Privacy Preserving Identification: Order Statistics List Decoding Perspective

Farzad Farhadzadeh, Sviatoslav Voloshynovskiy, and Oleksiy Koval
Computer Science Department
University of Geneva
Geneva, Switzerland
$\{$*Farzad.Farhadzadeh, svolos, Oleksiy.Koval*$\}$*@unige.ch*

***Abstract*—In this paper, the performance of privacy preserving Biometric/Physical Unclonable Function identification based on the order statistic list decoding is analyzed by evaluating the corresponding probabilities of correct identification and false acceptance. We demonstrate the impact of candidate list size on the identification system performance and compare it with the unique identification setup. Finally, the influence of privacy amplification on the system performance is analyzed.**

***Keywords*-biometric identification; privacy amplification; list decoding; order statistics; correct identification; false acceptance; receiver operating characteristic.**

I. Introduction

The present work addresses a privacy-preserving identification based on biometrics or Physical Unclonable Functions (PUFs). Both biometrics and PUFs are well-known in forensic applications [1] because of their ability to serve as a unique identifier for people and physical objects.

Design of biometric/PUFs identification system is a complex optimization problem that is subject to a set of conflicting constraints. Leaving some of them, i.e., security and complexity, outside of the scope of this paper, we present a solution to a performance-privacy trade-off in such a design.

Because of channel distortions, due to acquisition imperfection, compression *etc.*, the identification system should be able to cope with data degradations. In classical identification setups, the decoder estimates a unique index for a given query. Another approach, which can be considered as a generalization of the one mentioned above, firstly proposed by Elias [2] in information theory, is known as *list decoding*. The main feature of this type of decoding is to produce a fixed list size of the most likely candidates rather than a single one. The result of [2] was generalized by Forney to a variable list size [3]. In many identification problems, the final sink of information is a human being. This restriction makes variable list size decoding undesirable, due to the fact that in intensive degradation conditions the cardinality of candidates increaces drastically.

A special care should be addressed to the organization of the system codebook/database. Specifically, even gaining the physical access to the storage module should not provide an unauthorized party with a privacy relevant information ready for a malicious misuse. The privacy of the entire protocol is usually ensured by following the principles of cryptographic privacy amplification (PA) [7]. Finally, an additional randomization of the database content might be proposed targeting further privacy strengthening.

Accordingly to such a strategy, the main contribution of this paper can be summarized as follows. We introduce the identification setup based on an *order statistic list decoder* (OSLD) of a fixed maximum list size and analyze its performance versus unique decoding. For reasons of computational complexity, privacy and security, it is undesirable for an identification system to retain the biometrics or PUFs in their original form. We use random projections and binarization for the digital fingerprint generation and channel statistics conversion to a binary model [5]. The applied mechanism of binary data generation additionally provides minimization of the privacy leakage to unauthorized parties. For additional privacy enforcement, a complementary randomization mechanism is exploited [7]. Finally, we analyze the OSLD probabilities of correct identification and false acceptance over the binary symmetric channel (BSC) for various privacy constraints.

Notations: We use capital letters X to denote scalar random variables and $\mathbf{X}$ to denote vector random variables. Corresponding small letters x and $\mathbf{x}$ denote the realizations of scalar and vector random variables. All vectors without sign tilde are assumed to be of the length N and with the sign tilde of length L. $\mathcal{B}(N, p)$ denotes the Binomial distribution with N trials and probability of success p. $V_{(r:M)}$ stands for the r-th order statistics of M i.i.d. random variables.

II. Identification Setup

The identification setup under analysis shown in Fig. 1 consists of two main phases: *enrollment* and *identification*. In the enrollment phase, the digital fingerprints of biometrics or PUFs of items to be identified $\mathbf{x}(m) \in \mathcal{X}^N, m = 1, \ldots, M$, which are drawn independently from a common stationary distribution $p_{\mathbf{X}}(\mathbf{x})$, are stored in the *Database*. The digital fingerprints, which are short, robust and discriminative representative of items, are extracted in two steps. At the first step; by applying random projections [5], which are so-called *orthoprojectors* $\mathbf{W} \in \mathbb{R}^{L\times N}$ with $W_{ij} \sim \mathcal{N}(0, \frac{1}{N})$, $1 \leq i \leq L$ and $1 \leq j \leq N$, the dimensionality is reduced from N to L. At the second

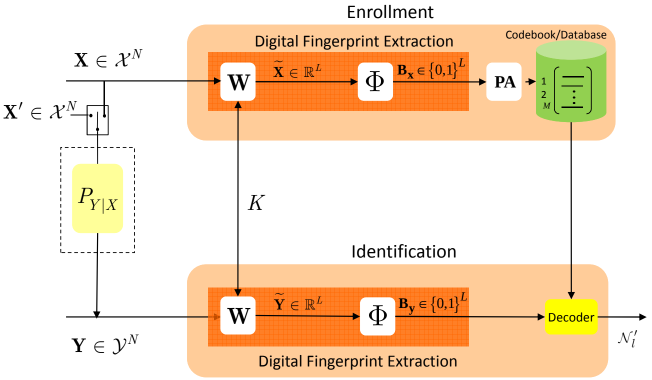
Enrollment
Digital Fingerprint Extraction
Codebook/Database
$\mathbf{X} \in \mathcal{X}^N$
$\mathbf{X}' \in \mathcal{X}^N$
W
$\tilde{\mathbf{X}} \in \mathbb{R}^L$
Φ
$\mathbf{B_x} \in \{0,1\}^L$
PA
1
2
M
$P_{Y|X}$
K
Identification
W
$\tilde{\mathbf{Y}} \in \mathbb{R}^L$
Φ
$\mathbf{B_y} \in \{0,1\}^L$
Decoder
$\mathcal{N}_l'$
$\mathbf{Y} \in \mathcal{Y}^N$
Digital Fingerprint Extraction

Figure 1. Identification problem based on binary templates.

step, L-length binary data are derived from the projected data by taking the sign of the lower dimensional data, i.e., $B_{\tilde{\mathbf{x}}} = \Phi(\tilde{\mathbf{X}})$, where $\tilde{\mathbf{X}} = \mathbf{WX}$.

In the identification phase; for a given query $\mathbf{Y}$, which can be a noisy version of the biometrics or PUFs or originated from the data irrelevant to the stored in the database, the digital fingerprint is extracted following the same stages as in the enrollment phase, i.e., $B_{\tilde{\mathbf{y}}} = \Phi(\tilde{\mathbf{Y}})$, where $\tilde{\mathbf{Y}} = \mathbf{WY}$. Afterwards, the decoder detects that the query is related to some entries of the database, and if so, identifies to which ones.

The decoding process in the identification setup is accomplished in two steps. At the first step, the primary candidates are chosen by the OSLD. At the second step, a threshold is applied to all candidates, and the candidates which satisfy the constraint remain on the list. The decoding procedure can be summarized as follows[1]:

1) The likelihood functions, $p(\mathbf{y}|\mathbf{x}(m)), 1 \leq m \leq M$, for all database entries are computed.
2) The computed likelihood functions are sorted in ascending order.
3) The indices of the N_l largest likelihood functions are chosen which form the primary list $\mathcal{N}_l$. The parameter $|\mathcal{N}_l| = N_l$ is referred as a primary list size.
4) The final output set of the decoder is defined as $\mathcal{N}_l' = \{m \in \mathcal{N}_l : p(\mathbf{y}|\mathbf{x}(m)) \geq e^{N\gamma}\}$ where the parameter γ controls the number of final candidates.

The performance of the decoder is evaluated by the probability of correct identification, i.e., $P_c = \Pr\{m \in \mathcal{N}_l' | \mathcal{H}_m\}$, where $\mathcal{H}_m$ corresponds to $\mathbf{Y}$ related to the m^{th} entry of the database, and the probability of false acceptance, i.e., $P_f = \Pr\{\mathcal{N}_l' \neq \varnothing | \mathcal{H}_0\}$, where $\mathcal{H}_0$ corresponds to $\mathbf{Y}$ unrelated to any database entry.

III. Order Statitics

Before considering error events, we will shortly review *Order Statistics*, which will be used in the computation of the probability of errors. We suppose that $V(1), V(2), \ldots, V(M)$ are M i.i.d. random variables, each with a cumulative distribution function (CDF) $F(v)$. Let $F_{(r:M)}(v)$ denote the CDF of the r-th order statistics $V_{(r:M)}$, which corresponds to the r-th position of $v_{(1:M)} \leq \ldots \leq v_{(r:M)} \leq \ldots \leq v_{(M:M)}$ for a specific outcome, Then [4]:

$$\begin{aligned} F_{(r:M)}(v) =& \Pr\left\{V_{(r:M)} \leq v\right\} \\ =& \Pr\Big\{\text{at least } r \text{ of } V(1), V(2), \ldots, V(M), \\ & \quad \text{are less than or equal to } v\Big\} \\ =& \sum_{i=r}^{M} \binom{M}{i} F^i(v)[1 - F(v)]^{M-i} \qquad (1) \end{aligned}$$

since the term in the summand is the binomial probability that *exactly* i of $V(1), V(2), \ldots, V(M)$ are less than or equal to v.

IV. Probability of Correct Identification

Once the primary list of candidates is constructed by the OSLD, the final candidates are extracted by thresholding their likelihoods. A correct identification event occurs when the index of the database entry related to the query is on the final list. The probability of correct identification, P_c, is given by:

$$\begin{aligned} P_c =& \sum_{m=1}^{M} \Pr\{(m \in \mathcal{N}_l) \cap (p(\mathbf{y}|\mathbf{x}(m)) \geq e^{N\gamma}) | \mathcal{H}_m\} \\ & \times \Pr\{\mathcal{H}_m\} \qquad (2) \end{aligned}$$

where $\mathcal{N}_l$ is the primary list of candidate indeces. As the entries of the database are identically distributed and equally likely to be queried, the overall probability of correct identification does not depend on the particular index and hence:

$$P_c = \Pr\{(1 \in \mathcal{N}_l) \cap (p(\mathbf{y}|\mathbf{x}(1)) \geq e^{N\gamma}) | \mathcal{H}_1\}. \qquad (3)$$

After dimensionality reduction, binarization and PA, we have the binary data of length L, where $L < N$. In the binary domain, the link between $\mathbf{b_x}$ and $\mathbf{b_y}$ and between $\mathbf{b_x}$ and $\mathbf{b_u}$ can be considered based on the BSC models with corresponding crossover probability P_b and λ, respectively [8]. The parameter λ corresponds to the BSC serving as a test channel for the compressed version $\mathbf{b_u}$ considered as the PA [7]. Under the above assumption, these two BSCs $\mathbf{b_x} \rightarrow \mathbf{b_y}$ and $\mathbf{b_x} \rightarrow \mathbf{b_u}$ 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, for any $\mathbf{b_u}(m), \mathbf{b_y} \in \{0,1\}^L$, the likelihood function

$$p(\mathbf{b_y}|\mathbf{b_u}(m)) = P_{b_e}^{d_H(\mathbf{b_y}, \mathbf{b_u}(m))} (1 - P_{b_e})^{L - d_H(\mathbf{b_y}, \mathbf{b_u}(m))} \qquad (4)$$

[1] The low-complexity identification of OSLD decoding based on the concept of bit reliability is given in [6].

is a decreasing function of the Hamming distance $d(m) \triangleq d_H(\mathbf{b_y}, \mathbf{b_u}(m))$ for $0 \leq P_{b_e} \leq 0.5$. In the following, we will consider the above Hamming distance as a realization of the random variable $D(m)$ where m refers to the index of $\mathbf{b_u}(m)$. Given a query related to the m^{th} entry of the database, the event $m \in \mathcal{N}_l$ occurs, if $d(m)$ is among the N_l smallest distances $\{d(1), d(2), \dots, d(M)\}$.

The dimensionality reduction and binarization modify the statistics of the database generated from $p_{\mathbf{X}}(\mathbf{x})$ to the Binomial distribution, i.e., $\mathbf{B_u} \sim \mathcal{B}(L, 1/2)$ for any $p_{\mathbf{X}}(\mathbf{x})$ with a diagonal covariance matrix.

Conditioned on $\mathcal{H}_1$, the sufficient statistics can be expressed as follows:

$$D(m) \sim \begin{cases} \mathcal{B}(L, P_{b_e}), & \text{for } m = 1, \\ \mathcal{B}(L, \frac{1}{2}), & \text{for } m \neq 1. \end{cases} \tag{5}$$

From (1), the complement of the CDF, $F^c_{(N_l:M-1)}(d)$, of N_l^{th} order statistics of the i.i.d random variables $D(m),\ m \neq 1$ is given by:

$$\begin{aligned} F^c_{(N_l:M-1)}(d) =& \Pr\{D_{(N_l:M-1)} > d\} \\ =& \sum_{p=0}^{N_l-1} \binom{M-1}{p} s(d)^p (1-s(d))^{(M-1)-p}, \end{aligned} \tag{6}$$

where $s(d) \triangleq \sum_{x=0}^{d} \binom{L}{x} \left(\frac{1}{2}\right)^L$. From (5) and (6), the probability of correct identification (3) over the BSC can be expressed by:

$$\begin{aligned} P_c &\overset{(a)}{=} \Pr\{(D_{(N_l:M-1)} > D(1)) \cap (D(1) \leq \eta) | \mathcal{H}_1\} \\ &= \sum_{d=0}^{\eta} \Pr\{D_{(N_l:M-1)} > d | \mathcal{H}_1, D(1) = d\} p_{D(1)}(d) \\ &= \left\{ \sum_{d=0}^{\eta} \binom{L}{d} P_{b_e}^d (1-P_{b_e})^{L-d} \right. \\ &\quad \left. \times \sum_{p=0}^{N_l-1} \binom{M-1}{p} s(d)^p (1-s(d))^{(M-1)-p} \right\} \end{aligned} \tag{7}$$

where from (3) and (4) $\eta = L\frac{\gamma - \ln(1-P_{b_e})}{\ln(P_{b_e}/(1-P_{b_e}))}$, (a) follows from (3) and the fact that the likelihood function is a decreasing function of the Hamming distance, and $p_{D(1)}(d)$ denotes the PMF of $D(1)$.

V. Probability of False Acceptance

The main reason to consider the probability of false acceptance is to show the reliability of the decoding process with respect to various attacking strategies. There are different scenarios to investigate the reliability of the decoder in identification setups:

- The PDF of database generation is known.
- The PDF of database generation is not known.
- The database entries are partially known.
- The database entries are totally known.

In this paper, we consider the scenario in which the PDF is fully known by the attacker. In the following, the false acceptance event is analyzed over the BSC according to the justification presented above.

To evaluate the probability of false acceptance, we define the following events:

$$E_{D_{(i:M)}} = \{D_{(i:M)} \leq \eta | \mathcal{H}_0\}, \tag{8}$$

where $1 \leq i \leq N_l$, $D(m) \sim \mathcal{B}(L, \frac{1}{2}), 1 \leq m \leq M$, and $E_{D_{(i:M)}}$ is the event that the i^{th} ascending ranked Hamming distance between the query and an entry of database is smaller than the threshold. From (1), the probability of false acceptance is found as:

$$\begin{aligned} P_f =& \Pr\Big\{ \bigcup_{i=1}^{N_l} E_{D_{(i:M)}} | \mathcal{H}_0 \Big\} = 1 - \Pr\Big\{ \bigcap_{i=1}^{N_l} E^c_{D_{(i:M)}} | \mathcal{H}_0 \Big\} \\ &\overset{(a)}{=} 1 - \Pr\big\{ E^c_{D_{(1:M)}} | \mathcal{H}_0 \big\} = \Pr\big\{ E_{D_{(1:M)}} | \mathcal{H}_0 \big\} \end{aligned} \tag{9}$$

where $E^c_{D_{(i:M)}}$ is the complement of $E_{D_{(i:M)}}$ and (a) follows from the fact that as the event $E^c_{D_{(1:M)}}$ occurs the rest events occur. From (1), the probability of false acceptance can be computed by:

$$\begin{aligned} P_f =& \Pr\big\{ \min_{1 \leq m \leq M} D(m) \leq \eta | \mathcal{H}_0 \big\} \\ =& 1 - \left[1 - \sum_{x=0}^{\eta} \binom{L}{x} \left(\frac{1}{2}\right)^L \right]^M . \end{aligned} \tag{10}$$

The interesting point to note is that the probability of false acceptance is unexpectedly independent of the primary list size.

VI. Simulation Results

Without loss of generality, as it is mentioned before the identification is performed in the space of binary templates. Therefore, we will assume that the database entries can be represented by the Binomial distribution. Thus, the performance of the proposed decoder is evaluated by using databases of synthetic data with different sizes that are independently and identically distributed according to $\mathbf{B_x} \sim \mathcal{B}(L, \frac{1}{2})$ that is as well optimal from a privacy point of view. The input to the identification system is obtained at the output of the BSC with a crossover probability $P_{b_e} = P_b * \lambda$ and $P_b = \frac{1}{\pi} \arccos \rho_{\tilde{X}\tilde{Y}}$ [6], where $\rho_{\tilde{X}\tilde{Y}}$ is a cross-correlation coefficient between $\tilde{X}$ and $\tilde{Y}$. Fig. 2 confirms the accuracy of the derived model for the P_c and P_f.

From (7) and (10), the probability of correct identification and false acceptance over the BSC are computed and *receiver operating characteristic* (ROC) curves have been shown for different SNRs, database sizes M and primary list sizes N_l (Fig. 3 and Fig. 4).

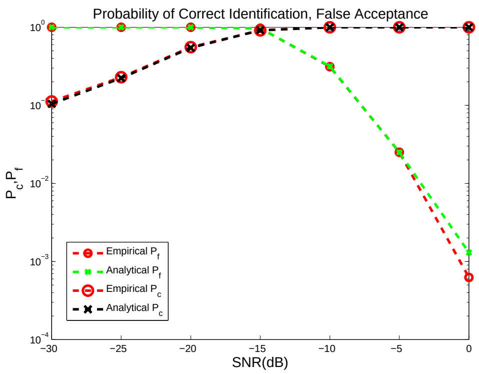

Figure 2. The probability of correct identification and false acceptance of OSLD with $N_l = 8$ over the BSC, where $L = 1024$, $M = 256$ and $\lambda = 0$.

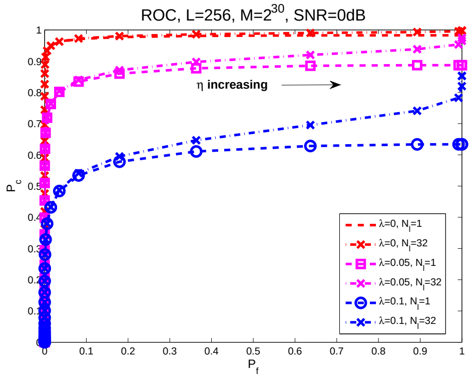

Figure 3. The OSLD performance improvement for the database with binary entries.

Fig. 3 presents the performance of the OSLD and unique decoder $(N_l = 1)$ under various constraints (λ) on PA. It is confirmed that the privacy improvement comes for the price of a certain performance loss in terms of P_c as a function of λ. Secondly, the obtained results allow to conclude that the OSLD improves the performance while the probability of correct identification P_c is not close to one.

Fig. 4 shows the impact of N_l and M on the identifier performance. Increasing primary list size N_l improves the identifier performance but after a certain value of N_l it does not change significantly. Finally, increasing database size M decreases the performance.

VII. Conclusions

In this paper, we investigated the performance of the maximum fixed list size OSLD-based biometric/PUF privacy preserving identification system in terms of the probability of correct identification and false acceptance.

The obtained theoretical and simulation results show that on the one hand, the OSLD can improve the identifier performance in very low SNR scenarios. On the other hand, this improvement is restricted by a certain range of list sizes. Finally, the loss in performance in terms of P_c due to PA is demonstrated for the current system implementation. We are planning to consider the ways of improving the performance/privacy trade-of in our future research.

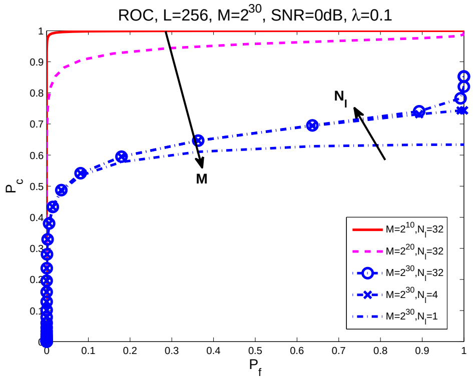

Figure 4. The effect of primary list size and database size over the OSLD performance.

Acknowledgments

The authors would like to thank SIP group members F. Beekhof and T. Holotyak for stimulating discussions. This paper was partially supported by SNF project 200021-119770.

References

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

[2] P. Elias, *List decodeing for noisy channels*, Tech. Rept. 335, Research Labratoary of Electronics, M.I.T, 1955.

[3] G. D. Forney, Jr, *Exponential error bounds for erasure, list and decision feedback schemes*, IEEETrans Inf. Theory, vol. IT-14, no. 2, pp. 206-220, Mar. 1968.

[4] H. A. David, H. N. Nagaraja, *Order Statistics*, 3rd ed, pp. 9, Wiley-Interscience, 2003.

[5] S. Voloshynovskiy, O. Koval, F. Beekhof, and T. Pun, *Conception and limits of robust perceptual hashing: toward side information assisted hash functions*, SPIE Photonics West, San Jose, USA, 2009.

[6] F. Beekhof, S. Voloshynovskiy, O. Koval, and T. Holotyak, *Fast Identification Algorithms for Forensic Applications*, IEEE International Workshop on Information Forensics and Security, London, 2009.

[7] C. H. Bennett, G. Brassard, C. Crepeau, and U. Maurer, *Generalized privacy amplification*, IEEE Transactions on Information Theory, vol. 41, no. 6, pp. 19151923, Nov. 1995.

[8] 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.* International Workshop on Signal Processing in the EncryptEd Domain, Lausanne, Switzerland, 2009.