: the more buyers involved, the higher the probability of remaining anonymous in one DNA relationship test.

It may appear that the decrease in the percentage of buyers involved in DNA relationship tests in the course of an investigation decreases to zero because of the exponential increase in population occurring at each generation. However, this is not the case. The decrease of this ratio of tested buyers depends on the population and not on the particular way it grows. To illustrate this behavior, the following simulations have been performed with a population growing linearly at each generation:

1) The first generation is, again, formed by the $M = 10$ seed buyers who obtain their fingerprinted contents from the merchant.
2) At each new generation, $M = 10$ new buyers obtain their contents from a variable number of parents between two and four (and thus, the average number of parents is, again, three). Again, parents are chosen at random from the previous generations.
3) With this scenario, the population N increases linearly with the number of generations: there are $N = kM$ buyers after the k-th generation.

We present simulation results comparing the linear and exponential growths scenarios *for the same population*. The results are shown in Table II and Figure 3. It can be observed that, when populations are of the same size, the results are almost identical irrespective of the number of generations and the growth model. In these tests, the seeds of the pseudo-

TABLE II: Average number and percentage of DNA relationship tests on non-seed buyers: comparison between exponential and linear growth for the same population

Population	Exponential growth		Linear growth	
	Gener.	Average tests (100 simul.)	Gener.	Average tests (100 simul.)
$N = 20$	$k = 2$	3.71 (37.1%)	$k = 2$	3.71 (37.1%)
$N = 40$	$k = 3$	7.29 (24.3%)	$k = 4$	6.90 (23.0%)
$N = 80$	$k = 4$	11.69 (16.7%)	$k = 8$	10.62 (15.2%)
$N = 160$	$k = 5$	17.05 (11.4%)	$k = 16$	15.43 (10.3%)
$N = 320$	$k = 6$	23.76 (7.7%)	$k = 32$	22.23 (7.2%)

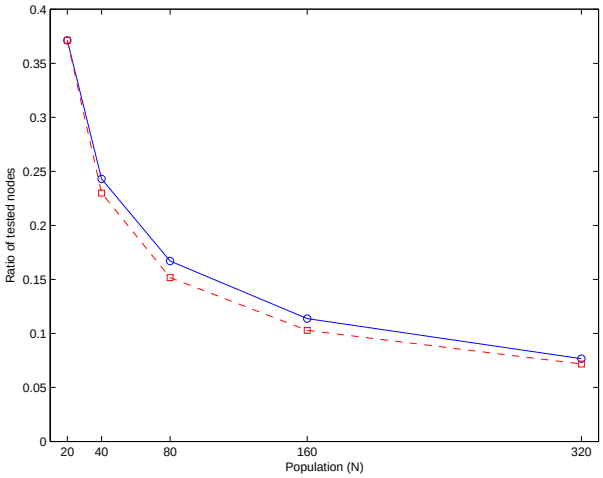


Fig. 3: Average fraction of non-seed buyers affected by DNA relationship tests: comparison between exponential growth (circle, solid) vs. linear growth (square, dashed) for the same population. Abscissae is population size

random number generator have been adjusted such that the results for two generations ($N = 20$) are the same with both growth models.

VI. Conclusion

A DNA-inspired fingerprinting scheme designed for P2P content distribution is presented. The proposed scheme allows the merchant to trace traitors who redistribute the content illegally. The merchant knows at most the fingerprinted copies of the seed buyers, but not the fingerprinted copies of non-seed buyers (the vast majority). Hence, the merchant does not know the identities of non-seed buyers. Whenever a traitor needs to be traced, only a small fraction of honest users must co-operate by providing their fingerprinted copies (quasi-privacy). Collusion resistance against dishonest buyers trying to create a forged copy without any of their fingerprints is also discussed. Finally, buyer frameproofness is guaranteed since a malicious merchant does not have access to the fingerprinted copies of non-seed nodes. Thus, he will not be able to frame an honest buyer since random guess is not an option to construct a valid fingerprint (due to combinatorial explosion).

Future research will involve designing backtrack-free protocols for traitor tracing in such a way that the fraction of honest buyers who must co-operate in case of an illegal redistribution is reduced. The security analysis of the proposed scheme against malicious proxies, who may even collude with other parties is also left for the future research.

Acknowledgment

This work was partly funded by the European Commission under FP7 projects "DwB" and "Inter-Trust", by the Spanish Government through projects TSI200765406-C03-01/03 "E-AEGIS", TIN2011-27076-C03-01/02 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-0004 "ARES", and by the Government of Catalonia through grant 2009 SGR 1135. The second author is partly supported as an ICREA-Acadèmia researcher by the Government of Catalonia; also, he holds the UNESCO Chair in Data Privacy, but the views expressed in this paper are his own and do not commit UNESCO.

References

[1] Y. Bo, L. Piyuan, and Z. Wenzheng. An efficient anonymous fingerprinting protocol. In *Computational Intelligence and Security*, LNCS 4456, Springer, pp. 824-832, 2007.

[2] D. Boneh and J. Shaw. Collusion-secure fingerprinting for digital data. In *Advances in Cryptology-CRYPTO'95*, LNCS 963, Springer, pp. 452-465, 1995.

[3] J. Camenisch. Efficient anonymous fingerprinting with group signatures. In *Asiacrypt 2000*, LNCS 1976, Springer, pp. 415-428, 2000.

[4] D. L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84-90, 1981.

[5] B. Cohen. The BitTorrent Protocol Specification. 2008. Available at http://www.bittorrent.org/beps/bep_0003.html

[6] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker. *Digital Watermarking and Steganography*. Burlington MA: Morgan Kaufmann, 2008.

[7] J. Domingo-Ferrer. Anonymous fingerprinting based on committed oblivious transfer. In *Public Key Cryptography-PKC 1999*, LNCS 1560, Springer, pp. 43-52, 1999.

[8] J. Domingo-Ferrer and J. Herrera-Joancomartí, Short collusion-secure fingerprints based on dual binary Hamming codes. *Electronics Letters*, 36(20):1697-1699, 2000.

[9] D. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston: Addison-Wesley, 1989.

[10] O. Heckmann and A. Bock. The eDonkey 2000 Protocol. KOM Technical Report 08/2002, Ver. 0.8. Department of Electrical Engineering & Information Technology & Department of Computer Science. Darmstadt University of Technology, Germany. 2002.

[11] P. Maymounkov and D. Mazières. Kademlia: a peer-to-peer information system based on the XOR metric. In *IPTPS 2002-First International Workshop on Peer-to-Peer Systems*, LNCS 2429, Springer, pp. 43-65, 2002.

[12] D. Megías and J. Domingo-Ferrer. Privacy-aware peer-to-peer content distribution using automatically recombined fingerprints. *Multimedia Systems*, in press.

[13] D. Megías, J. Serra-Ruiz, and M. Fallahpour. Efficient self-synchronised blind audio watermarking system based on time domain and FFT amplitude modification. *Signal Processing*, 90(12):3078-3092, 2010.

[14] K. Nuida, S. Fujitsu, M. Hagiwara, T. Kitagawa, H. Watanabe, K. Ogawa, and H. Imai. An improvement of Tardos's collusion-secure fingerprinting codes with very short lengths. In *Proceedings of the 17th international conference on Applied algebra, algebraic algorithms and error-correcting codes (AAECC'07)*, Springer, pp. 80-89, 2007.

[15] B. Pfitzmann and M. Waidner. Anonymous fingerprinting. In *Advances in Cryptology-EUROCRYPT'96*, LNCS 1233, Springer, pp. 88-102, 1997.

[16] G. Tardos. Optimal probabilistic fingerprint codes. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing (STOC '03)*, ACM, pp. 116-125, 2003.

[17] H. Xie, Y.R. Yang, A. Krishnamurthy, Y. G. Liu and A. Silberschatz. P4P: provider portal for applications. *SIGCOMM Comput. Commun. Rev.*, 38(4):351-362, 2008.