EXPERIMENTATION OF "ADVANCED INFTHEO" MODULE FOR "R" ON THE EXAMPLE OF BIOMETRIC GENERATED SECRET KEY SHARING SYSTEM

Mariam Haroutunian, Narek Pahlevanyan

Abstract: *Many information-theoretical results in practice are difficult to compute, because of the large volume of distributions. To perform computations of complex formulas of Information Theory authors have developed new module (Advanced Inftheo) for R language. For high performance the main functions of module use parallel algorithms and tools of C++ to dynamically and optimally allocate memory. In this paper we demonstrate some results of computations that are done by Advanced Inftheo module. As an example we compute and represent the lower and upper bounds of E-achievable secret key rate of the biometric generated secret key sharing system obtained in [Haroutunian-Pahlevanyan, 2014]. E-achievable secret key rate is the generalization of the secret key rate, studied by [Ignatenko-Willems, 2012].*

Keywords: *E-achievable secret key rate, R language, parallel computations*

ACM Classification Keywords: *H.0 Information Systems - Conference proceedings*

Introduction

The usage of biometric secrecy systems in modern society is growing very quickly. Biometric secrecy systems are based on the person's physiological or behavioral characteristics. Physiological characteristics include fingerprints, hand geometry, facial, voice, iris, retinal features etc. [Chen-Vinck, 2011]. Behavioral characteristics include the dynamics of signatures, keystrokes etc. Biometric secrecy systems capture and process person's unique characteristics, and then authenticate that person's identity based on comparison of the record of captured characteristics with a biometric sample presented by the person to be authenticated. Biometric characteristics cannot be lost or forgotten, they are difficult to copy, share and distribute. Therefore, the advantage of biometric secrecy systems against the traditional password based security systems is evident.

Biometric secrecy systems can be used in various applications, such as in authentication, identification, examinations, payment processing, secure travel documents, visas et al. In many applications, such as for example, examinations the person is required to be present at the time and point of authentication. Moreover, there are access scenarios, which require participation of multiple previously registered users for a successful authentication or to get an access grant for a certain entity. For instance there are

cryptographic constructions known as secret sharing schemes, where a secret key is split into shares and distributed amongst users in such a way that it can be reconstructed only when the necessary number of the secret key holders comes together. The revealed secret can then be used for encryption or authentication. One of such applications could be sharing of a bank account by family members.

As mentioned in [Ignatenko-Willems, 2012], biometric secrecy systems are grouped around two classes: cancelable biometrics and fuzzy encryption. In fuzzy encryption class systems a secret key is generated/chosen during an enrollment procedure, in which the biometric data are observed for the first time. The secret key is to be reconstructed after these biometric data are observed again, during an attempt to get access. Reliable biometric secret key sharing systems extract helper data from the biometric information at the time of enrollment. These helper data contributes to reliable reconstruction of the secret key.

However, the usage of biometric secrecy systems has its own disadvantages. Since biometric data are gathered from individuals under environmental conditions and the channels are exposed to noise the biometric secrecy system may accept an impostor or reject an authorized individual. It's not possible to build ideal biometric secrecy system, it can be information-theoretical secure up to a certain level. From information-theoretical point of view biometric secrecy systems were studied by O'Sullivan and Schmid [O'Sullivan - Schmid, 2002], Willems et al [Willems et al, 2003; Haroutunian-Pahlevanyan, 2014]. Willems [Willems et al, 2003] investigated the fundamental properties of biometric identification system. It has been shown that it is impossible to reliably identify more persons than capacity which is an inherent characteristic of any identification system. By analogy with notion of E-capacity or rate-reliability function introduced by E. Haroutunian [Haroutunian, 2007; Haroutunian et al, 2008] in [Haroutunian-Pahlevanyan, 2014] we introduce the new concept of E-achievable secret key rate for biometric generated secret key sharing system. The authors derived the upper and the lower bounds for E-achievable secret key rate of biometric generated secret key sharing system. E-achievable secret key rate expresses the dependance of the main characteristics of the system.

The construction of biometric secrecy systems that are both reliable and secure is strongly connected with investigation of rate-reliability function. In practice the investigation of rate-reliability function is complex and computational results are complicated to obtain, mostly because of the large volume of distributions. There are many statistical software packages that can help in computations, the popular ones are SAS, STATA, SPSS, Matlab, Mathematica. Moreover, statistical libraries are available in most of programming languages, for instance Pandas in Python, Alglib in C++ and C#. In [Pahlevanyan, 2014] the author made a comparative analysis of the R language with other statistical packages and demonstrated several significant advantages of R. Furthermore, for data analysis, large companies such as Google, Facebook, and Twitter use R.

R is a language for statistical computing, data manipulation, data mining and graphics. Robert Gentleman and Ross Ihaka started development of R in 1993, but it became popular last years, particularly for data scientists, as it contains a number of built-in functions for organizing data, running calculations on the information and creating elegant graphical representations of data sets. R provides a lot of different techniques for statistical linear and nonlinear modeling, time-series analysis, classification, clustering as well as graphical packages for creating high quality, and sophisticated, customized plots with very simple syntax. The capabilities of R can be extended through user-created modules. Modules are libraries developed in C++ that include specific functions for usage in certain applications. A core set of packages included with the installation of R, with more than 5,800 additional packages and 120,000 functions are free available for download [Venables et al, 2014]. R already had an extension for calculating various measures of Information Theory, but there was a need in creation of new module for estimation and computation of more complex formulas mentioned above.

To perform computations of complex formulas of Information Theory authors have developed new module for R, called Advanced Inftheo. Module Advanced Inftheo was developed in C++, because in R there is no multi-threading support, and there are restrictions of memory management. For high performance the main functions of module use parallel algorithms and tools of C++ to dynamically and optimally allocate memory. Moreover, in parallel algorithms there are often problems associated with the usage of same system resources by parallel running threads, to overcome such problems the module uses thread interaction techniques known as semaphores and critical sections. The module provides functionality for computation of the lower and upper bounds of E-achievable secret key rate, as well as functionality for computation of mutual information, conditional mutual information, Kullback - Leibler (divergence) distance and other quantities of Information Theory. It has an option to connect with cluster (using the library MPI) and execute all computational functions on cluster.

In this paper we demonstrate some results of computations that are done by Advanced Inftheo module. As an example we compute the lower and upper bounds of E-achievable secret key rate of the biometric generated secret key sharing system for various distributions. We give graphical representations of the computations to simplify the solutions in building of applications.

Biometric generated secret key sharing model

Let's define some conventions that are applied within this paper. Capital letters are used for random variables (RV) X, Y taking values in the finite alphabets $\mathcal{X}, \mathcal{Y}$ correspondingly. The cardinality of the alphabet $\mathcal{X}$ is denoted by $|\mathcal{X}|$. The notation $|a|^+$ is used for $\max(a, 0)$. Biometric generated secret key sharing model is represented in Figure 1.

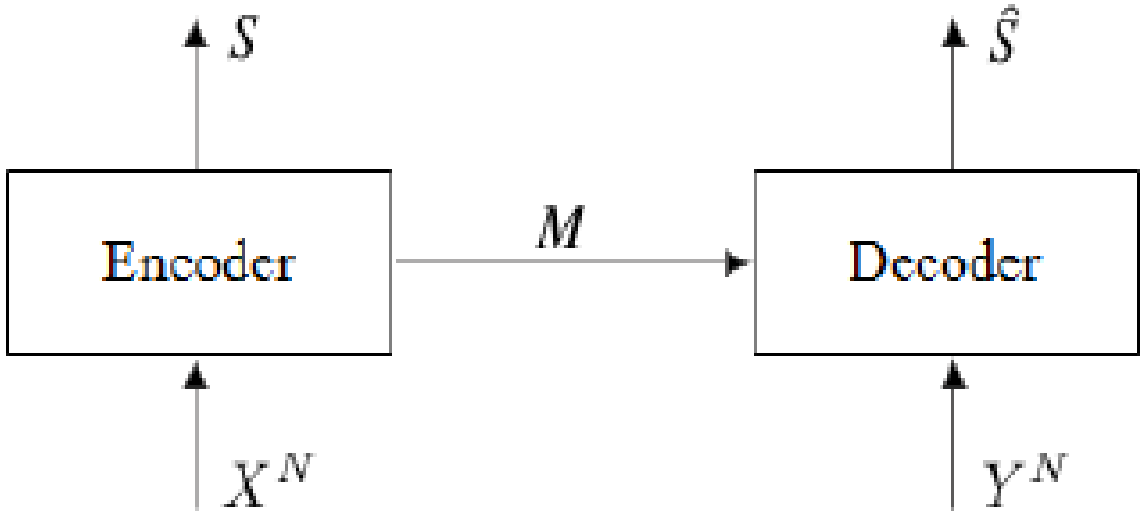

Figure 1. Biometric generated secret key sharing model

The model is based on biometric source with distribution $\{Q(x,y), x \in \mathcal{X}, y \in \mathcal{Y}\}$. This source produces

$\mathbf{x} \equiv x^N = (x_1, x_2, \dots, x_N)$ of N symbols from the finite alphabet $\mathcal{X}$ and a second sequence $\mathbf{y} \equiv y^N = (y_1, y_2, \dots, y_N)$ of N symbols from the finite alphabet $\mathcal{Y}$. The first sequence is called the enrollment sequence, and the second sequence the authentication sequence. Moreover, the second sequence Y^N is a noisy version of the first sequence X^N. Let us denote

$$Q(x,y) = Q_1(x)Q_2(y|x), x \in \mathcal{X}, y \in \mathcal{Y}.$$

We assume that

$$Q^N(\mathbf{x},\mathbf{y}) = \prod_{n=1}^{N} Q(x_n, y_n).$$

Then consider an encoder that explores enrolment sequence X^N. From this sequence in biometric generated secret key sharing model the encoder generates a secret $S \in \{1,2,\dots,|S|\}$ and then a public helper data $\in \{1,2,\dots,|M|\}$. That means that

$$f(X^N) = (S, M),$$

where by $f(\cdot)$ we denote the encoder function. The helper data is sent to decoder. The decoder explores the authentication sequence Y^N and produces an estimate $\hat{S}$ of the secret S using the received helper data M, hence

$$g(Y^N, M) = \hat{S},$$

where by $g(\cdot,\cdot)$ we denote the decoder function. The channel between encoder and decoder is expected to be public. We assume that attacker has an access to that channel, so he can see all the public information but cannot modify it. The information outflow is described in terms of mutual information, and the size of the secret key in terms of entropy. Fingerprints and irises can be modeled as such biometric sources.

The important quantitative measures of a biometric secrecy system are reliability E, error probability, secret key rate, size of secret key and the information that the helper data leak on the biometric observation. That leak of biometric information is called privacy leakage. The privacy leakage should be small, to avoid the biometric data of an individual to become compromised. Moreover, the secret key length should be large to minimize the probability that the secret key is guessed. The goal of encoder and decoder is to produce a secret key as large as possible, that satisfies to condition $Pr\{S \neq \hat{S}\} \approx 0$, this means that probability that the estimated secret $\hat{S}$ is not equal to generated secret S is close to zero.

The definition for achievable secret key rate can be found in [Ignatenko-Willems, 2012]. We investigate the exponentially high reliability criterion in biometric generated secret key sharing systems. The new performance concept introduced in [Haroutunian-Pahlevanyan, 2014] for biometric generated secret key sharing system, takes into account a stronger requirement on authentication fault events with extremely small probability. In terms of practical applications an exponential decrease in error probability (namely, authentication fault events) is more desirable. Here is the definition of E-achievable secret key rate [Haroutunian-Pahlevanyan, 2014].

Definition. A secret key rate $R(E)$, for $R(E) \geq 0$, is called E-achievable if for all $\delta > 0, E > 0$ and N large enough, there exists a code such that

$$Pr\{S \neq \hat{S}\} \leq 2^{-N(E-\delta)},$$

$$\frac{1}{N}H(S) + \delta \geq \frac{1}{N}\log_2|S| \geq R(E) - \delta,$$

$$\frac{1}{N}I(S \wedge M) \leq \delta.$$

We shall use the following PD in the formulation of result:

$$Q_1 = \{Q_1(x), x \in \mathcal{X}\}, Q_2 = \{Q_2(y|x), y \in \mathcal{Y}, x \in \mathcal{X}\},$$

$$P_1 = \{P_1(x), x \in \mathcal{X}\}, P_2 = \{P_2(y|x), y \in \mathcal{Y}, x \in \mathcal{X}\},$$

$$Q = \{Q(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\},$$

$$P = \{P(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

We refer to [Haroutunian, 2007; Haroutunian et al, 2008; Csiszar, 1998] and [Cover-Thomas, 2006] for notions of divergence $D(P||Q)$, mutual information $I_P(X \wedge Y)$, information-theoretic quantities.

The main result found in [Haroutunian-Pahlevanyan, 2014] is the theorem.

Theorem. For biometric generated secret key sharing model the largest E-achievable secret key rate $R(E)$ is lower bounded by

$$R_r(E) = \min_{P:D(P||Q)\leq E} |I_P(X \wedge Y) + D(P||Q) - E|^+$$

And upper bounded by:

$$R_{sp}(E) = \min_{P:D(P||Q)\leq E} I_P(X \wedge Y).$$

The proof of theorem can be found in [Haroutunian-Pahlevanyan, 2014].

Corollary. When $E \to 0$, the limits of lower and upper bounds coincide and equal the largest achievable secret key rate defined in [Ignatenko-Willems, 2012]

$$\lim_{E\to 0} R_r(E) = \lim_{E\to 0} R_{sp}(E) = I_Q(X \wedge Y).$$

Graphical representations of computations

We have performed computation of above formulas in the theorem using Advanced Inftheo module. In this section we present some graphical representations of results. Let's consider binary symmetric channel and denote d as a parameter. Let $Q(y|x)$ conditional distribution matrix be defined as

$$Q(1|1) = Q(0|0) = 1 - d,$$
$$Q(1|0) = Q(0|1) = d.$$

From the Figures 2, 3 and 4 of lower and upper bounds of E-achievable secret key rate we obtain the dependence of achievable secret key rate from reliability E for various d.

The computations have been done on machine with medium parameters (Intel Core 2 Duo 2 x 2.00GHz, with 2.5GB RAM). In worst case when only single thread is used inside Advanced Inftheo module to perform calculations the computation time, for instance when $d = 0.2$ and E changes from 0 to 1.2 with step 0.001, would be around 8.36 seconds. The above computation for same instance (when $d = 0.2, E = \overrightarrow{0; 1.2}$ with step 0.001) took 3.25 seconds on multithreaded environment of Advanced Inftheo module. As we can see, we have around 61% time gain. If we increase the precision of computations the time gain will be significant. It's worth mentioning that threads inside Advanced Inftheo are being allocated based on pc's technical capabilities and on range and step of E. A lot more time gain can be achieved if computations would be done on cluster and then forwarded to the user. The

considered dependence of achievable secret key rate from reliability E will help to design practical biometric generated secret sharing systems.

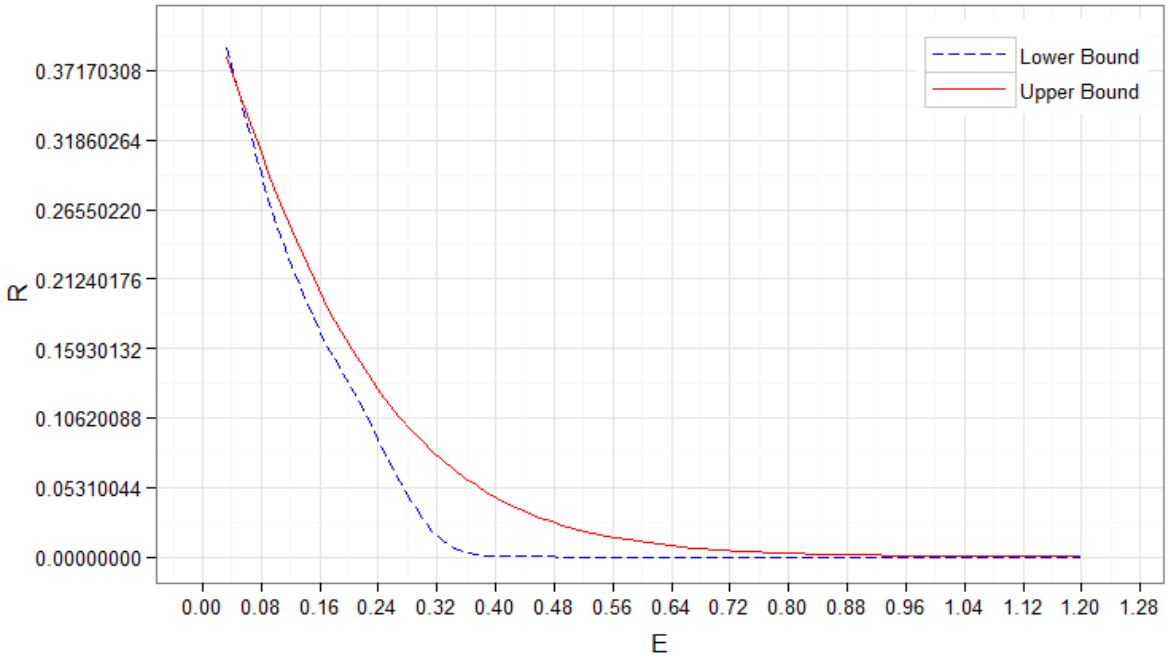

Figure 2. Bounds of of E-achievable secret key rate, when $d = 0.1$

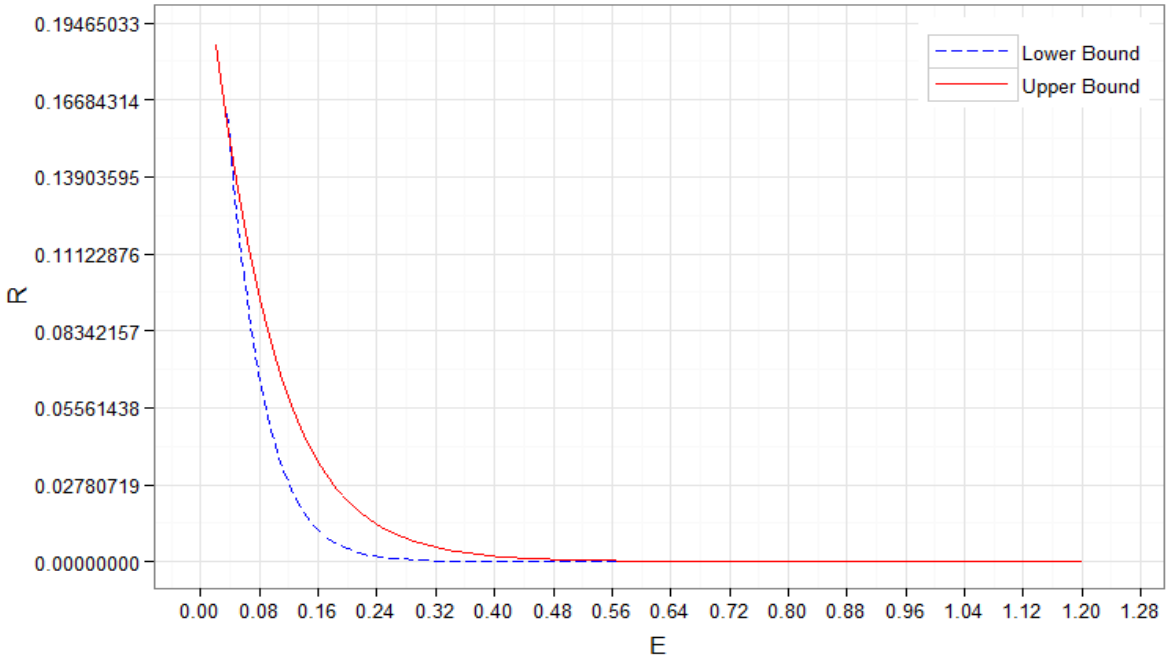

Figure 3. Bounds of of E-achievable secret key rate, when $d = 0.2$

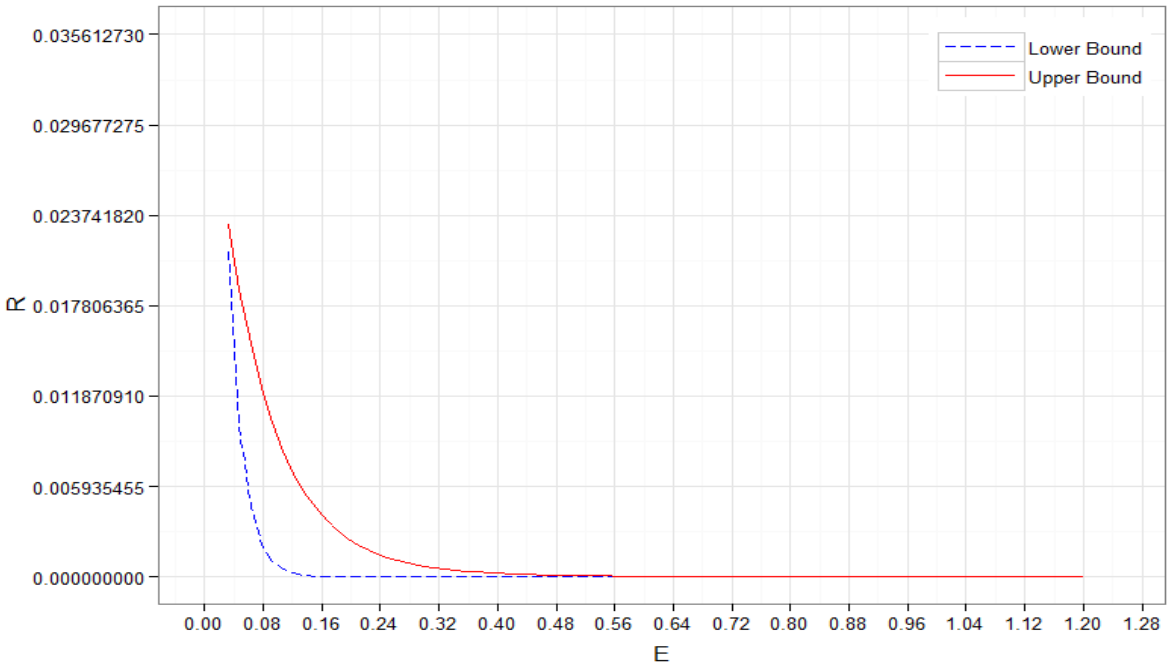

Figure 4. Bounds of of E-achievable secret key rate, when $d = 0.3$

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Chen-Vinck, 2011] Y. Chen and A. H. Vinck, "From password to biometrics: How far can we go," in 7th Asia-Europe Workshop on Concepts in Information theory, Boppard, Germany, pp. 1–8, 2011.

[Cover-Thomas, 2006] T. M. Cover and J. A. Thomas, Elements of Information Theory 2nd Edition. New York, NY, USA: Wiley-Interscience, 2006.

[Csiszar, 1998] I. Csiszar, "The method of types," IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2505–2523, 1998.

[Haroutunian et al, 2008] E. A. Haroutunian, M. E. Haroutunian, and A. N. Harutyunyan, "Reliability criteria in information theory and in statistical hypothesis testing," Foundations and Trends in Communications and Information Theory, vol. 4, no. 23, pp. 97–263, 2008.

[Haroutunian, 2007] E. Haroutunian, "On bounds for E - capacity of dmc," IEEE Transactions on Information Theory, vol. 53, no. 11, pp. 4210–4220, 2007.

[Haroutunian-Pahlevanyan, 2014] M.E. Haroutunian, N. S. Pahlevanyan "Information theoretical analysis of biometric secret key sharing model," Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science, vol.42, pp. 17-27, 2014.

[Ignatenko-Willems, 2012] T. Ignatenko and F. M. Willems, "Biometric security from an information-theoretical perspective," Foundations and Trends in Communications and Information Theory,vol. 7, no. 2-3, pp. 135-316, 2012.

[OSullivan-Schmid, 2002] J. A. OSullivan and N. A. Schmid, "Large deviations performance analysis for biometrics recognition." Proc. 40th Annual Allerton Conf. on Communication, Control, and Computing, pp. 1–10, Oct. 2002.

[Pahlevanyan, 2014] N. S. Pahlevanyan "Comparison of R language with other statistical tools," Armenian Mathematical Union Annual Session, p. 20, 2014.

[Venables et al, 2014] W. N. Venables, D. M. Smith and the R Core Team. "An Introduction to R", version 3.1.1, pp. 51-77, 2014.

[Willems et al, 2003] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in Information Theory, 2003. Proceedings. IEEE International Symposium on Information Theory, Yokohama, Japan, 2003, p. 82.

Authors' Information

Mariam Haroutunian *– Professor, Doctor of Physical and Mathematical Sciences, Leading Researcher and Head of department for Information Theory and Cognitive Models at Institute for Informatics and Automation Problems, National Academy of Sciences, Armenia;*
e-mail: armar@ipia.sci.am

Major Fields of Scientific Research: Information theory, Probability theory and Mathematical Statistics, Information-theoretic aspects of information security.

Narek Pahlevanyan *– Studying for PhD; Gyumri State Pedagogical Institute, Armenia; e-mail: narek@ravcap.com*

Major Fields of Scientific Research: Information theory, cryptography, cloud computing, machine learning, biometric secrecy systems.