Information-Theoretical Analysis of Private Content Identification

S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, T. Holotyak
Stochastic Information Processing (SIP) Group
University of Geneva
7, route de Drize, 1227 Geneva, Switzerland
Email: {svolos, oleksiy.koval, fokko.beekhof, farzad.farhadzadeh, taras.holotyak}@unige.ch

***Abstract*—In recent years, content identification based on digital fingerprinting attracts a lot of attention in different emerging applications. At the same time, the theoretical analysis of digital fingerprinting systems for finite length case remains an open issue. Additionally, privacy leaks caused by fingerprint storage, distribution and sharing in a public domain via third party outsourced services cause certain concerns in the cryptographic community. In this paper, we perform an information-theoretic analysis of finite length digital fingerprinting systems in a private content identification setup and reveal certain connections between fingerprint based content identification and Forney's erasure/list decoding [1]. Along this analysis, we also consider complexity issues of fast content identification in large databases on remote untrusted servers.**

I. Introduction

In the last 10 years, content identification based on digital fingerprinting performed an impressive evolution from just an alternative solution to digital watermarking in copyright protection to a stand-alone domain of research. In many applications, where content modifications caused by watermark embedding are *undesirable*, *hardly possible* without severe and unpredictable consequences (human biometrics including DNA) or *conflicting with the assumed protocol* such as physical uniqueness and unclonability (physical unclonable functions (PUFs)), digital identification is the only possible solution. Besides, digital content identification based on fingerprinting became a *de facto* standard in various multimedia security and management applications such as copyright protection, content filtering and automatic identification, authentication, broadcast monitoring, content tagging, etc.

A digital fingerprint represents a short, robust and distinctive content description allowing fast and privacy-preserving operations. In this case, all operations are performed on the fingerprint instead of on the original large and privacy-sensitive data.

Some important practical and theoretical achievements were reported during last years. The main efforts on the side of practical algorithms have been concentrated on robust feature selection and fast indexing techniques mostly borrowed from content-based retrieval applications [2], [3]. The information-theoretic limits of content identification under infinite length and ergodic assumptions have been investigated by Willems et. al. [4] using the jointly typical decoder. The detection-theoretic limits have been first studied in [5] under geometrical desynchronization distortions and a further extension of this framework was proposed in [6] for the case of finite-length fingerprinting and null hypothesis. The used decision rule is based on minimum Hamming distance decoder with a fidelity constraint under binary symmetric channel model. Since this decision rule requires the computation of likelihoods/distances between the query and all database entries, the complexity of the considered identification is exponential with the input length. Due to the additional fact that identification services are often outsourced to third parties and state authorities, the privacy of data owners is an important issue and remains largely unexplored.

Therefore, in this paper we introduce an information-theoretic framework for the analysis of private content identification based on finite length fingerprinting. Contrary to previous works, we consider alternative decoding rules and demonstrate their capability to achieve the identification capacity limit under asymptotic assumptions. Finally, we will show that content identification is closely related to the problem of erasure and list decoding [1] and further investigation of this connection might reveal many interesting insights to the analysis and design of future identification systems.

Notations. We use capital letters to denote scalar random variables X, bold capital letters to denote vector random variables $\mathbf{X}$, corresponding small letters x and small bold letters $\mathbf{x}$ to denote the realizations of scalar and vector random variables, respectively, i.e., $\mathbf{x} = \{x(1), x(2), ..., x(N)\}$. $\mathbf{b_x}$ is used to denote the binary version of $\mathbf{x}$. We use $X \sim p(x)$ to indicate that a random variable X follows $p_X(x)$.

II. Identification problem formulation

We will assume that the *data owner* has M entries in the database indexed by an index m, i.e., $\mathbf{x}(m) \in \mathbb{R}^N$, $1 \leq m \leq M$, where $M = 2^{LR}$ with R to be the identification rate of fingerprinting code-(M, L) and L stands for the fingerprint length. The index m is associated to all identification information (ownership, time of creation, distribution channel, etc.) and the data $\mathbf{x}(m)$ is some privacy sensitive part of the database represented by image, video, audio, biometric, PUFs, etc. . At the same time, the *data user* has a query data $\mathbf{y} \in \mathbb{R}^N$ that can be in some relationship with $\mathbf{x}(m)$ via a probabilistic model $p(\mathbf{y}|\mathbf{x})$ or can represent some irrelevant input $\mathbf{x}'$. The data user wishes to retrieve the identification

978-1-4244-8264-1/10/$26.00 ©2010 IEEE

information of $\mathbf{x}(m)$ that is the closest to the query $\mathbf{y}$ or reject the query, if no relevant database entry is found. For complexity and privacy reasons, the above identification is performed in the domain of digital fingerprints $\mathbf{b_x} \in \{0,1\}^L$ and $\mathbf{b_y} \in \{0,1\}^L$ that are short length, secure and robust counterparts of $\mathbf{x}$ and $\mathbf{y}$, respectively (Fig. 1). Moreover, to ensure adequate privacy protection of digital fingerprints, the data owner applies privacy amplification (PA) to produce protected version $\mathbf{b_u}(m)$ of $\mathbf{b_x}(m)$. The resulting fingerprints can be shared with third parties for various security and management services. In particular, the storage of the resulting codebook/database of protected fingerprints $\mathbf{b_u}(m)$, $1 \le m \le M$, and the content identification can be performed on a remote server that can be honest in terms of claimed functionalities but curious in terms of observing, analysis or leaking the stored data. The result of identification should be an estimate of index $\hat{m}$ of the corresponding closest entry or the erasure, i.e., null hypothesis. If the query is properly identified, the corresponding encrypted content $\mathbf{x}(m)$ or associated identification information is delivered to the data user using the predefined data exchange protocol.

In the scope of this paper, we will assume that the binary fingerprints are obtained by a dimensionality reduction transform $\mathbf{W}$ and binarization Q (Fig.1). The projected vectors of lower dimensionality $\tilde{\mathbf{x}}(m) \in \mathbb{R}^L$ and $\tilde{\mathbf{y}} \in \mathbb{R}^L$ are obtained from $\mathbf{x}(m)$ and $\mathbf{y}$ based on the dimensionality reduction transform:

$$\tilde{\mathbf{x}}(m) = \mathbf{W}\mathbf{x}(m), \tag{1}$$

$$\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{L \times N}$ and $L \le N$ and $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L)^T$ consists of a set of projection basis vectors $\mathbf{w}_i \in \mathbb{R}^N$ with $1 \le i \le L$. The dimensionality reduction transform is based on any randomized orthogonal matrix $\mathbf{W}$ (random projection transform) whose elements $w_{i,j}$ are generated from some specified distribution. An $L \times N$ random matrix $\mathbf{W}$ whose entries $w_{i,j}$ are independent realizations of Gaussian random variables $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ presents a particular interest for our study. In this case, such a matrix can be considered as an almost *orthoprojector*, for which $\mathbf{W}\mathbf{W}^T \approx \mathbf{I}_L$.[1] The selection of basis vectors with a Gaussian distribution also guarantees the Gaussian distribution of the projected coefficients. This will also be true for other statistics of the projection coefficients for sufficiently large N according to the Central Limit Theorem.

The binarization is performed as:

$$b_{\mathbf{x}_i} = \mathrm{sign}(\mathbf{w}_i^T \mathbf{x}), \tag{3}$$

where $b_{\mathbf{x}_i} \in \{0,1\}$, with $1 \le i \le L$ and $\mathrm{sign}(a) = 1$, if $a \ge 0$ and 0, otherwise. Since all projections are independent, it can be assumed that all bits in $\mathbf{b_x}$ will be independent and equiprobable for sufficiently large L.[2]

[1]Otherwise, one can apply special orthogonalization techniques to ensure perfect orthogonality.

[2]This assumption is only possible for independent input data. Since the transformed vectors will closely follow the Gaussian pdf but will not necessarily be decorrelated, one can apply the principle component analysis to decorrelate them, that, for the case of Gaussian data, will also provide their independence.

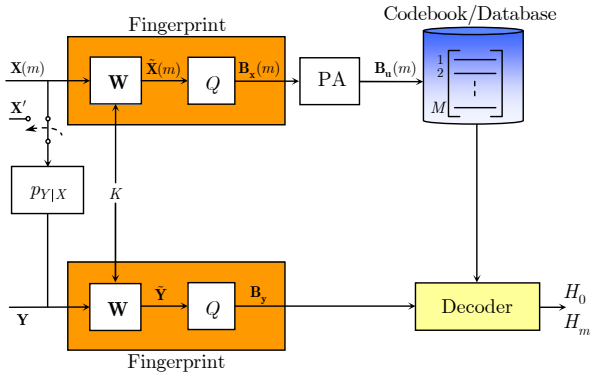

Fig. 1. Generalized block-diagram of private content identification based on digital fingerprinting.

The mismatch between the data owner fingerprint $\mathbf{b_x}$ and data user query fingerprint $\mathbf{b_y}$ can be modeled based on memoryless binary symmetric channel (BSC) model with a probability of bit error P_b. In the Appendix we show that $P_b = \frac{1}{\pi}\arccos(\rho_{\tilde{X}\tilde{Y}})$, where $\rho_{\tilde{X}\tilde{Y}}$ is a correlation coefficient between $\tilde{X}$ and $\tilde{Y}$.

Therefore, the main issues are: (a) the accuracy of identification defined in terms of probability of false acceptance of irrelevant entries and probability of wrong estimation of queries corresponding to the existing entries; (b) the complexity of identification; (c) the memory storage of resulting fingerprints; (d) the maximum number of correctly identifiable entries under the measures defined in (a) and length of fingerprints L; (e) the identification capacity under infinite L and (f) the privacy leak due to the fingerprint disclosure.[3]

To introduce a uniform consideration of the above issues, we define the generic content identification problem as a composite hypothesis test:

$$\begin{cases} H_0: & p(\mathbf{y}|H_0) = p(\mathbf{y}|\mathbf{x}'), \\ H_m: & p(\mathbf{y}|H_m) = p(\mathbf{y}|\mathbf{x}(m)),\ m = 1, \ldots, M, \end{cases} \tag{4}$$

and the corresponding private content identification based on binary fingerprinting as:

$$\begin{cases} H_0: & p(\mathbf{b_y}|H_0) = p(\mathbf{b_y}|\mathbf{b_x}'), \\ H_m: & p(\mathbf{b_y}|H_m) = p(\mathbf{b_y}|\mathbf{b_u}(m)),\ m = 1, \ldots, M. \end{cases} \tag{5}$$

In the binary fingerprinting domain, the link between $\mathbf{b_x}$ and between $\mathbf{b_y}$ and $\mathbf{b_x}$ and $\mathbf{b_u}$ can be considered based on the BSC models with corresponding bit error probabilities P_b and λ. The parameter λ corresponds to the BSC serving as a test channel for the compressed version $\mathbf{b_u}$ considered as the privacy amplification [7]. Under the above assumption, these two BSCs $\mathbf{b_x} \to \mathbf{b_y}$ and $\mathbf{b_x} \to \mathbf{b_u}$ can be considered as an equivalent channel obtained by their concatenation with the cross-probability P_{b_e} equals to the convolution $P_{b_e} = P_b * \lambda = P_b(1-\lambda) + \lambda(1-P_b)$. Under these conditions,

[3]In this paper, we do not analyze the identification from partial data such as block of image or frame of video due to the straightforward extension of our results to these cases under corresponding matching conditions.

the corresponding hypothesis (5) are:

$$\begin{cases} H_0\colon\ p(\mathbf{b_y}|\mathbf{b_x}') = \frac{1}{2^L}, \\ H_m\colon\ p(\mathbf{b_y}|\mathbf{b_u}(m)) = P_{b_e}^{d^H(\mathbf{b_y},\mathbf{b_u}(m))}(1-P_{b_e})^{L-d^H(\mathbf{b_y},\mathbf{b_u}(m))}, \end{cases} \tag{6}$$

where $d^H(.,.)$ denotes the Hamming distance.

Let the decision rule based on the public version $\mathbf{b_u}$ of $\mathbf{b_x}$ corresponds to the Forney's erasure decoder [1]:

$$p(\mathbf{b_y}|\mathbf{b_u}(m)) \geq 2^{\tau L}, \tag{7}$$

where τ is the threshold. We will show that this threshold should satisfy $\tau \leq -H_2(P_{b_e})$, where $H_2(.)$ denotes the binary entropy, for the unique decoding of index m and rejection hypothesis H_0.

Under (6), the decision rule (7) can be rewritten as:

$$d^H(\mathbf{b_y}, \mathbf{b_u}(m)) \leq L\gamma, \tag{8}$$

where $\gamma = \frac{-\tau+\log_2(1-P_{b_e})}{\log_2 \frac{1-P_{b_e}}{P_{b_e}}}$. We will refer to this decision rule as a *bounded distance decoder* (BDD) that produces a unique $\hat{m}$. It should be pointed out that under the proper selection of the threshold τ, one can also convert the content identification scheme based on the fingerprinting into a content-based retrieval system that produces multiple candidates in the proximity to $\mathbf{b_y}$ that corresponds to the list decoding formulation of Forney [8]. Therefore, we consider the content identification problem as a classical channel decoding problem. The benefits of soft information for the reduction of decoding complexity will be considered in the next section.

III. Identification system performance analysis

Proposition 1. The optimal threshold τ for the unique content identification under rule (7) should satisfy $\tau \leq -H_2(P_{b_e})$ to guarantee the minimum of overall identification error P_e.

Proof: We define the probability of false acceptance of some $\mathbf{y}$ produced by the database unrelated input $\mathbf{x}'$ as:

$$\begin{aligned} P_f(\gamma) &= \Pr[\bigcup_{m=1}^{M} d^H(\mathbf{B_u}(m), \mathbf{B_y}) \leq \gamma L|H_0] \\ &\leq_{(a)} \sum_{m=1}^{M} \Pr[d^H(\mathbf{B_u}(m), \mathbf{B_y}) \leq \gamma L|H_0] \\ &= M \Pr[d^H(\mathbf{B_u}(m), \mathbf{B_y}) \leq \gamma L|H_0] \\ &\leq_{(b)} 2^{-L(1-H_2(\gamma)-R)}, \end{aligned} \tag{9}$$

where (a) follows from the union bound and (b) from the Chernoff bound on the tail of binomial distributions $\mathcal{B}(L, 0.5)$ that results from $d^H(\mathbf{B_u}(m), \mathbf{B_y}) \sim \mathcal{B}(L, 0.5)$ under the hypothesis H_0.

The probability of incorrect identification is defined as:

$$\begin{aligned} P_{ic}(\gamma) &= \Pr[d^H(\mathbf{B_u}(m), \mathbf{B_y}) > \gamma L \\ &\quad \cup \bigcup_{n\neq m}^{M} d^H(\mathbf{B_u}(n), \mathbf{B_y}) \leq \gamma L|H_m] \\ &\leq_{(a)} \Pr[d^H(\mathbf{B_u}(m), \mathbf{B_y}) > \gamma L|H_m] \\ &\quad + \sum_{n\neq m}^{M} \Pr[d^H(\mathbf{B_u}(n), \mathbf{B_y}) \leq \gamma L|H_m] \\ &= \Pr[d^H(\mathbf{B_u}(m), \mathbf{B_y}) > \gamma L|H_m] \\ &\quad + (M-1)\Pr[d^H(\mathbf{B_u}(n), \mathbf{B_y}) \leq \gamma L|H_m] \\ &\leq_{(b)} 2^{-LD(\gamma||P_{b_e})} + 2^{-L(1-H_2(\gamma)-R)}. \end{aligned} \tag{10}$$

where $D(\gamma||P_{b_e}) = \gamma \log_2 \frac{\gamma}{P_{b_e}} + (1-\gamma)\log_2 \frac{1-\gamma}{1-P_{b_e}}$ is the divergence and where (a) follows from the union bound and (b) from the Chernoff bounds on the tails of binomial distribution $\mathcal{B}(L, P_{b_e})$ and $\mathcal{B}(L, 0.5)$.

Thus, combing bounds on P_f and P_{ic}, one obtains the overall identification probability of error[4]:

$$P_e = \frac{1}{2}P_f + \frac{1}{2}P_{ic} \tag{11}$$

$$\leq \frac{1}{2}\left(2^{-LD(\gamma||P_{b_e})} + 2\cdot 2^{-L(1-H_2(\gamma)-R)}\right), \tag{12}$$

that is minimized with $\gamma_{\text{opt}} = \frac{1-R+\log_2(1-P_{b_e})-1/L}{\log_2\left(\frac{1-P_{b_e}}{P_{b_e}}\right)}$ defines the optimal threshold minimizing the above probability of error. Comparing the BDD threshold γ of (8) with the obtained γ_{opt} for large L that diminishes as $1/L$ in the nominator, one can conclude that $\tau \leq -(1-R) = -H_2(P_{b_e})$ for the identification rates $R \leq 1 - H_2(P_{b_e})$. ∎

Remark 1. For the identification rate satisfying $R \leq 1 - H_2(P_{b_e})$, the above optimal threshold yields $\gamma_{\text{opt}} \leq P_{b_e}$. This means that the decoding region around each codeword is defined by the radius close to $P_{b_e}L$.

IV. Identification capacity and privacy leak

In this section, we will consider the case of asymptoticly large L.

Proposition 2. For $P_{b_e} \leq \gamma \leq \frac{1}{2}$ and if $H_2(\gamma) \leq 1 - R$ there exist codes with rate R and error probability P_e such that:

$$\lim_{L\to\infty} P_e = 0. \tag{13}$$

As soon as γ is arbitrarily close to P_{b_e}, the rate $R = 1 - H_2(P_{b_e})$ is achievable, and it is referred to as private identification capacity:

$$C_{id} = 1 - H_2(P_{b_e}). \tag{14}$$

The privacy leak L_p about $\mathbf{B_x}$ from the public $\mathbf{B_u}$ is defined by the mutual information between them[5]:

$$L_p = I(B_{\text{u}}; B_{\text{x}}) = 1 - H_2(\lambda). \tag{15}$$

[4]We assume the probability of submission of database related and unrelated queries to be same due to the lack of reliable prior models. If such priors are known, the overall probability is minimized by the corresponding threshold.

[5]A more conservative definition of privacy leak would be $I(B_{\text{u}}; X)$.

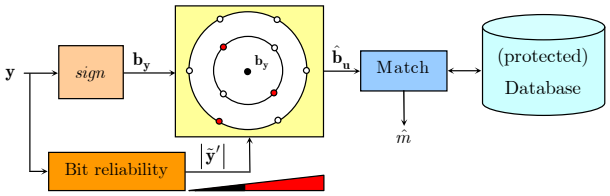

Fig. 2. The BDD based content identification protocol.

Remark 2. If privacy amplification is not applied, i.e., $\lambda = 0$ and $\mathbf{b_u} = \mathbf{b_x}$, one is interested in the maximization of identification capacity that yields:

$$C_{id} = I(B_\mathrm{x}; B_\mathrm{y}) = 1 - H_2(P_b), \tag{16}$$

$$L_p = I(B_\mathrm{u}; B_\mathrm{x}) = 1. \tag{17}$$

This result coincides with the identification capacity of biometric systems obtained in [4] derived based on jointly typical decoder. Therefore, in the analyzed setup, the privacy amplification is considered as the degradation of private data $\mathbf{b_x}$ by passing data via the BSC with a cross-probability λ and the identification capacity is influenced by the privacy amplification resulting in the increased P_{b_e}.

V. Complexity of fingerprint decoding

In this section, we consider the complexity of the BDD based content identification. An efficient decoding of random codes, i.e., the identification based on the completely unstructured fingerprints, represents a challenging computational problem. The exhaustive implementation of decoding rule (8) to verify all candidates $\mathbf{b_u}(m)$, $1 \leq m \leq M$ with $M \leq 2^{L(1-H_2(P_{b_e}))}$ would require $\mathcal{O}(L2^{L(1-H_2(P_{b_e}))})$ verifications that is a NP-hard problem. This is obviously prohibitively high for asymptoticly large L.

Alternatively, one can design the content identification decoder based on the assumption that the most likely candidates $\mathbf{b_u}(\hat{m})$ are within the Hamming sphere of radius γL around the codeword $\mathbf{b_y}$ (Fig. 2). In this case, the decoder just generates all possible codewords $\hat{\mathbf{b}}_\mathbf{u}$ within this sphere and sequentially validates their presence by querying the database for an exact match. If such a match is found, the index $\hat{m}$ of the corresponding codeword is declared as the result of identification. Otherwise, an erasure is declared. Such an identification protocol can be implemented on the remote server that only performs the matching while the list of possible candidates is generated on the side of data user [9].

Proposition 3. The cardinality of the list of candidates contained in the Hamming sphere of radius γL, where $0 \leq \gamma \leq \frac{1}{2}$, is defined by the partial sum of the binomial coefficients that can be bounded as:

$$\sum_{t=0}^{\gamma L} \binom{L}{t} \leq 2^{LH_2(\gamma)}, \tag{18}$$

where $0 \leq t \leq \gamma L$. Furthermore, asymptotically we have:

$$\lim_{L\to\infty} \sum_{t=0}^{\gamma L} \binom{L}{t} \doteq 2^{LH_2(\gamma)}. \tag{19}$$

Remark 3. According to Remark 1, $\gamma_{opt} \leq P_{b_e}$ that results in the identification complexity $\mathcal{O}(L2^{LH_2(P_{b_e})})$. The complexity of the BDD based on the Hamming sphere decoding is lower than the complexity of the above exhaustive search decoding for P_{b_e} satisfying the condition $H_2(P_{b_e}) < 1 - H_2(P_{b_e})$, i.e., $P_{b_e} < H_2^{-1}(0.5) \approx 0.11$.

The decoding complexity can be further reduced by analyzing the reliability of each bit based on the observed magnitude $|\tilde{y}|$. It is shown in the Appendix that the bit error probability for a given projection is proportional to $Q\left(\frac{|\tilde{x}|}{\sigma_z}\right)$, where σ_z represents the variance of noise in the projected domain and $|\tilde{x}|$ stands for the magnitude of projected coefficients. Unfortunately, $|\tilde{x}|$ is not available at the decoder side. That is why the distorted version $|\tilde{y}|$ is used to pick up the most likely codewords within the Hamming sphere and perform the matching only for them. Moreover, to avoid the computation of likelihoods for all codewords within the Hamming sphere, one can use another soft-strategy for content identification similar in spirit to classic GMD and Chase-2 decoding [10], [11]. One can sort all bits according to their reliabilities computed according to the sorted magnitude $|\tilde{y}'|$ (Fig. 2) and compute the maximum number of predicted errors in the observation $\mathbf{b_y}$ in the assumption of average bit error probability P_{b_e} as:

$$t_{b_{\mathrm{Max}}} = \mathcal{B}^{-1}(1-\epsilon, L, P_{b_e}), \tag{20}$$

where $\mathcal{B}^{-1}(.)$ is inverse binomial cumulative density function and ϵ is a arbitrarily small chosen probability that the number of error bits exceeds $t_{b_{\mathrm{Max}}}$.

Keeping the $L - t_{b_{\mathrm{Max}}}$ most reliable bits unchanged, one can quickly validate the remaining unreliable $t_{b_{\mathrm{Max}}}$ bits in the defined positions by performing $\mathcal{O}(L2^{t_{b_{\mathrm{Max}}}})$ verifications.

Remark 4. For large L, $t_{b_{\mathrm{Max}}}$ closely approaches LP_{b_e} and the complexity of the BDD based on the soft information about bit reliabilities is $\mathcal{O}(L2^{LP_{b_e}})$ that is smaller than the complexity of the BDD based on the Hamming sphere decoding $\mathcal{O}(L2^{LH_2(P_{b_e})})$. The obtained complexity is still exponential in L. However, it critically depends on the P_{b_e}, i.e., on the level of content degradation, contrary to the exhaustive search based decoding, used for the minimum distance based identification strategy [6], where the decoding complexity does not depend on the level of expected data degradation.

The indicative numbers for the requested fingerprint length, complexity and privacy leak for about 1 billion contents are shown in Table 1. It is clear that the privacy amplification increases the number of candidates on the list of codewords to be verified. At the same time, the BDD based on the bit reliability outperforms the exhaustive search decoding in terms of complexity for $P_{b_e} < 1 - H_2(P_{b_e})$, i.e., $P_{b_e} \leq 0.2271$. The use of block-based decoding [2] together with the described approach can further reduce the complexity.

Remark 5. We only discuss the reduction of decoding complexity of random fingerprinting codes by exploring soft information about the bit reliability. However, additional gain

TABLE I
INDICATIVE NUMBERS FOR REQUESTED FINGERPRINT LENGTH, DECODING COMPLEXITY AND PRIVACY LEAK FOR $M = 2^{30}$

	$\lambda = 0.05$			$\lambda = 0.10$		
	L	$\mathcal{O}(.)$	L_p	L	$\mathcal{O}(.)$	L_p
$P_b = 0.05$	55	$2^{5.2}$	0.71	73	$2^{10.1}$	0.53
$P_b = 0.10$	73	$2^{10.1}$	0.71	94	$2^{16.9}$	0.53
$P_b = 0.15$	97	$2^{18.0}$	0.71	125	$2^{27.5}$	0.53
$P_b = 0.20$	135	$2^{31.1}$	0.71	173	$2^{45.0}$	0.53

is also expected in the identification rate while moving from the binary version of B_{y}, i.e., $I(B_{\mathrm{u}}; B_{\mathrm{y}})$, to the soft version of $\tilde{Y}$, i.e., $I(B_{\mathrm{u}}; \tilde{Y})$ with $\tilde{Y} = B_{\mathrm{y}}|\tilde{Y}|$ the analysis of which is out of the scope of this paper.

VI. CONCLUSIONS

In this paper, we have presented a framework for private content identification based on binary fingerprints. The proposed approach is closely related to the erasure/list decoding of Forney and results in the bounded distance decoding for the binary case. Along this analysis we have introduced privacy amplification and derived the achievable identification rate for the finite-length fingerprints based on the theoretical performance analysis. We have also obtained the asymptotic results for long fingerprints and shown that they coincide with the previous results obtained for the jointly typical decoding. The complexity issues of private content identification on remote servers have been considered based on the soft information about the bit reliability. In our future research, we will also concentrate on the extension of our results to a broader family of distortions including geometrical transformations as well as propose even more efficient search strategies that are currently under testing for large databases.

ACKNOWLEDGMENT

This paper was partially supported by SNF project 119770.

APPENDIX

The bit error probability indicates the mismatch of signs between $\tilde{x}$ and $\tilde{y}$, i.e., $\Pr[\mathrm{sign}(\tilde{x}) \neq \mathrm{sign}(\tilde{y})]$:

$$P_b = \Pr[\tilde{Y} \geq 0 | \tilde{X} < 0] \Pr[\tilde{X} < 0] \quad (21)$$
$$+ \Pr[\tilde{Y} < 0 | \tilde{X} \geq 0] \Pr[\tilde{X} \geq 0], \quad (22)$$

or by symmetry for $\Pr[\tilde{X} < 0] = \Pr[\tilde{X} \geq 0] = \frac{1}{2}$ it can be rewritten as:

$$\begin{aligned} P_b &= \Pr[\tilde{Y} < 0 | \tilde{X} \geq 0] \\ &= 2 \int_0^{\infty} \int_{-\infty}^{0} p(\tilde{y}|\tilde{x}) p(\tilde{x}) d\tilde{y} d\tilde{x} \\ &= 2 \int_0^{\infty} P_{b|\tilde{x}} p(\tilde{x}) d\tilde{x}, \end{aligned} \quad (23)$$

where:

$$\begin{aligned} P_{b|\tilde{x}} &= \int_{-\infty}^{0} p(\tilde{y}|\tilde{x}) d\tilde{y} \\ &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{\frac{-(\tilde{y}-\tilde{x})^2}{2\sigma_Z^2}} d\tilde{y} \\ &= Q\left(\frac{|\tilde{x}|}{\sigma_Z}\right), \end{aligned} \quad (24)$$

stands for the bit error probability for a given projection coefficient $\tilde{x}$ under the assumption that $p(\tilde{x}, \tilde{y})$ corresponds to jointly Gaussian distribution in the random projection domain. The modulo sign is used for completeness of the consideration for the above symmetrical case when $\tilde{X} < 0$. One can immediately note that some projections can be more reliable in terms of bit error probability than others and the equation (24) can be a good measure of bit *reliability*.

Substituting (24) into (23), one obtains:

$$\begin{aligned} P_b &= 2 \int_0^{\infty} Q\left(\frac{|\tilde{x}|}{\sigma_Z}\right) \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{\frac{-\tilde{x}^2}{2\sigma_X^2}} d\tilde{x} \\ &= \frac{1}{\pi} \arccos(\rho_{\tilde{X}\tilde{Y}}), \end{aligned} \quad (25)$$

where $\rho_q^2 \tilde{X}\tilde{Y} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_Z^2}$ is the squared correlation coefficient between $\tilde{X}$ and $\tilde{Y}$.

REFERENCES

[1] G. D. Forney, "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, pp. 206–220, March 1968.

[2] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," in *International Workshop on Content-Based Multimedia Indexing*, Brescia, Italy, September 2001, pp. 117–125.

[3] F. Lefebvre and B. Macq, "Rash : RAdon Soft Hash algorithm," in *Proceedings of EUSIPCO - European Signal Processing Conference*, Toulouse, France, 2002.

[4] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in *Proc. 2003 IEEE Int. Symp. Inform. Theory*, Yokohama, Japan, June 29 - July 4 2003, p. 82.

[5] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun, "Robust perceptual hashing as classification problem: decision-theoretic and practical considerations," in *Proceedings of the IEEE 2007 International Workshop on Multimedia Signal Processing*, Chania, Greece, October 1–3 2007.

[6] A. L. Varna, A. Swaminathan, and M. Wu, "A decision theoretic framework for analyzing hash-based content identification systems," in *ACM Digital Rights Management Workshop*, Oct. 2008, pp. 67–76.

[7] C. H. Bennett, G. Brassard, C. Crépeau, and U. Maurer, "Generalized privacy amplification," *IEEE Transactions on Information Theory*, vol. 41, no. 6, pp. 1915–1923, Nov. 1995.

[8] F. Farhadzadeh, S. Voloshynovskiy, and O. Koval, "Performance analysis of identification system based on order statistics list decoder," in *IEEE International Symposium on Information Theory*, June 13-18 2010.

[9] S. Voloshynovskiy, F. Beekhof, O. Koval, and T. Holotyak, "On privacy preserving search in large scale distributed systems: a signal processing view on searchable encryption," in *Proceedings of the International Workshop on Signal Processing in the EncryptEd Domain*, Lausanne, Switzerland, 2009.

[10] G. D. Forney, "Generalized minimum distance decoding," *IEEE Transactions on Information Theory*, vol. 12, pp. 125–131, 1966.

[11] D. Chase, "A class of algorithms for decoding block codes with channel measurement information," *IEEE Transactions on Information Theory*, vol. 18, pp. 170–181, 1972.