Trading-off performance and complexity in identification problem

Taras Holotyak, Svyatoslav Voloshynovskiy, Oleksiy Koval, and Fokko Beekhof

University of Geneva, Department of Computer Science,
7 route de Drize, CH-1227, Geneva, Switzerland

ABSTRACT

In this paper, we consider an information-theoretic formulation of the content identification under search complexity constrain. The proposed framework is based on soft fingerprinting, i.e., joint consideration of sign and magnitude of fingerprint coefficients. The fingerprint magnitude is analyzed in the scope of communications with side information that results in channel decomposition, where all bits of fingerprints are classified to be communicated via several channels with distinctive characteristics. We demonstrate that under certain conditions the channels with low identification capacity can be neglected without considerable rate loss. This is a basis for the analysis of fast identification techniques trading-off theoretical performance in terms of achievable rate and search complexity.

Keywords: content identification, identification rate, complexity, storage memory, privacy, soft fingerprinting, bit reliability, channel polarization

1. INTRODUCTION

The amount of information stored in digital form grows exponentially. The Internet already contains billions of multimedia files available on-line in the form of digital images, video or audio. For example, the latest estimations evaluate Picasa or Flicker to have nowadays about 3 billion images and a similar number of video clips on YouTube and even larger number of images in the Google Image Search database.[1,2] At the same time, most of the Internet multimedia files are unlabeled and provided by different users. Therefore, there is a great interest in the development of systems allowing flexible management of these collections such as content-based retrieval, content filtering and automatic tagging.[3,4] Besides, some multimedia security applications require multimedia copyright protection, content origin identification, content tracking and broadcast monitoring.[5,6]

Similar problems also exist in the physical world where either humans or physical objects should be reliably identified based on their unique features or characteristics. In the case of humans these unique features correspond to biometrics (fingerprint, iris, *etc.*) that should be handled with special care to satisfy privacy-preserving requirements.[7,8] In the case of physical objects, the unique features are represented by specific unclonable characteristics, which can be acquired but can not be duplicated or reproduced with sufficient precision.[9]

Finally, numerous genetics and proteomics applications require either accurate identification of DNA sequences, proteins or peptides or detection of certain post-translation modifications considered to be as deviation from the baseline templates.[10,11] Curse of dimensionality in large scale databases, noise modifications and distortions raise numerous concerns in the search community regarding fast and accurate identification of these sequences.

Despite the different domains and origins, all these problems have in common the necessity to find the best matches to a given query according to the certain defined measure of similarity. The result of the search is given in the form of either the best unique match or a list of matches. Additionally, the list size may vary (for example as search results produced by Google) or be fixed to a certain value considered to be feasible for further manual processing. Therefore, identification with unique match is very close in formulation to pattern recognition, while list-based identification displays some remarkable similarity with nearest neighbor (NN) search. Additionally, unique identification can be considered as the NN problem with a result presented by the index on the top of

The contact author is S. Voloshynovskiy, E-mail: svolos@unige.ch the list. In any case, the NN problem in a multidimensional space is known to be NP-hard due to the curse of dimensionality.[12, 13]

Several multi-dimensional indexing methods, such as the popular KD-tree[14] or branch-and-bound techniques, *etc.*, have been proposed to reduce search complexity. However, for large dimensional problems it turns out[15] that such approaches are not more efficient than the brute-force exhaustive distance calculation, whose complexity is $\mathcal{O}(ML)$, where M is the size of the database and L stands for the length of a feature vector used for content representation.

Therefore, current state-of-the-art techniques overcame this issue by performing approximate matching.[16–18] The key idea shared by these algorithms is to find the best matches with *only* high probability close to $1 - \epsilon$, where ϵ is a small positive value, instead of the exact match with probability 1. In this respect the most common distance for matching is considered to be the Euclidian one.

One of the first techniques of approximate matching in the Euclidian distance metric is Euclidian *Locality Sensitive Hashing* (LSH),[16, 19] which has been successfully used for image search based on local descriptors,[20] 3D object indexing[21] and manually prefiltered proteomics data.[22] However, for real data, LSH is outperformed by heuristic methods.[18]

Moreover, identification systems in general are facing not only matching an accuracy-complexity trade-off. As soon as these systems are applied to practical problems, the memory storage of the indexing structure and its update due to new entries start to play a significant role. For example, in the case of Euclidian LSH, the memory usage may even be higher than that of original data vectors. The same critical argument also applies to audio search systems based on robust fingerprinting[23] and image indexing[24] that both essentially represent a concatenated one-dimensional form of the LSH strategy. Similar approaches are based on small binary codes,[25] semantic hashing[26] and spectral hashing.[27] Only relatively recently, researchers have tried to design memory limited identification systems. This is a key criterion problem involving large scale applications,[28, 29] where millions to billions of images have to be indexed.

It is also worth noting that due to the large scale and required huge computational power, the identification is either outsourced or considered to be securely executed on third party servers. The examples of existing outsourced systems are numerous and include outsourcing of email services, P2P data sharing, Google-like search architectures and social networks. At the same time, all the above multimedia, biometrics, genomics or proteomics applications more broadly and intensively enter into the field of human privacy that is a very sensitive issue. Any naive use and implementation of large scale identification systems might therefore lead to a severe privacy problem.

The privacy model in the identification framework can be considered in a simplified setting,[30] when the data owner stores some possibly privacy-sensitive information on an untrusted server, which models the outsourced service provider. At the same time, these data should be provided to several authorized users, who are allowed to access and search it. The data owner and data user might possess a common secret. Since the server might be honest-but-curious or even malicious, the data provided by the data owner to the server should be properly handled to avoid any privacy disclosure, but at the same time allow accurate search.

Therefore, the data user query for identification search should reveal as little information as possible to the server and unauthorized parties about the stored data itself and query of interest to the data user. Moreover, the data user might possess only some inexact distorted or noisy version of data owner's data. Thus, exact matching strategies are not applicable in this case. The state-of-the-art techniques are therefore mostly based on a *searchable encryption*, which consists in the identification in the protected domain directly on the server. Not pretending to be exhaustive in our review, we refer the reader to[31] for the classification of different secure search strategies in the encrypted domain. The existing techniques can be classified in four groups: *indices based search*, when the actual search is performed based on an added index (hash of encrypted index);[32] *trapdoor encryption based search*, when the search is performed based on a predefined codebook of encrypted codewords;[33–35] *secret sharing based search*, when the data is distributed overall several servers, which are assumed to not collude;[36] and *homomorphic encryption based search*, when the search is performed directly in the encrypted domain using similarity features in the class of homomorphic encryption.[37–39] Being attractive in terms of both communication and security, all these approaches currently work only for exact matching or require brute force matching like in the case of homomorphic encryption. It is obvious that for many real world applications these techniques are unfeasible.

Therefore, one of the goals of the proposed project is to consider privacy preserving protocols enabling identification in the protected domain based on possibly distorted or inexact queries with a limited number of communication sessions per query.

Another very important problem recently emerged in multimedia applications is security of content identification/search systems. In recent years, content identification systems are used to spot and filter upload of copyrighted materials on sharing platforms such as YouTube to either block or monetize it. Content identification systems are also considered as preventions of downloads in P2P networks in a so-called *graduated response.* Under such conditions, the originally friendly environment of content search for exchange systems turns out to be quite aggressive in sense of reply to the introduced filtering, restriction and controlling functionalities. Since the system aims at protecting value, serious hackers will try to circumvent these systems. Therefore, it is very timely to formulate the question about the security side of content identification systems in general. To our knowledge, this subject remains completely unexplored besides several recent results presented in.[40]

In fact, based on the knowledge about the robustness of identification systems to different modifications attackers might develop new attacks. Since practically all identification systems are well described in literature and there is no secret key used, the hacker might exploit this valuable knowledge in his attacks too. To exemplify these strategies, one can mention *content concealment* (uploading illegal materials, the hacker learns the operational possibility of system) and *abnormally frequent identification* (as pirates tweak HTML pages to get ranked higher in textual search engines in *black hat SEO* attacks), one can tweak for example the visual or audio contents such that they always get (artificially) ranked high in the resulting list. Thus, a dishonest content owner may increase his revenue thanks to this exaggerated advertisement.

Finally, the last but not the least problem is the investigation of performance (identification accuracy) under the above requirements. When the identification systems are applied to large-scale problems, it is not always sufficient to validate system performances on small test databases used in benchmarking as it is done in most of scientific publications besides some rare exceptions.[41] From the other side, it is also unfeasible that a small group of researchers can practically test billion-size applications by themselves. Therefore, development of accurate information-theoretic models of these systems and corresponding methods allowing to predict the system's performance under database scaling is of great practical importance.

The state-of-the-art information-theoretic contributions to the identification problem solution can be classified in three groups:(a) investigation of theoretical performance limits; (b) investigation of the performance-complexity trade-off; (c) investigation of the performance-storage memory trade-off. The information-theoretic performance limits of content identification under infinite length and ergodic assumptions has been investigated by Willems et. al.[42] using the jointly typical decoder. Along similar lines, the detection-theoretic limits have been first studied in[43] under geometrical desynchronizing distortions and a further extension of this framework was proposed in[44] 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 the 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 scheme is exponential with the input length.

The second group of identification-theoretic methods addresses the performance-complexity trade-off. Unfortunately, this very important problem received little attention in the literature. Only recently, Willems published the first paper[45] dedicated to this problem. The main idea behind this approach is to split M database entries into $\sqrt{M}$ groups with $\sqrt{M}$ codewords in each group. The group index is deduced based on quantization and the remaining task is to find a valid candidate within the identified group based on $\sqrt{M}$ checks. Therefore, preserving the achievable identification rate to be close to the identification capacity limit, C_{id}, one can find the correct index with complexity $\mathcal{O}(L\sqrt{M})$ or $\mathcal{O}(L\sqrt{2^{LC_{id}}})$. Unfortunately, for large M this fractional complexity might be still prohibitively high. At the same time, it is not clear what is the relationship of this approach to the reviewed above heuristic methods.

Finally, the third group of methods addresses the problem of identification rate R_{id} - memory (or storage rate) R_s trade-off without considering previous issues. Westover and O'Sullivan considered this trade-off in a pattern recognition formulation[46] and Tuncel analyzed it in a large-scale database management setup.[47] Unfortunately, one could admit that the security and privacy constraints within the information-theoretic framework remain largely unexplored.

To address the above issues, we will consider the data stored in the database as digital fingerprints, i.e., compressed, robust and secure representations of digital contents. In summary, to design a practical identification system one should find an optimal trade-off between performance in terms of probability of error (a.k.a. robustness), security, privacy, memory storage, identification rate and complexity as schematically shown in Figure 1. To our best knowledge the systematic analysis of this trade-off was not performed.

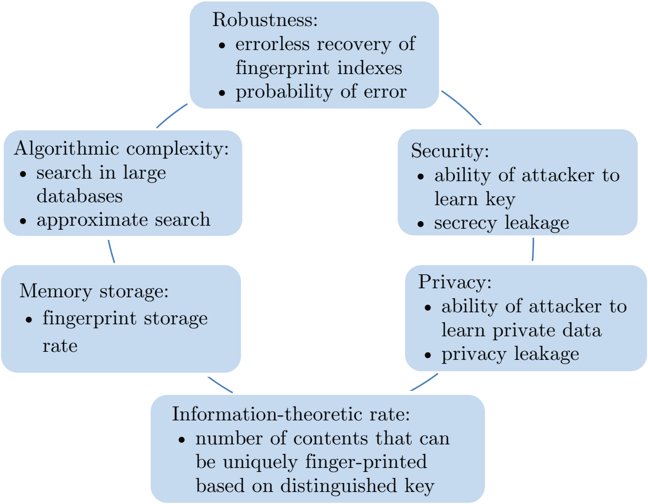

Robustness:
• errorless recovery of fingerprint indexes
• probability of error
Algorithmic complexity:
• search in large databases
• approximate search
Security:
• ability of attacker to learn key
• secrecy leakage
Memory storage:
• fingerprint storage rate
Privacy:
• ability of attacker to learn private data
• privacy leakage
Information-theoretic rate:
• number of contents that can be uniquely finger-printed based on distinguished key

Figure 1: Schematic relationship between main trade-off requirements to identification systems.

New information-theoretic approaches to fast, secure, memory efficient and accurate identification should be proposed and investigated. This is a very challenging problem that should address compromises between various above conflicting requirements.

This is why one of the main objectives of this paper is to introduce an information-theoretic framework able to properly model, analyze and finally optimally trade-off these requirements. In turn, this should lead to the development of new practical methods so urgently needed in many applications.

Moreover, many practical fingerprinting techniques neglect information about bit reliability provided by the magnitude of fingerprint coefficient. Therefore, we will try to fulfill this gap by considering soft-fingerprinting that can resolve the above trade-offs. The proposed methodology is based on a framework of sign-magnitude decomposition that naturally leads to a concept of channel splitting into reliable and unreliable sub-channels. We show that under certain conditions most of identification rate can be concentrated in the reliable channel. We call this effect a *channel polarization*. However, this requires still exponential complexity of decoding for the considered random codebooks. Therefore, to relax this critical constrain we demonstrate that slight deviation from the optimal channel splitting, accompanied with the minor rate loss, might provide a considerable gain in identification complexity. Such a methodology can be considered as a theoretical basis for the analysis and generalization of approximate search strategies. The paper is organized as follow. In Section 2, we provide the identification problem formulation. Section 3 introduces the framework of sign-magnitude decomposition and Section 4 explains the channel polarization. Fast approximate identification trading-off achievable rate and complexity is presented in Section 5. Section 6 concludes the paper.

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 designate 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 $p_X(x)$.

2. IDENTIFICATION PROBLEM FORMULATION

To resolve the performance-complexity trade-off in the identification problem, a dimensionality reduction based on random projections and the concept of bit reliability were proposed.[48] The main idea behind the random projection application consists in the removal of ambiguity about the data *prior* statistics, while information about bit reliabilities is used to reduce the identification complexity. The schematic diagram of the proposed approach is shown in Fig. 2.

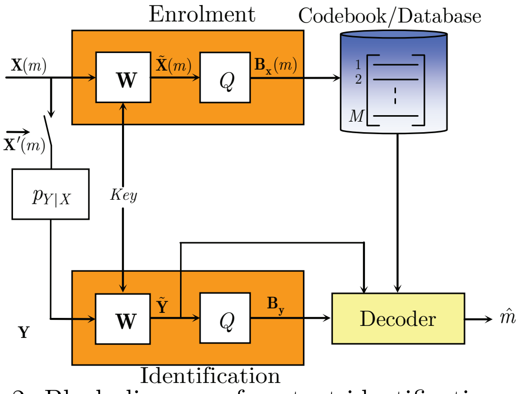


Figure 2: Block-diagram of content identification system.

Under such a formulation, the identification system can be analyzed within digital communication framework. It functions in two operating modes: enrollment and identification. During the enrollment, the contents denoted as $\mathbf{x}(m) \in \mathbb{R}^N$, $m = 1, ..., M$, are transformed into the fingerprints using the following two stage procedure. First, $\mathbf{x}(m)$ of dimensionality N are projected onto a lower dimensional J ($J \leq N$) space via:

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

where $\mathbf{W} \in \mathbb{R}^{J \times N}$, $\mathbf{W} = (\mathbf{w}_1, ..., \mathbf{w}_J)^T$, and $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$. The reason for such a design of the projection matrix is to ensure the invariance of the system to the deviation and lack of *prior* statistics of $\mathbf{X}(m)$. It is not difficult to demonstrate that for any i.i.d. generated $\mathbf{x}(m)$, $\widetilde{\mathbf{x}}(m)$ will have Gaussian statistics with approximately preserved diagonal covariance matrix, i.e., $cov\left[\widetilde{\mathbf{X}}\right] \approx \sigma_X^2 \mathbb{I}_J$, where $\mathbb{I}_J$ is an identity matrix with $\mathbf{W}\mathbf{W}^{\mathbf{T}} \approx \mathbb{I}_{\mathbf{J}}$. Secondly, the projection output is converted to the binary form as follows:

$$\mathbf{b}_{\mathbf{x}}(m) = sign\left(\mathbf{W}\mathbf{x}(m)\right), \tag{2}$$

where *sign* denotes function that extracts the sign of a real number. The main purpose of the binarization stage is to protect the privacy of stored contents ($I(B_{\mathrm{x}}; \tilde{X}) = 1$), as well as to tackle data storage and computational complexity aspects.

In the identification mode, the query, which is distorted by discrete memoryless channel (DMC), $p\left(\mathbf{y}|\mathbf{x}\right)$, version of $\mathbf{x}$ or $\tilde{\mathbf{x}}$, is converted to the binary form according to:

$$\widetilde{\mathbf{y}} = \mathbf{W}\mathbf{y} = \mathbf{W}(\mathbf{x} + \mathbf{z}),\ \mathbf{b}_{\mathbf{y}} = sign\left(\mathbf{W}\mathbf{y}\right). \tag{3}$$

If only binary part $\mathbf{b}_{\mathbf{y}}$ of the query $\mathbf{y}$ is used for the identification, we will refer to it as a *hard fingerprinting.* However, if in addition the magnitude $|\widetilde{\mathbf{y}}|$ is utilized, we will call this a *soft fingerprinting.*

It is important to note that, similarly to $\mathbf{x}$, any additive i.i.d. noise $\mathbf{z}$ will be converted to the additive Gaussian one with $cov\left[\widetilde{\mathbf{Z}}\right] \approx \sigma_Z^2 \mathbb{I}_J$. Finally, the decoder that observes $\mathbf{y}$ and has access to the enrolled database should decide, which one out of M alternatives is present at the system input. In most identification system designs, the bounded distance decoding (BDD) is used.[49] In order to find a match with the channel output

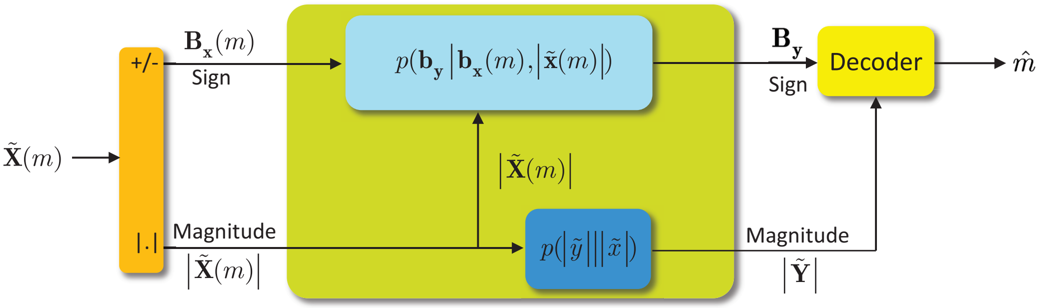

$\mathbf{B}_{\mathbf{x}}(m)$
Sign
+/-
$\tilde{\mathbf{X}}(m)$
|.|
Magnitude
$|\tilde{\mathbf{X}}(m)|$
$p(\mathbf{b_y}|\mathbf{b_x}(m),|\tilde{\mathbf{x}}(m)|)$
$|\tilde{\mathbf{X}}(m)|$
$p(|\tilde{y}||\tilde{x}|)$
$\mathbf{B}_{\mathbf{y}}$
Sign
Decoder
$\hat{m}$
Magnitude
$|\tilde{\mathbf{Y}}|$

Figure 3: Fingerprinting model of communications based on sign-magnitude decomposition.

in the given codebook, this method performs an exhaustive search over the entire codebook and produces the estimate $\hat{m} = \{\epsilon, 1, 2, \ldots, M\}$, where ϵ is the erasure or rejection

$$d^H(\mathbf{b_y}, \mathbf{b_x}(m)) \leq \gamma L, \tag{4}$$

where $d^H(\cdot,\cdot)$ denotes Hamming distance between binary vectors and product γL defines the BDD threshold, more detailed description of which is given in.[49,50] In the case of additive Gaussian model of observation $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} + \tilde{\mathbf{Z}}$, the decoding achieves the identification capacity:[42]

$$C_{id} = I\left(\tilde{X};\tilde{Y}\right) = \frac{1}{2}log_2\left(1+\frac{\sigma^2_{\tilde{X}}}{\sigma^2_{\tilde{Z}}}\right). \tag{5}$$

Operation with real-valued $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ leads to the optimal in terms of performance solution of the identification problem, but characterized by complexity $\mathcal{O}(MJ)$. However, the computational complexity of this decoder is prohibitively high for the practical usage. Additionally, the storage of $\tilde{\mathbf{X}}$ is not desirable for the privacy reasons. Thus, in the following consideration we will concentrate on the setup, where query $\tilde{\mathbf{y}}$ is comparing with only binary counterparts $\mathbf{b_x(m)}$, $1 \leq m \leq M$. It is important to point out that the probability of bit error between the stored binary fingerprints and binarized query is defined by[49]

$$P_{b|\tilde{x}} = Q\left(\frac{|\tilde{x}|}{\sigma_Z}\right), \tag{6}$$

where $Q(\cdot)$ stands for Q-function and σ_Z denotes the standard deviation of the Gaussian channel.

3. SIGN-MAGNITUDE DECOMPOSITION

In the case of real-valued signals, projected coefficients can be decomposed as it is shown in Fig. 3. Keeping in mind Gaussianity of $\tilde{X}$ and independence of sign and magnitude components one can split source $\tilde{X}$ with $h(\tilde{X}) = 1/2log_2(2\pi e\sigma^2_X)$ into independent subsources $|\tilde{X}|$ with differential entropy $h(|\tilde{X}|) = 1/2log_2(1/2\pi e\sigma^2_X)$ and B_X with entropy $H(B_X) = 1$, where $h(\tilde{X}) = H(B_X) + h(|\tilde{X}|)$. Therefore, general identification channel $p(\tilde{y}|\tilde{x})$ can be decomposed into two sub-channels $b_{\mathrm{x}} \rightarrow b_{\mathrm{y}}$ and $|\tilde{x}| \rightarrow |\tilde{y}|$ with corresponding interrelations. Storing only $\mathbf{B_x}$, the information transmission is performed using binary symmetrical channel (BSC), whose state is determined by $|\tilde{x}|$, while information transmission thought the $|\tilde{x}| \rightarrow |\tilde{y}|$ channel is skipped.

Since the application of random projection mapper $\mathbf{W}$ with $W_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$ leads to the additive Gaussian noise observation model $\tilde{y} = \tilde{x} + \tilde{z}$, where $\tilde{z}$ represents the zero-mean Gaussian noise with the variance σ^2_Z in the projected domain. Therefore, the behavior of bits transmission in BSC $b_{\mathrm{x}} \rightarrow b_{\mathrm{y}}$ is completely characterized by probability of bit error $P_{b|\tilde{x}}$ (6). Under such a setting, many existing fingerprinting systems can be considered as *hard* fingerprinting, i.e., those that do not use information about the channel state, and *soft* fingerprinting, i.e., those that benefit from this knowledge. Therefore, depending on the availability of *channel state information* (CSI) one can distinguish 3 major cases:

- *no CSI* ($S = \oslash$), where only $p(\tilde{x})$ is known (*hard fingerprinting*). The identification rate under the hard fingerprinting is:

$$R_{id|0} = I(B_{\mathrm{x}}; B_{\mathrm{y}}|\oslash) = 1 - H_2(\bar{P}_b), \tag{7}$$

  where $\bar{P}_b = \int_{-\infty}^{+\infty} Q\left(\tilde{x}/\sigma_Z\right) p(\tilde{x}) d\tilde{x}$ is the average probability of bit error;

- *perfect CSI* ($S = |\tilde{x}|$) (*soft fingerprinting* with perfect CSI). The identification rate under the soft fingerprinting with perfect CSI is:

$$R_{id|\tilde{x}} = \int_{-\infty}^{+\infty} I(B_{\mathrm{x}}; B_{\mathrm{y}}||\tilde{X}| = |\tilde{x}|)p(\tilde{x})d\tilde{x} = 1 - 2\int_{0}^{+\infty} H_2\left[Q\left(\frac{\tilde{x}}{\sigma_Z}\right)\right] p(\tilde{x})d\tilde{x}; \tag{8}$$

- *partial CSI* ($S = |\tilde{y}|$) (*soft fingerprinting* with partial CSI). The identification rate under the soft fingerprinting with partial CSI is:

$$R_{id|\tilde{y}} = 1 - 2\int_{0}^{+\infty} H_2\left[\int_{-\infty}^{+\infty} Q\left(\frac{\tilde{x}+z}{\sigma_Z}\right) p(\tilde{z})d\tilde{z}\right] p(\tilde{x})d\tilde{x}. \tag{9}$$

The identification rates under hard fingerprinting, soft fingerprinting with perfect and partial CSI are shown in Fig. 4. The observation model is considered in terms of the signal-to-noise ratio (SNR) defined as $SNR = 10\log_{10}(\sigma_X^2/\sigma_Z^2)$. For the comparison reasons the capacity of the AWGN identification channel (5) is also shown in Fig. 4. Based on the obtained results one can conclude that presence of perfect CSI at the decoder enhances the identification rate with respect to the hard fingerprinting. Also, partial CSI at the decoder enhances the identification rate for the high SNRs, i.e., in the region where it is not severely corrupted by the observation noise, and contrarily slightly degrades the rate with respect to the hard fingerprinting for the low SNRs. Achievable identification rate for all considered fingerprinting techniques is saturated at 1 for the high SNR. The gap between the identification rate of the AWGN channel and fingerprinting one is in part of $I(|\tilde{X}|; |\tilde{Y}|)$.

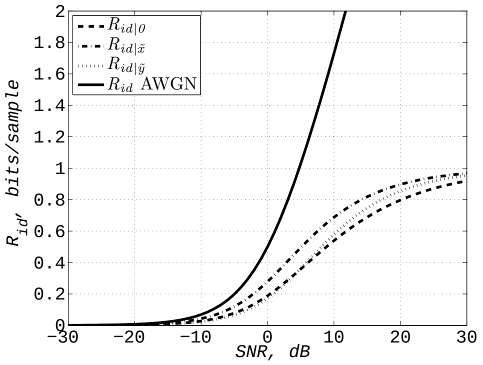


Figure 4: The identification rates under hard fingerprinting and soft fingerprinting with perfect and partial CSI.

We also introduce the practical model achieving the above theoretical limits based on a *channel splitting*. The channel splitting model assumes that bits are transmitted via several BSCs with parameters defined by the CSI. In the most simple case of 2-channel splitting, two BSCs are considered. At this moment, we assume that $S = |\tilde{\mathbf{x}}|$. The channel splitting can be performed based on the thresholding of coefficient magnitudes with the threshold T (Fig. 5). The L fingerprint bits related to the large magnitude coefficients are considered as those belonging the good BSC with the cross-over probability P_b^G and the remaining to the bad one with P_b^B. The cross-over probabilities for good and bad channels based on the perfect CSI ($S = |\tilde{\mathbf{x}}|$) are:

$$P_{b|\tilde{x}}^G = \frac{2}{\Pr^G}\int_{T}^{+\infty} Q\left(\frac{\tilde{x}}{\sigma_Z}\right) p(\tilde{x})d\tilde{x}, \tag{10}$$

$$P_{b|\tilde{x}}^B = \frac{2}{\Pr^B}\int_{0}^{T} Q\left(\frac{\tilde{x}}{\sigma_Z}\right) p(\tilde{x})d\tilde{x}, \tag{11}$$

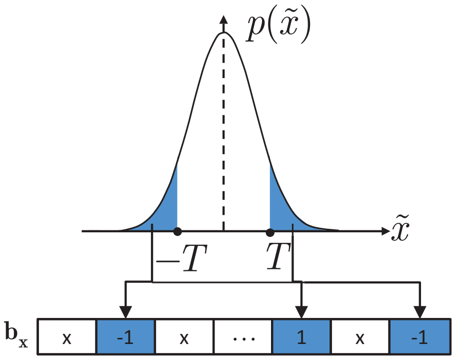


Figure 5: Channel splitting based on thresholding.

where $\Pr^G = 2\int_T^{+\infty} p(\tilde{x})d\tilde{x}$ and $\Pr^B = 2\int_0^T p(\tilde{x})d\tilde{x}$ correspond to the probabilities of observing the good and bad channels, respectively. The corresponding identification rates are:

$$R^G_{id|\tilde{x}} = \Pr^G\left(1 - H_2\left(P^G_{b|\tilde{x}}\right)\right), \tag{12}$$

$$R^B_{id|\tilde{x}} = \Pr^B\left(1 - H_2\left(P^B_{b|\tilde{x}}\right)\right), \tag{13}$$

and the total rate is:

$$R^{2\mathrm{Ch}}_{id|\tilde{x}} = R^G_{id|\tilde{x}} + R^B_{id|\tilde{x}}. \tag{14}$$

The cross-over probabilities for good and bad channels based on the partial CSI ($S = |\tilde{\mathbf{y}}|$) are:

$$P^G_{b|\tilde{y}} = \frac{2}{\Pr^G}\int_T^{+\infty}\left[\int_{-\infty}^{+\infty} Q\left(\frac{\tilde{x}+\tilde{z}}{\sigma_Z}\right)p(\tilde{z})d\tilde{z}\right]p(\tilde{x})d\tilde{x}, \tag{15}$$

$$P^B_{b|\tilde{y}} = \frac{2}{\Pr^B}\int_0^{T}\left[\int_{-\infty}^{+\infty} Q\left(\frac{\tilde{x}+\tilde{z}}{\sigma_Z}\right)p(\tilde{z})d\tilde{z}\right]p(\tilde{x})d\tilde{x}. \tag{16}$$

The corresponding identification rates are:

$$R^G_{id|\tilde{y}} = \Pr^G\left(1 - H_2\left(P^G_{b|\tilde{y}}\right)\right), \tag{17}$$

$$R^B_{id|\tilde{y}} = \Pr^B\left(1 - H_2\left(P^B_{b|\tilde{y}}\right)\right), \tag{18}$$

and the total rate is:

$$R^{2\mathrm{Ch}}_{id|\tilde{y}} = R^G_{id|\tilde{y}} + R^B_{id|\tilde{y}}. \tag{19}$$

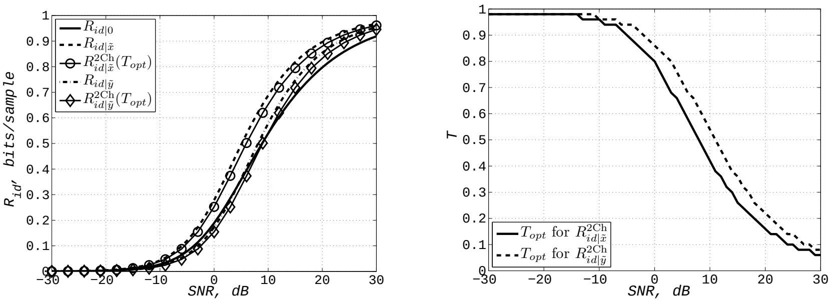
The total cross-over probability P_b remains the same regardless the value of threshold T:

$$P_b = 2\int_0^{+\infty} Q\left(\frac{\tilde{x}}{\sigma_Z}\right)p(\tilde{x})d\tilde{x} = P^G_{b|\tilde{x}}\Pr^G + P^B_{b|\tilde{x}}\Pr^B. \tag{20}$$

4. CHANNEL POLARIZATION

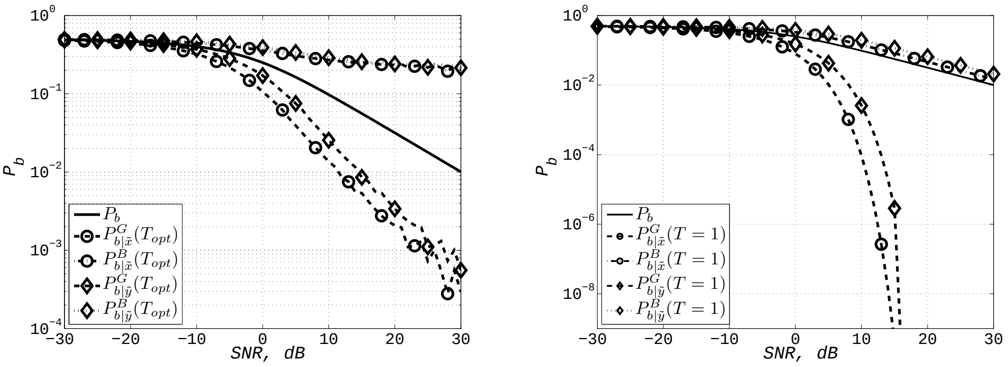
The channel splitting by the selection of threshold T can be performed according to several strategies:

- **Strategy 1**: maximize the total rate $R^{2\mathrm{Ch}}_{id|\tilde{x}}$ or $R^{2\mathrm{Ch}}_{id|\tilde{y}}$ to approach upper theoretical limits $R_{id|\tilde{x}}$ or $R_{id|\tilde{y}}$, respectively, that gives optimal values of thresholds $T_{\mathrm{opt}|\tilde{x}}$ and $T_{\mathrm{opt}|\tilde{y}}$ for each SNR;
- **Strategy 2**: minimize probabilities $P^G_{b|\tilde{x}}$ or $P^G_{b|\tilde{y}}$ for search complexity reasons[48] that creates a sort of *channel polarization* after some SNR, when certain amount of bits can be communicated without errors.



(a) (b)

Figure 6: Approaching theoretical rates based on 2-channel splitting model for the optimal threshold selection: (a) achievable identification rates under different CSIs, (b) optimal thresholds for perfect and partial CSI.



(a) (b)

Figure 7: Probabilities of bit errors for: (a) T_{opt} and (b) $T = 1$.

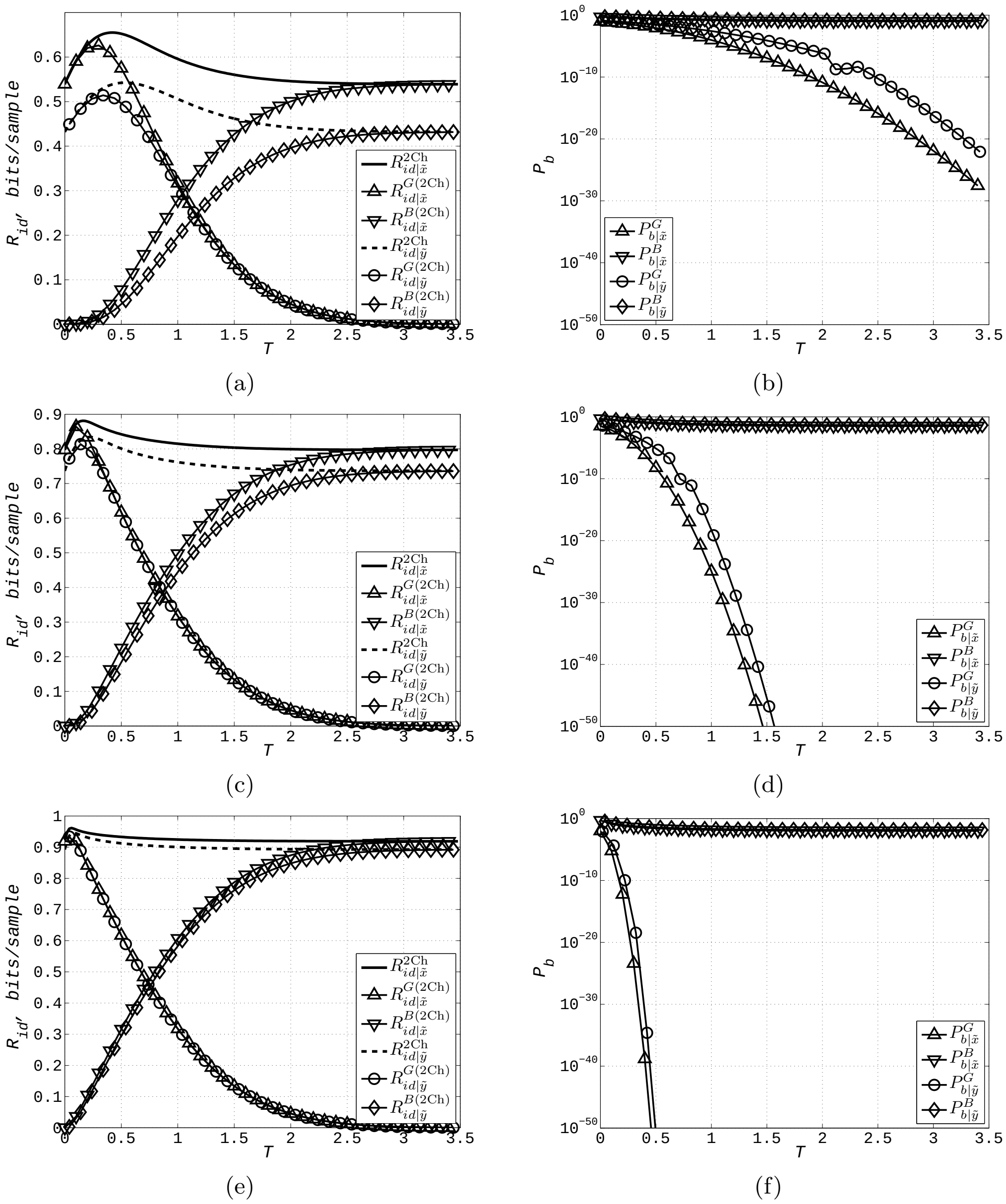
According to strategy 1, the binary channel splitting approaches theoretical performance limits under the optimal threshold selection as shown in Fig. 6a. The remaining gap is easily compensated by more accurate models using more than 2 channels splitting model. The optimal thresholds $T_{\mathrm{opt}|\tilde{x}}$ and $T_{\mathrm{opt}|\tilde{y}}$ are shown in Fig. 6b.

To exemplify strategy 2 we show in Fig. 7a the pairs of cross-over probabilities for good and bad channels under perfect and partial CSI that resulted from the strategy 1. Fig. 7b shows the same pairs for the fixed threshold T, where one can clearly observe the significant reduction of $P^G_{b|\tilde{x}}$ or $P^G_{b|\tilde{y}}$ that asymptotically goes to zero for $SNR > 15dB$.

Obviously, this strategy is not optimal in terms of total rate maximization. However, the expected loss in the approaching the total rate with the minimization of $P^G_{b|\tilde{y}}$ is minor. The dependence of achievable rates and cross-over probabilities on the threshold T are shown in Fig. 8 for SNR in the range of $10 \div 30dB$. The rate of convergence of $P^G_{b|\tilde{y}}$ to zero is exponential with respect to the loss in the identification rate. The plots also clearly demonstrate the polarization effect, when almost all useful rate is concentrated in the good channel and the bad channel can be completely disregarded from the data transmission point of view.

5. LOW COMPLEXITY IDENTIFICATION

In this Section, we introduce the practical framework for trading-off performance and complexity in the identification. Contrarily to most state-of-the-art approximate search strategies, where the identification rate remains undefined,[16, 27, 51] our goal is to develop a technique, where performance of the system will be evaluated not only in terms of the probabilities of making incorrect decisions, but also assessing the identification rate. The main advantage one can obtain by including into the consideration this digital communication parameter is the ability to determine the database cardinality, below which all quires can be uniquely identified. Under these



(a) (b)

(c) (d)

(e) (f)

Figure 8: Achievable rates and cross-over probabilities for channel splitting at SNR=10, 20, and 30dB.

constraints, the implementation will target the exact matching requiring only reading/writing operations rather than multiplication summation operations.

Prior to the consideration of the proposed setup, we would like to point out that optimal performance in communication protocols over the BSC with the cross-over probability P_b in terms of achievable rates is attained by the minimum distance decoding counterpart of (4) that requires $\mathcal{O}(MJ)$ distance computations for a sufficiently large J. Although, efficient hardware tools are currently available for Hamming distance computation,[52] search complexity can still represent an issue preventing system implementation in real time. Therefore, our objective is to benefit from the available CSI extracted from $\mathbf{y}$ targeting reduction of the computational burden.

Our main motivation comes with the channel splitting paradigm that was introduced and developed in the above Section. According to this paradigm, one can assume that a vanishing probability of error $P^{G}_{b|\tilde{y}}$ can be achieved in good channels under a proper selection of the threshold T. Therefore, one can claim that the outputs of these good channels will almost surely replicate their inputs and the corresponding bits in a set of positions of $\mathbf{b}_{\mathbf{x}}(m), 1 \leq m \leq M$, and $\mathbf{y}$, which can be identified based on the CSI, coincide.

Using such an errorless transmission phenomenon, we propose the following identification system architecture based on channel polarization and sign-magnitude decomposition (Fig.2). According to the advocated approach, it is assumed that the entire database obtained in the enrollment stage is permanently stored in the RAM of the identification server. The database is composed with M binary vectors of length J that are obtained by applying random projections and binarization (2) to the contents $\mathbf{x}(m), 1 \leq m \leq M$.

At the identification stage, the query is converted to the binary format according to (3) and its magnitude is stored accordingly. Then, the positions of $L(L < J)$ most reliable bits are determined based on the imperfect CSI that is represented by sorted $|\tilde{\mathbf{y}}|$. Afterwards, the originally stored codebook of MJ binary entries is modified to the reduced one wit ML entries only. Finally, a match of the length L binary query versus the reduced database is performed targeting unique identification of $\hat{m} \in \{\epsilon, 1, 2, ..., M\}$.

It is easy to demonstrate using an information-theoretic argument that the maximum achievable rate of identification in the original setup with M length J codewords is given by the capacity of the BSC with the cross-over probability P_b that coincides with $R_{id|0}$ (7). Therefore, one can claim unique identification of $2^{JR_{id|0}} = 2^{J(1-H_2(P_b))}$ codewords that is achieved with complexity upper limited by $\mathcal{O}(MJ) = \mathcal{O}(J(1-H_2(P_b)))$ binary distance computations.

Oppositely, in the case only L good channels are used for identification, the maximum identification rate $R^G_{id|\tilde{y}}$ (17) is defined by the average probability of error in these channels $P^G_{b|\tilde{y}}$ (15). Such a result is achieved with the corresponding upper bound on the complexity that is defined as $\mathcal{O}(ML) = \mathcal{O}(L(1-H_2(P^G_{b|\tilde{y}})))$ matching operations. Therefore, the overall identification complexity depends on the $P^G_{b|\tilde{y}}$. It defines the number of bit flips that happens in a binary vector of length L at the output of the corresponding BSC via $\mathbb{B}^{-1}(1-\xi, L, P^G_{b|\tilde{y}})$, where $\mathbb{B}^{-1}(\xi, L, p)$ denotes the inverse cumulative distribution function of the Binomial distribution of the parameters (α, L, p) and ξ is a small positive constant. The number of errors in the entire codebook can be then evaluated as follow:

$$t_{error} = \mathbb{B}^{-1}(1-\xi, L \cdot M, P^G_{b|\tilde{y}}). \tag{21}$$

Thus, in the case t_{error} approaches zero, one can claim that the observed query exactly coincides with the database entry and in such a situation any distance computations are no required. For instance, if one would like obtain $t_{error} = 0$ with $\xi = 10^{-5}$ in database with about 7 millions entries, $P^G_{b|\tilde{y}} = 10^{-15}$ has to be satisfied in the reduced codebook with a codeword length $L = 618$. In order to investigate a feasibility of getting such low values of $P^G_{b|\tilde{y}} = 10^{-15}$, we analyzed a joint behavior of this probability of error versus $R_{id|\tilde{y}}$ as functions of the threshold for various SNRs that separates good and bad channels. First, one can admit that in case the system operates in a blind mode, i.e., no CSI is taken into account, the corresponding achievable rate is attained for the probability of bit error that is far apart from the sought order of 10^{-15} (Fig. 9). Therefore, no exact matching is possible in this regime. Alternatively, as one can easily observe from Fig. 9, the required reduction of the probability of error could be achieved in the imperfect CSI assisted setup in price of certain $R^G_{id|\tilde{y}}$ loss. Remarkably, for the case of $SNR = 25$ dB, the target ($P^G_{b|\tilde{y}} = 10^{-15}$) is attained for the $R^G_{id|\tilde{y}}$ reduction from 0.89 bits/sample to 0.6 bits/sample, i.e., approximately for a 30% loss of the identification rate.

Evidently, reduction of the probability of error is achieved by a more conservative good channel separation threshold. Therefore, in order to guarantee that the above analysis is valid and the number of good channels equals L, one should guarantee a sufficiently large J. This phenomenon is relevant to the enrolled database memory storage since according to the main assumption the entire collection of codewords is stored in the identification server RAM.

6. CONCLUSIONS

In this paper, we have presented the theoretical analysis of performance-complexity trade-off in the identification problem based on the channel splitting paradigm. Considering identification as communication thought the parallel channels with partially known channel state information, the effect of channel polarization, i.e., separation

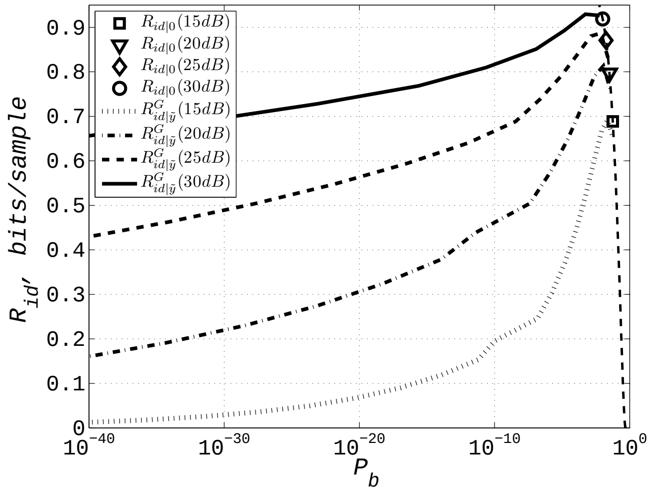


Figure 9: Identification operational characteristics.

of the channels to the subsets of channels with good (negligible small P_b) and bad (high P_b) characteristics, was investigated and used to build identification algorithms with reduced complexity. The above analysis considers identification as an joint optimization with respect to its accuracy (performance) and speed (computational complexity) with the fixed level of privacy leakage. Our future work should be conducted to incorporate privacy into this joint trading-off procedure.

REFERENCES

[1] Deng, J., Dong, W., Socher, R., jia Li, L., Li, K., and Fei-fei, L., "Imagenet: A large-scale hierarchical image database," in [*CVPR*], (2009).

[2] Hays, J. and Efros, A., "Scene completion using millions of photographs," *Commun. ACM* **51**(10), 87–94 (2008).

[3] Hua, G. and Tian, Q., "What can visual content analysis do for text based image search?," in [*Proceedings of the 2009 IEEE international conference on Multimedia and Expo*], *ICME'09*, 1480–1483, IEEE Press, Piscataway, NJ, USA (2009).

[4] Kalantidis, Y., Tolias, G., Spyrou, E., Mylonas, P., and Avrithis, Y., "Visual image retrieval and localization," in [*International Workshop on Content-Based Multimedia Indexing*], (2009).

[5] Haitsma, J., van der Veen, M., Kalker, T., and Bruekers, F., "Audio watermarking for monitoring and copy protection," in [*Proceedings of the ACM workshops on Multimedia*], 119–122, ACM, New York, NY, USA (2000).

[6] Lu, C.-S. and Hsu, C.-Y., "Geometric distortion-resilient image hashing scheme and its applications on copy detection and authentication," *Multimedia Systems* **11**, 159–173 (2005).

[7] Ignatenko, T. and Willems, F., "Biometric systems: Privacy and secrecy aspects," *Information Forensics and Security, IEEE Transactions on* **4**, 956 –973 (dec. 2009).

[8] Kalker, T., Ignatenko, T., Willems, F., Schmid, N., Vetro, A., Jain, A., Rane, S., and Wechsler, H., "Biometric security: An overview," in [*In IS&T/SPIE Electronic Imaging: Media Forensics and Security*], 17–21 (2010).

[9] Tuyls, P., Skoric, B., and Kevenaar, T., [*Security with noisy data: On Private Biometrics, Secure Key Storage and Anti-Counterfeiting*], Springer (2007).

[10] Halperin, E., Buhler, J., Karp, R., Krauthgamer, R., and B.Westover, "Detecting protein sequence conservation via metric embeddings," *BIOINFORMATICS* **1**(1), 1–8 (2003).

[11] Seo, J. and Lee, K.-J., "Post-translational modifications and their biological functions: Proteomic analysis and systematic approaches," *Biochemistry and Molecular Biology* **37**(1), 35–44 (2004).

[12] Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U., "When is "nearest neighbor" meaningful?," in [*In Int. Conf. on Database Theory*], 217–235 (1999).

[13] Böhm, C., B., S., and Keim, D. A., "Searching in high-dimensional spaces : Index structures for improving the performance of multimedia databases," *ACM computing surveys* **33**(3), 322–373 (2001).

[14] Friedman, J. H., Bentley, J. L., and Finkel, R. A., "An algorithm for finding best matches in logarithmic expected time," *ACM Trans. Math. Softw.* **3**(3), 209–226 (1977).

[15] Weber, R., Schek, H.-J., and Blott, S., "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in [*Proceedings of the 24rd International Conference on Very Large Data Bases*], 194–205, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1998).

[16] Datar, M. and Indyk, P., "Locality-sensitive hashing scheme based on p-stable distributions," in [*In SCG 04: Proceedings of the twentieth annual symposium on Computational geometry*], 253–262, ACM Press (2004).

[17] Gionis, A., Indyk, P., and Motwani, R., "Similarity search in high dimensions via hashing," 518–529 (1997).

[18] Muja, M. and Lowe, D. G., "Fast approximate nearest neighbors with automatic algorithm configuration," in [*In VISAPP International Conference on Computer Vision Theory and Applications*], 331–340 (2009).

[19] Shakhnarovich, G., Darrell, T., and Indyk, P., [*Security with noisy data: On Private Biometrics, Secure Key Storage and Anti-Counterfeiting, CH. 3*], MIT Press (2006).

[20] Ke, Y., Sukthankar, R., and Huston, L., "Efficient near-duplicate detection and sub-image retrieval," in [*In ACM Multimedia*], 869–876 (2004).

[21] Matei, B., Shan, Y., Sawhney, H. S., Tan, Y., Kumar, R., Huber, D., and Hebert, M., "Rapid object indexing using locality sensitive hashing and joint 3d-signature space estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**, 1111 – 1126 (July 2006).

[22] Li, Y., Chi, H., Wang, L.-H., and et al., H.-P. W., "Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing," *Rapid Communications in Mass Spectrometry : RCM* **24**, 807–814 (Feb. 2010).

[23] Haitsma, J. and Kalker, T., "Robust audio hashing for content identification," in [*In Content-Based Multimedia Indexing (CBMI*], 117–125 (2001).

[24] Seo, J. S., Haitsma, J., Kalker, T., and Yoo, C. D., "A robust image fingerprinting system using the radon transform," *Signal Processing: Image Communication* **19**, 325–339 (April 2004).

[25] Torralba, A., Fergus, R., and Weiss, Y., "Small codes and large image databases for recognition," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* **0**, 1–8, IEEE Computer Society, Los Alamitos, CA, USA (2008).

[26] Salakhutdinov, R. and Hinton, G., "Semantic hashing," *Int. J. Approx. Reasoning* **50**, 969–978 (July 2009).

[27] Weiss, Y., Torralba, A., and Fergus, R., "Spectral hashing," in [*Proceedings of "Advances in Neural Information Processing Systems"*], (2008).

[28] Nister, D. and Stewenius, H., "Scalable recognition with a vocabulary tree," in [*In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*], 2161 – 2168 (2006).

[29] Silpa-Anan, C. and Hartley, R., "Optimised kd-trees for fast image descriptor matching," in [*In IEEE Conference on Computer Vision and Pattern Recognition*], 1–8 (2008).

[30] Voloshynovskiy, S., Beekhof, F., Koval, O., and Holotyak, T., "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*], (2009).

[31] Brinkman, R., *Searching in encrypted data*, PhD thesis, University of Twente, University of Twente, Twente, The Netherlands (2007).

[32] Iyver, B., HacGumus, H., and S., S. M., "Efficient execution of aggregation queries over encrypted relational databases," in [*In the International Conference on Database Systems for Advanced Applications*], 633–650 (2004).

[33] Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y., "Order preserving encryption for numeric data," in [*SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*], 563–574 (2004).

[34] Boneh, D., Crescenzo, G. D., Ostrovsky, R., and Persiano, G., "Public key encryption with keyword search," in [*Advances in Cryptology - EUROCRYPT 2004*], Cachin, C. and Camenisch, J., eds., *Lecture Notes in Computer Science* **3027**, 506–522, Springer Berlin / Heidelberg (2004).

[35] Song, D. X., Wagner, D., David, S., and Perrig, A., "Practical techniques for searches on encrypted data," in [*IEEE Symposium on Security and Privacy*], 44–55 (2000).

[36] Kushilevitz, E. and Ostrovsky, R., "Replication is not needed: single database, computationally-private information retrieval," in [*Annual Symposium on Foundations of Computer Science,*], 364 –373 (oct. 1997).

[37] Chor, B. and Gilboa, N., "Computationally private information retrieval (extended abstract)," in [*STOC '97: Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*], 304–313, ACM, New York, NY, USA (1997).

[38] Chor, B., Kushilevitz, E., Goldreich, O., and Sudan, M., "Private information retrieval," *J. ACM* **45**(6), 965–981 (1998).

[39] Domingo-Ferrer, J., "A new privacy homomorphism and applications," *Information Processing Letters* **60**, 277–282 (1996).

[40] Do, T.-T., Kijak, E., Furon, T., and Amsaleg, L., "Challenging the security of content based image retrieval systems," in [*IEEE International Workshop on Multimedia Signal Processing*], (2010).

[41] Torralba, A., Fergus, R., and Freeman, W. T., "80 million tiny images: a large dataset for non-parametric object and scene recognition," *Information Processing Letters* **30**(11), 1958–1970 (2008).

[42] Willems, F., Kalker, T., Goseling, J., and Linnartz, J.-P., "On the capacity of a biometrical identification system," in [*Information Theory, 2003. Proceedings. IEEE International Symposium on*], 82 (Jun 2003).

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

[44] Varna, A. L., Swaminathan, A., and Wu, M., "A decision theoretic framework for analyzing binary hash-based content identification systems," in [*DRM '08: Proceedings of the 8th ACM workshop on Digital rights management*], 67–76, ACM, New York, NY, USA (2008).

[45] Willems, F., "Searching methods for biometric identification systems: Fundamental limits," in [*IEEE International Symposium on Information Theory*], 2241 –2245 (jun. 2009).

[46] Westover, M. and O'Sullivan, J., "Achievable rates for pattern recognition," *IEEE Transactions on Information Theory* **54**, 299 –320 (Jan. 2008).

[47] Tuncel, E., "Capacity/storage tradeoff in high-dimensional identification systems," *Information Theory, IEEE Transactions on* **55**, 2097 –2106 (may. 2009).

[48] Holotyak, T., Voloshynovskiy, S., Beekhof, F., and Koval, O., "Fast identification of highly distorted images," in [*Proceedings of SPIE Photonics West, Electronic Imaging 2010 / Media Forensics and Security XII*], (January 21–24 2010).

[49] Voloshynovskiy, S., Koval, O., Beekhof, F., Farhadzadeh, F., and Holotyak, T., "Information-theoretical analysis of private content identification," in [*IEEE Information Theory Workshop, ITW2010*], (Aug.30-Sep.3 2010).

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

[51] Raginsky, M. and Lazebnik, S., "Locality-sensitive binary codes from shift-invariant kernels," in [*Twenty-Third Annual Conference on Neural Information Processing Systems*], 1509–1517 (2009).

[52] Daugman, J., "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons," *Proceedings of the IEEE* **94**(11), 1927–1935 (2006).