PROCEEDINGS

International Workshop on

Information Theory and Data Science

From Information Age to Big Data Era

A Claude Shannon Centenary Event

Yerevan, Armenia

October 3-5, 2016

Editors

A.J. Han Vinck
Institute of Digital Signal Processing
University of Duisburg Essen

and

Ashot N. Harutyunyan
VMware

Proceedings of the Workshop “From Information Age to Big Data Era”
VMware Armenia Training Center
Yerevan, Armenia
October 3-5, 2016
With refs.
ISBN/EAN: 978-90-74249-28-7
Titel: From Information Age to Big Data Era:
Auteur: Vinck, A.J. Han/Harutyunyan, Ashot N.
Uitgever: Shannon, Stichting
Bibliografische imprint: Shannon, Stichting
NUR-code: 986
NUR-omschrijving: Datacommunicatie en netwerken
Druk: 1
Aantal paginas: 94
Taal: Engels
Verschijningsvorm: Paperback/softback

Message from Organizers

Era of Big Data has started with a tremendous transformative impact on technologies, business, and our daily life. Intrinsically, this era is enabled by the preceding revolutionary Age of Information, when the science and technology of digital communications rapidly progressed and changed the reality we live in.

2016 is 100th anniversary of the founder of the communication science and the Information Age – Claude E. Shannon. This is a special occasion for researchers from information theory, coding, and security communities to get together with professionals working on data science problems for a discussion forum on trends and opportunities connecting those Eras.

Our goal is to initiate a conversation between academia and industry, as well as engage students into modern and rapidly growing areas of research and technology innovation.

Hopefully this is a good start for establishing a series of similar meetings on a regular basis.

Please visit https://informationbigdata.wordpress.com/ for more information about the event.

Organizers**,**

Ashot N. Harutyunyan (VMware)

Davit A. Sahakyan (Monitis)

Honorary Chair Prof. A.J. Han Vinck (University of Duisburg-Essen).

Acknowledgements. Organizers are grateful to

VMware Armenia for hosting the event

and

Alexander von Humboldt Foundation for supporting Prof. Vinck's visit to Humboldt alumnus A. Harutyunyan for a collaboration week and participation in the workshop.

Contributions

Information Theory and Big Data

typical or non-typical, that is the question

A.J. Han Vinck
University of Duisburg-Essen, Germany

Summary of the Presentation

Big-Data can be seen as the collection/generation, storage/communication, processing and interpretation of big volumes of data. The collection, storage, processing of data is part of the problems described in field of information theory. However, in the famous paper by Shannon, in 1948, “a mathematical theory of communications” the interpretation or semantics are explicitly excluded in the definition of information theory, [1]. Shannon states: semantic aspects of communication are irrelevant to the engineering problem. Furthermore, messages are to be selected from a set of possible messages. Shannon clearly tackled the communication problem from an engineering point of view.

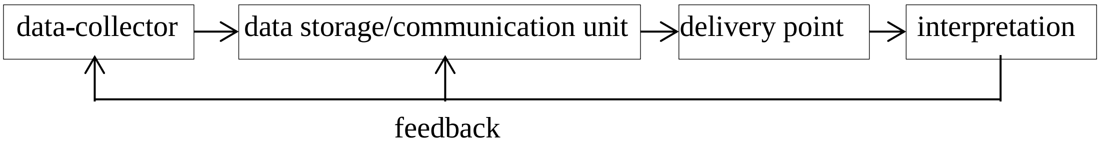


Figure 1. Semantics included in the Big-Data communication problem.

Figure 1 gives a practical situation, where data is collected/generated and communicated/stored. At the delivery point we have an interpretation of the data. The feedback can be introduced to select or generate specific data to be used for further interpretation. The feedback can also be used to reduce the complexity of storing/transmitting the data. For instance the interpretation can be used to improve the compression or data reduction efficiency before storing. Again, this semantic type of feedback is not included in the classical information theory.

Anomaly detection plays an important role in Big-Data. An anomaly is an event that is unknown beforehand and is outside the definition of the expected data values. Therefore, the detection of anomalies, as opposed to outliers, is not possible with the standard information theory arguments. Outliers can be detected by using the theory of typicality, see also Cover [2]. One can say that an outlier is an event that occurs very rare, with low probability. However, in the theory of typicality, one assumes that the underlying probability distributions are known and for instance the entropy of the collected data can be calculated. This is often not the case. The theory of estimating the entropy of a finite collection of data is a topic of recent information theoretical research projects in for instance neuro-science, see [3,4].

For larger volumes of data, the use of large distributed memory becomes important. The general problem is how to store and retrieve the data from a distributed storage. Reliability is an important issue since, with large memories, the probability of having also a correct memory goes to zero. Terabyte memories are not exceptional anymore. Therefore the use of error correcting - or error detecting coding is a necessity in modern memory systems. As an example, a simple one error correcting BCH code is used in SSD memory. An interesting question arises how much the application of coding improves on the lifetime of these memories. As a figure of merit, one can investigate the Mean Time to Failure (MTBF). In [5], we show that the improvement in the MTBF for memories of size N words of length n can be improved with a factor η,

$$\eta = \frac{k}{n} N^{dmin-1}/_{dmin-1},$$

where k/n and d_{min} are the efficiency and the minimum distance of the error correcting code, respectively. For a simple minimum distance $d_{min} = 3$ code the gain is proportional to $\sqrt{N}$.

Another important issue is that of data compression and data reduction. Data compression is the exact representation of the collected data in a reduced and efficient way. Many compression algorithms are available with different kind of applications. The performance of these algorithms depends on the stability of the data. Adaptive algorithms like Lempel-Ziv are preferred above algorithms that are static. Data reduction removes data that seems to be irrelevant for the user. Efficiency can be very high, but this is a technique that should be considered with care. In big-data we are interested in irregular behavior of data (anomaly) and thus reduction for efficiency may lead to removal of important data. The problem for the algorithms in general is that the source model is often not - or partly known. In information theory, the technique of typicality is a popular technique to determine whether data belongs to a typical set of data or to a non-typical set of data. In data compression we give special code words to the members of a non-typical set. In data reduction, we neglect this set. The main drawback of this technique is that it assumes pre-knowledge of the statistical properties of the data. This is in practice not always the case or not possible to obtain. In Figure 2 we give a model for data compression/reduction for independent data. In data compression, the difference with the "best match" of one of the N code words from the data base can be encoded together with the $\log_2 N$ bits for the selected match from the data base.

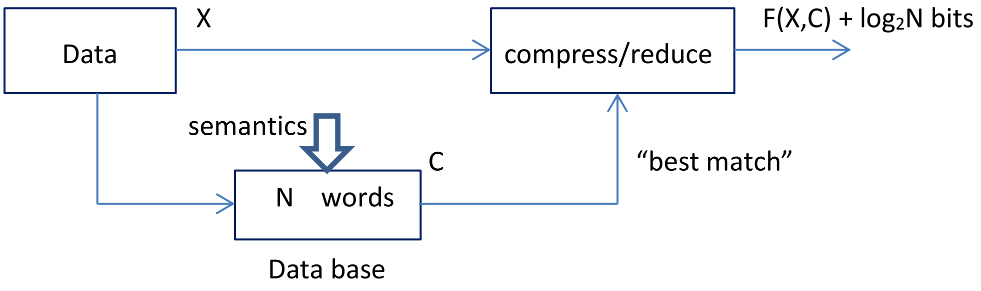


Figure 2. General source coding system.

For data reduction, using a distortion condition, we only encode the "best match" with $\log_2 N$ bits. We have to optimize N given the distortion condition. Adaptive algorithms (like Lempel-Ziv) update the data base depending on the received data X.

In Figure 3, we encode the dependent data with a predictor. At time T, the prediction from the predictor is used to determine and encode the difference between the data and the prediction. The input data with small error ε at time T is the new input to the predictor for time T+1. If the predictor at the receiver and transmitter side start with the same prediction, unique reconstruction at the receiver is possible, since only the value C and δ' are needed. We can also see where semantics can be used to improve the system performance.

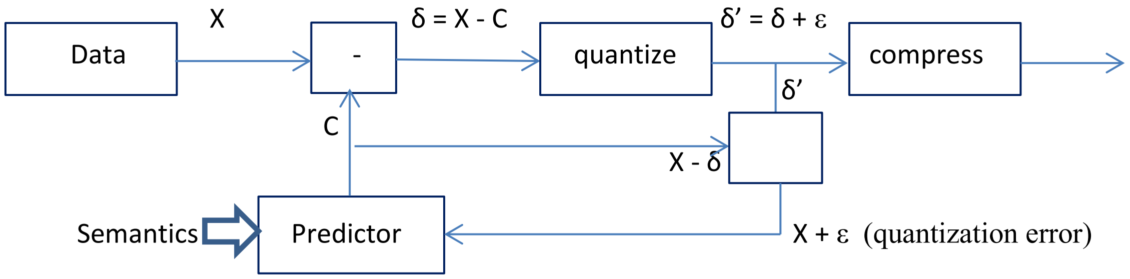


Figure 3. General source coding system with prediction and semantics.

One of the important applications in big-data is that of information retrieval. An ingenious system developed for libraries was presented in [4]. It clearly illustrates the difference between hard- and software. Suppose that we want to find an item in the data base that contains the keywords, alpha, beta, gamma. If all items have a list of key words, then a search can be executed to check for the specific keywords. Indexing techniques exist that solve this problem with low complexity. See for instance the – inverted indexes – technique, [6]. A well known technique is that of a Bloom filter, where each index term is hashed to k binary digits randomly positioned in an array of length n. This technique is similar to superimposed coding as introduced by Kautz and Singleton, [7]. In their work, a matrix of size N words of length n, where each word contains k nonzero digits is used. Every item in a system can have no more than T index terms. The item sum-index is then given by the Boolean OR of T rows, which should not cover any other row in the matrix. Extension to the q-ary OR are given in [8]. Other extensions consider covering of subsets containing the OR of more than one row.

Biometrics will play an important role in future systems. Face recognition in large crowds or universal biometrics for passports are examples of large amounts of data in which only a very small portion is of interest. A basic security scheme is presented by Shannon in his famous 1949 paper, see [9]. The principle is given in Figure 4, where we assume that the receiver has a noisy version of the encryption key. For perfect security, as defined by Shannon, the entropy H(M) ≤ H(K) – H(E), where H(E) is the entropy of the noise. For a noiseless key system, H(M) ≤ H(K). There is thus a price to pay, see [10]. We can translate this system directly to a biometric verification scheme, see Figure 5.

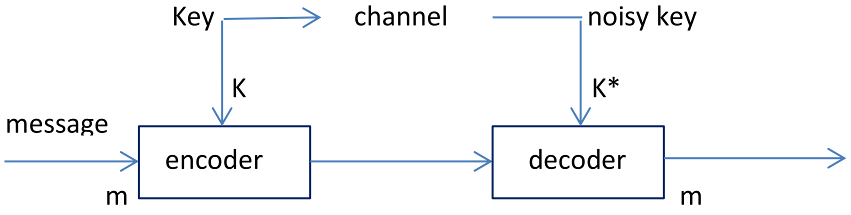


Figure 4. Shannon's cypher system with a noisy key.

Let c be a code word from an error correcting code, and (b XOR c) stored in a data base at the receiver. Then at the receiver side, for a particular user verification, we can calculate s = e XOR c and we can decode c and e, when e is within the decoding limits of the code. Using e, we can find back b, since b XOR e XOR e = b. A cryptographic hash function using b and c that is stored in a data base, can be checked at the receiver for validity. One could also derive a key from the corrected biometric. Note that the (b XOR c) in the data base should not give an opponent the opportunity to derive b. A good designed system must use a biometric and code that fit together. Code parameters must be chosen in such a way that the

False Acceptance Rate and the False Rejection Rate are low. The code redundancy and the error generating process are the key parameters for the design, [11].

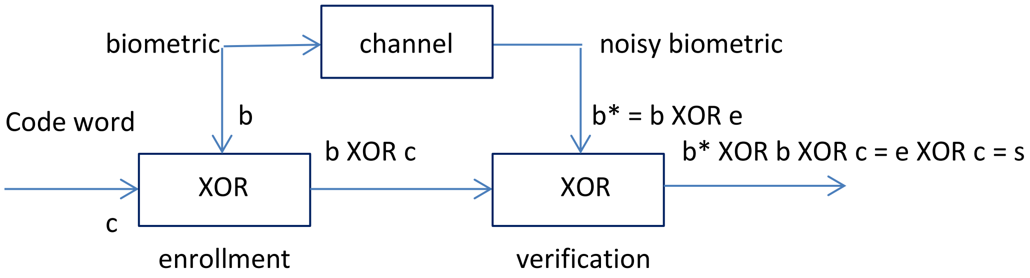


Figure 5. Shannon’s cypher system for a biometric key.

References

[1] C.E. Shannon, “A Mathematical Theory of Communication,” vol. 27, pp. 379–423, 623–656, October, 1948.

[2] Thomas M. Cover, Joy A. Thomas, Elements of information theory –2nd ed. “A Wiley Interscience publication.” July 2006.

[3] L. Paninski, *“*Estimation *of* entropy and mutual information,” *Neural Computation* 15: 1191-1254, 2013.

[4] Martin Vinck, Francesco P. Battaglia, Vladimir B. Balakirsky, A.J. Han Vinck, and Cyriel M. A. Pennartz, “Estimation of the entropy based on its polynomial representation,” *Phys. Rev.* E 85, 051139, 2012.

[5] A.J. Han Vinck and Karel Post ,“On the influence of coding on the mean time to failure for degrading memories with defects,” *IEEE Tr. on Information Theory*, pp. 902-906, July 1989.

[6] Calvin N. Mooers "Zatocoding and developments in information retrieval,” *Aslib Proceedings*, vol. 8, issue: 1, pp. 3 – 22, 1956.

[7] Black, Paul E., inverted index, Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Technology Oct 2006. Verified Dec 2006.

[8] A.J. Han Vinck and S. Martirossian , “On Superimposed Codes,” in *Numbers, Information and Complexity*, editors: Ingo Althöfer, Ning Cai, Gunter Dueck, 2013.

[9] C. E. Shannon, “Communication theory of secrecy systems,” *Bell System Technical Journal* 28, pp. 656 – 715, 1949.

[10] A.J. Han Vinck, Coding Concepts and Reed-Solomon Codes, 206 pages, ISBN 978-3-9813030-63 (hard copy can be ordered by author), 2013.

[11] Ulrike Korte, Michael Krawczak, Ullrich Martini, Johannes, Merkle, Rainer Plaga, Matthias Niesing, Carsten Tiemann, and Han Vinck, “A cryptographic biometric authentication system based on genetic fingerprints, (Extended Version),” GI-Edition - *Lecture Notes in Informatics* (LNI), P-128 , pp. 263-276, Bonner Köllen Verlag, ISBN 978-3-88579-222-2, 2008.

Security Challenges of Cloud Computing

Gurgen Khachatrian

American University of Armenia
Yerevan, Armenia

1 Extended Abstract

The emerging cloud technologies opened a new era in information storage and processing. Nowadays many companies and individuals are using public cloud storages such are Dropbox, Box, and Google Drive, to store their data instead of using private data storages. To avoid the security risks the user should store its data in the encrypted form. This will protect the data security in untrusted environment such as cloud, but will significantly change the users experience as they will not be able easily access any data by making search. To solve this problem new directions in cryptography have been emerged investigating the search possibilities over encrypted data. Two main directions can be distinguished, namely, so-called Secure Pattern Search and Searchable Symmetric Encryption.

1) The first important direction in this area is the Secure Pattern Search in the specific file or string. There is a Secure Database (SB) and different users need to search over SB for a specific data string. The objective of this search is twofold: firstly any user searching for a specific data string in SB should only get YES/No information exposing the existence of his queried string in database no other information in addition, secondly SB owner should not get any information what specific data string was searched. However secure pattern search is an advanced setting not employed yet in practice as the known solutions for this problems are not practical since they require to employ computationally expensive

public key operations which makes existing solutions highly impractical.In general this novel secure search engine will allow to search over encrypted data without tracking user's behavior.

2) Searchable Symmetric Encryption technology allows the data owner to protect its data consisting of many different files by using a symmetric encryption in the way that it is still possible to search over encrypted data in the case when all encrypted data is stored in the remote server (the cloud), at the same time allowing the cloud to learn as little as possible information about the search results. In nature the data owner creates a special encrypted index related to his files and stores the index among with the files in the cloud server. Later on the data owner will be able to send special search tokens corresponding to some keyword to the cloud and the cloud can search and get all the files containing the keyword. Consequently, the cloud should not learn anything about the searched keyword itself.

Because of its practical significance a Searchable symmetric encryption (SSE) has been an active area of research and development during the last decade. Besides some results mostly theoretical in nature the main focus was concentrated on practical SSE schemes. Any practical scheme obviously should render a reasonable trade-off between the following properties: search time, security, compact indexes and the ability to efficiently update the database i.e. add and delete files. From the other hand, the emerging cloud technologies require more efficient and more functional cryptographic primitives for security management and particularly key exchanges between cloud applications, software as a service providers and end users.

3) One of the major problems dealing with secure cloud computations which can be a bottleneck of efficient computations in a real time is the processing speed of public key operations used during an interaction between cloud entities and users .As such a development of novel public key systems with significantly faster processing speed compared with known algorithms becomes a critical issue. It can be called an High performance cryptographyin Big data.

4) White-box cryptography(WBC) is a relatively new cryptographic discipline developed in last decade which allows transforming any block cipher to public key encryption scheme resulting to about 1000 times

more efficient system compared with traditional public key schemes such as RSA or El-Gamal. WBC had found numerous applications in DRM systems widely used by Apple, IRDETO, NetFlix. Our own scientific research has shown its applicability in different security protocols including the secure pattern search application. The development of secure white-box scheme is a very challenging task. The major problem with WBC is its security. Up to date all publicly known WBC systems have been broken. We intend to develop a novel approach to the design of WBC systems with provable security.

5) Development of a new Public key systems based on permutation polynomials which run significantly faster compared with well known systems and have a comparable security. The papers [1] and [2] represent new results obtained recently in that direction. In particular a paper [2] represents performance results for the new cryptosystem based on permutation polynomials which show that with the comparable security it runs faster by 130 times compared with RSA-2048.

References

[1] G. Khachatrian, M. Kuregian, ”Permutation polynomials and a new public key encryption ” - *accepted for publication in Discrete Applied Mathematics journal*, February 2015

[2] G. Khachatrian, M. Karapetyan, ”On a public key encryption algorithm based on permutation polynomials and performance analysis” - *International Journal Information Theories and Applications* , Vol, 23, Number 1, pages 34-38, 2016

Information Theory Research in Armenia

Mariam Haroutunian and Evgueni Haroutunian
Institute for Informatics and Automation Problems
NAS RA

To Claude E. Shannon Centenary.
To the memory of Roland L. Dobrushin

Abstract

This survey summarizes the results of Armenian researchers in the field of Information Theory. The research is devoted to the problems of determination of interdependence between coding rate and error probability exponent for different information transmission scenario.

1 Introduction

This year the scientific world celebrated Shannon’s centennial as the father of the information age. Shannon is best known for developing the mathematical foundations of communication (establishing the field of information theory), data compression, digital computers, cryptography, circuit complexity, flow networks, and juggling, as well as laying foundations of artificial intelligence and human-computer interaction.

As the founder of Information theory Shannon mathematically addressed the basic problems in communications and gave their solutions, stating the three fundamental discoveries underlying the information theory concerning the transmission problem via noisy channel and its inherent concept – capacity, data compression with the central role of entropy in that, and source coding under fidelity criterion with specification of the possible performance limit in terms of the mutual information introduced by him [1, 2].

Under the term “Shannon Theory” it is generally accepted now to mean the subfield of information theory which deals with the establishment of performance bounds for various parameters of transmission systems.

The information theory research in Armenia was led by Dobrushin. He repeatedly visited Armenia with a series of lectures in Yerevan State University, which were later published in [3]. He had PhD students in Armenia, among which was E. Haroutunian.

Important properties of each communication channel are characterized by the **reliability function** $E(R)$, which was introduced by Shannon [4], as the optimal exponent of the exponential decrease

$$\exp\{-NE(R)\}$$

of the decoding error probability, when code length N increases, for given transmission rate R less than capacity C of the channel. In an analogous sense one can characterize various communication systems. The reliability function $E(R)$ is also called the **error probability exponent**. Besides, by analogy with the concept of the rate-distortion function [5], the function $E(R)$ may be called the **reliability-rate function.**

Because of principal difficulty of finding the reliability function for the whole range of rates $0 < R < C$, this problem is completely solved only in rather particular cases. The situation is typical when obtained upper and lower bounds for the function $E(R)$ coincide only for rates R in some interval, say $R_{crit} < R < C$, where R_{crit} is the rate, for which the derivative of $E(R)$ by R equals 1.

It seems, that the approach of studying the function $R(E) = C(E)$, inverse to $E(R)$, suggested by E. Haroutunian [6 - 8] and developed by his students is more effective and fruitful for investigation of more complicated information transmission systems [9]. This is not a simple mechanical permutation of roles of independent and dependent variables, since the investigation of optimal rates of codes, ensuring when N increases the error probability exponential decrease with given exponent (reliability) E, can be more expedient than the study of the function $E(R)$.

At the same time, there is an analogy with the problem from coding theory about bounding of codes optimal volume depending on their correction ability. This allows to hope for profitable application of results and methods of one theory in the other. The definition of the function $C(E)$ is in natural conformity with Shannon's notions of the channel capacity C. So, by analogy with the definition of the capacity, this characteristic of the channel may be called ***E*-capacity**. From the other side the name **rate-reliability function** is also logical. One of the advantages of our approach is the convenience in study of the optimal rates of source codes ensuring given exponential decrease of probability of exceeding the given distortion level of messages restoration. This is the **rate-reliability-distortion function** $R(E, \Delta)$ inverse to **exponent function** $E(R, \Delta)$ by Marton [10]. So the name shows which dependence of characteristics is in study. It makes possible to consider also other arguments, for example, coding rates on different inputs of channel or source, if their number is greater than one. This makes the theory more well-ordered and comprehensible.

We can note the following practically useful circumstance: the comparison of the analytical form of writing of the sphere packing bound for $C(E)$ with expression of the capacity C in some cases gives us the possibility to write down formally the bound for each system, for which the achievable rates region (capacity) is known. In rate-reliability-distortion theory an advantage of the approach is the technical ease of treatment of the coding rate as a function of distortion and error exponent which allows to convert readily the results from the rate-reliability-distortion area to the rate-distortion ones looking at the extremal values of the reliability, e.g. $E \to \infty; E \to 0$. This fact is especially important when one deals with multidimensional situation. Having solved the problem of finding the rate-reliability-distortion region of a multiterminal system, the corresponding rate-distortion one can be deduced without an effort.

One of the main contributions of E. Haroutunian to Information Theory is that he was the first who expressed different bounds by maxmin of functionals with entropy, information and divergence [7]. Before that the bounds were written in the more complicated form of Gallager. In the mentioned paper the coincidence of this two forms of the Sphere packimg bound was proven. Now the Haroutunian's form of presentation of different bounds is named "standard".

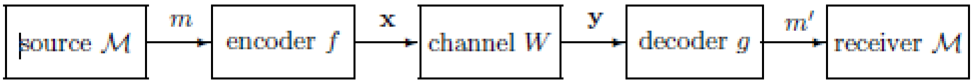
Many of results of E. Haroutunian are included in the textbooks of information theory [11, 12] and are cited by other authors, but more often the result from [13] on upper bound to the error exponent for channels with feedback is mentioned as Haroutunian's exponent or Haroutunian's bound (see for example [14]). Despite the numerous investigations this bound has not been improved till now.

Combinatorial methods developed in information theory are applied also for investigation of problems of the logarithmically asymptotically optimal testing of statistical hypotheses.
Authors of this survey have been teaching courses of Information Theory and particularly the concept of E-capacity in Yerevan State University and in Armenian State Engineering University for many years and prepared teaching aids in Armenian [15, 16].

2 Bounds for E-capacity of discrete memoryless channel (DMC)

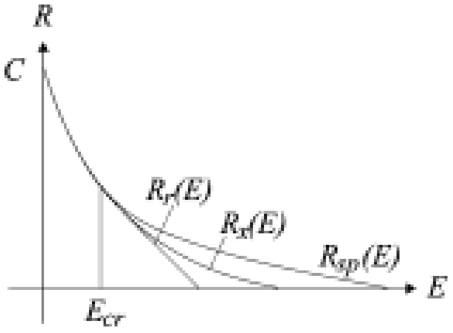
The concept of E-capacity was first considered by E. Haroutunian in [6].

The **DMC** is a one-way communication noisy channel with finite input and output alphabets and stochastic matrix of transition probabilities, which operate at each moment of time independently from the previous and next transmitted or received symbols.



Communication system with noisy channel

The survey of results for E-capacity of DMC is given in [8], where the random coding $R_r(E)$, expurgated $R_x(E)$, and sphere packing $R_{sp}(E)$, bounds are derived for maximal and average error probabilities and the comparison of the upper and the lower bounds is performed.



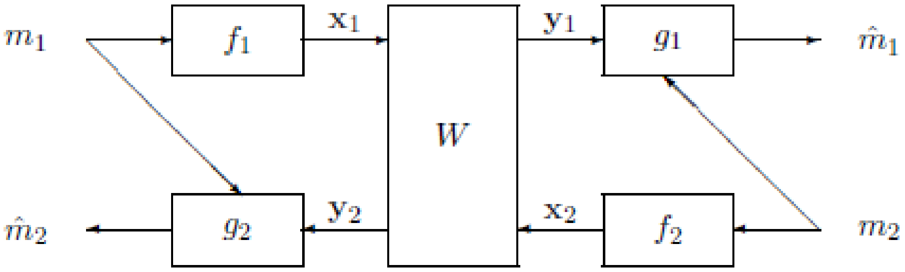
Typical behavior of the bounds for E-capacity of DMC.

Concerning methods for the bounds construction, it is found that the Shannon's random coding method [1] of proving the existence of codes with definite properties, can be applied with the same success for studying of the rate-reliability function. For the converse coding theorem type upper bounds deduction (so called sphere packing bounds) E. Haroutunian proposed a simple combinatorial method [8, 17], which one can apply to various systems. For the construction of expurgated bound the method of graph decomposition introduced by Csisza'r and Körner in [18] was applied. The proofs are based on the method of types [19, 20].

3 Multiterminal channels

Various multiterminal channels have been investigated and bounds for E-capacity regions have been constructed. The detailed reference list can be found in [9], so we omit it here, just remind the main results.

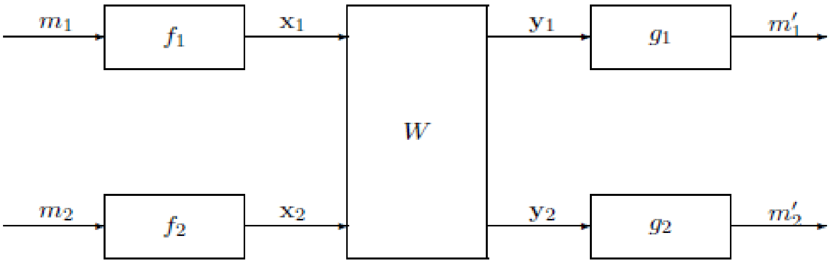
The **two-way channel** (TWC) has two terminals and the transmission in one direction interferes with the transmission in the opposite direction. The sources of two terminals are independent. In the general model the encoding at each terminal depends on both the message to be transmitted and the sequence of symbols received at that terminal. Similarly the decoding at each terminal depends on the sequence of symbols received and sent at that terminal. We have considered the restricted version of TWC, where the transmitted sequence from each terminal depends only on the message but does not depend on the received sequence at that terminal.



Restricted two-way channel

The Restricted TWC as well as the general TWC were first investigated by Shannon, who obtained the capacity region of the RTWC. The capacity region of the general TWC has not been found up to now. Important results relative to various models of two-way channels were obtained by many authors, particularly, it was demonstrated that the capacity regions of TWC for average and maximal error probabilities do not coincide. We have constructed the outer and inner bounds of E-capacity region [21].

Shannon considered also another version of the TWC, where the transmission of information from one sender to its corresponding receiver may interfere with the transmission of information from the other sender to its receiver, which was later called **interference channel** (IFC). The general IFC differs from the TWC in two respects: the sender at each terminal does not observe the outputs at that terminal and there is no side information at the receivers.

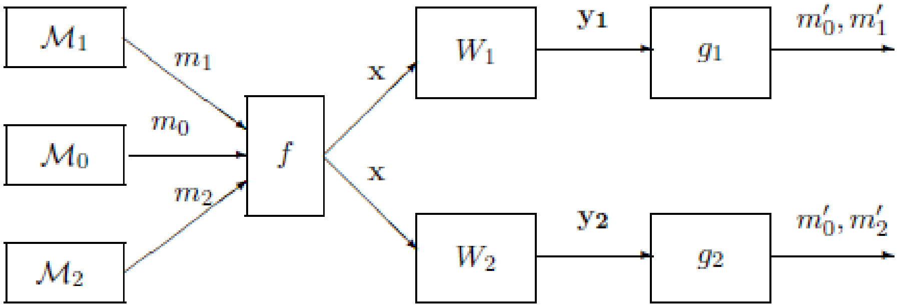


General interference channel

Ahlswede [22] obtained bounds for capacity region of general IFC but the capacity region is found only in particular cases. We have constructed the random coding bound of E-capacity region for general IFC, as well as for the IFC with cribbing encoders, when the second encoder learns from the first encoder the codeword, that will be sent in the present block.

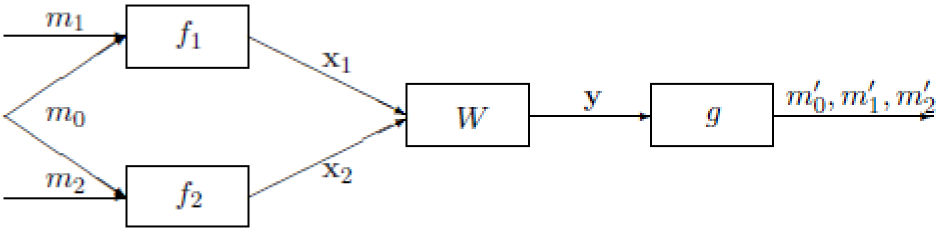
The next considered communication system is the **broadcast channel** (BC) with one encoder and two or more receivers. We study the BC with two receivers and three sources, two private and one common.

BC were first studied by Cover [23]. Despite the fact that in many works several models of BC were considered the capacity region of BC in the situation, when two private and one common messages must be transmitted, has not yet been found. In [24] Willems proved that the capacity regions of BC for maximal and average error probabilities are the same. We found the inner bound for the capacity region of BC.



Broadcast channel

The most general model of the discrete memoryless **multiple-access channel** (MAC) is the channel with correlated encoder inputs.



MAC with correlated encoder inputs

Three independent sources create messages to be transmitted by two encoders. One of the sources is connected with both encoders and each of the two others is connected with only one of the encoders. Dueck [25] showed that in general the maximal error capacity region of MAC is smaller than the corresponding average error capacity region. Determination of the maximal error capacity region of the MAC in various communication situations is still an open problem. We investigated the E-capacity region for the average error probability by constructing outer and inner bounds. Similar results are obtained for special cases, such as regular MAC with $M_0 = 1$, asymmetric MAC with $M_1 = 1$, MAC with cribbing encoders.

4 Varying channels

Channel in which the transition probabilities depend on a parameter s are called varying channels. Values of the parameter s can be changed by different rules, depending on which different models of channels are formed. Varying channels can be considered in different situations, when the state of the channel is known or unknown on the encoder and decoder.

The DMC is called **compound channel** (CC), if the state s of the channel is invariable during transmission of one codeword of length N, but can be changed arbitrarily for transmission of the next codeword. The capacity of this channel was found by Wolfowitz [26], who showed that the knowledge of the state s at the decoder does not improve the asymptotic characteristics of the channel. So it is enough to study the channel in two cases. As for DMC, the capacity of the compound channel for average and maximal error probabilities are the same. We have formulated the sphere packing, random coding and expurgated bounds of E-capacity of CC.

The **channel with random parameter** (CRP) is a family of discrete memoryless channels with the state, varying independently at each moment of the channel action with the same known PD $Q(s)$ on S. From mathematical point of view the most interesting is the situation with the additional information at the encoder. In [27] Gelfand and Pinsker determined the capacity of CRP for average error probability in the case of non-causal side information. For all four situations the upper and lower bounds of E-capacity are obtained by M. Haroutunian [28]. She also studied the generalized CRP where the PD Q of the states is invariable during the transmission of length N, but can be changed arbitrarily during the next transmission.

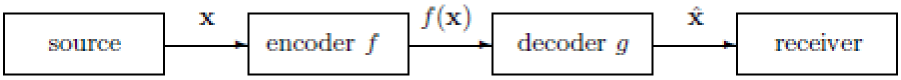
The **multiple-access channel with random parameter** (MACRP) is a family of discrete memoryless multiple-access channels, where s is the channel state, varying independently at each moment with the same PD $Q(s)$ on S. M. Haroutunian studied the E-capacity of the MAC with random parameter in various situations, when the whole state sequence s is known or unknown at the encoders and at the decoder [29].

The **arbitrarily varying channel** (AVC) is a discrete memoryless transmission system, which depends on state s that may change in an arbitrary manner within a finite set S. Bounds of E-capacity are obtained for this channel in the situation of informed encoder [30].

The detailed proofs of above mentioned results are included in the doctoral thesis of M. Haroutunian [31].

5 Source coding

Now we expound the concept of the **rate-reliability-distortion function** [32] for discrete memoryless sources (DMS).



Noiseless communication system

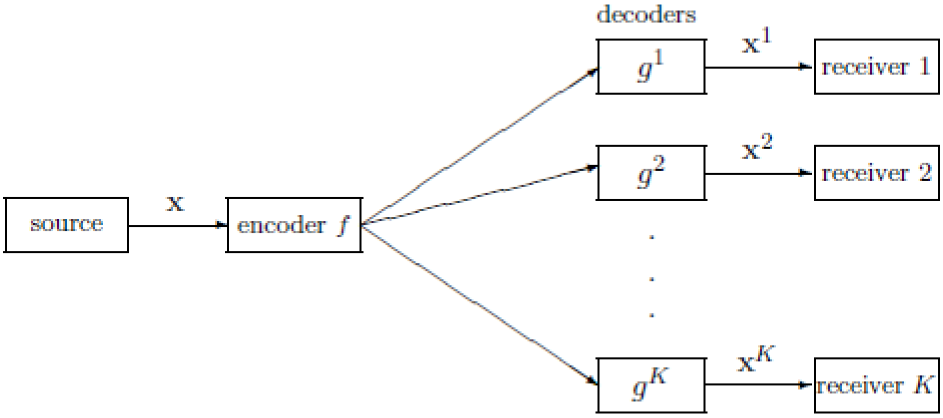
Shannon rate-distortion function [2] shows the dependence of the asymptotically minimal coding rate on a required average fidelity (distortion) threshold for source noiseless transmission. Another characteristic in source coding subject to a distortion criterion can be considered, namely, an exponential decrease in error probability with a desirable exponent or reliability. The maximum error exponent as a function of coding rate and distortion in the rate-distortion problem was specified by Marton [10].

An alternative order dependence of the three parameters was examined by E. Haroutunian and his students. They define the rate-reliability-distortion function as the minimal rate at which the messages of a source can be encoded and then reconstructed by the receiver with distortion level and an error probability that decreases exponentially with the codeword length. Therefore, the achievability of the coding rate R is considered as a function of a fixed distortion level $\Delta \geq 0$ and an error exponent $E > 0$. In a series of works E. Haroutunian, R. Maroutian, A. Harutyunyan, A. Kazarian successively extended this idea into the multiuser source coding problems (the full list of publications can be found in [9], see also [33-35]).

Actually, the approach brings a technical ease. Solving a general problem on rate-reliability-distortion one can readily convert the results from that area to the rate-distortion one looking at the extremal values of the reliability, e.g. $E \to \infty$, $E \to 0$. It is more useful when we deal with a multiuser source coding problem.

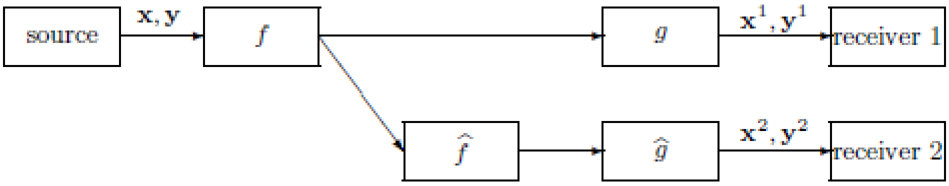
The error exponent criterion in lossy coding for more general class of sources – **arbitrarily varying sources** (AVS) was considered [36]. Here the source outputs distribution depends on the source state, which varies within a finite set from one time instant to the next in an arbitrary manner. The problem statement for the AVS coding subject to fidelity and reliability criteria is totally identical to the DMS coding problem under the same constraints, with only difference in the source model.

The source coding problem with fidelity and reliability criteria for the **robust descriptions system** with one encoder and many decoders was considered. Messages of a DMS encoded by one encoder must be transmitted to K different receivers. Each of them, based upon the same codeword, has to restore the original message within a distortion with a reliability both acceptable to him. The rate-reliability-distortion function is specified [37].



Robust description system

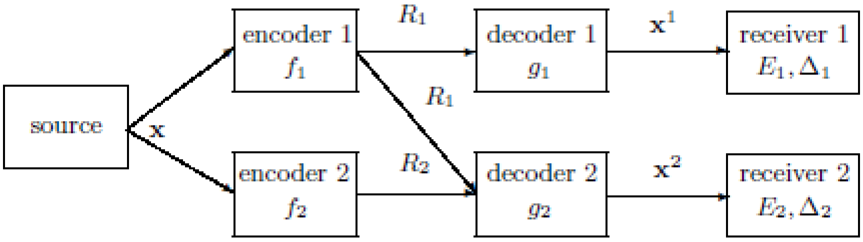
The next considered model is the **cascade communication system**, where two correlated sources coded by a common encoder and a separate encoder must be transmitted to two receivers by two common decoders within prescribed distortion levels.



Cascade communication system

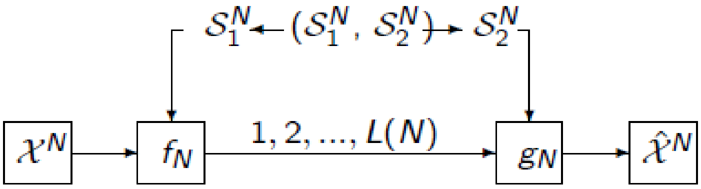
The region of all achievable rates of the best codes ensuring reconstruction of messages within given distortion levels with error probabilities exponents is illustrated.

The **hierarchical source** coding problem and the relevant concept of successive refinement of information subject to reliability criterion was also discussed [38, 39]. The characterization of the rate-reliability-distortion region for the hierarchical source coding and the conditions for successive refinability subject to reliability criterion were derived.



The hierarchical communication system

The **generalized model of the DMS with two-sided state information** has been investigated in the sense of rate-reliability-distortion function.



Source with two-sided state information

6 Logarithmically asymptotically optimal testing of statistical hypotheses

The usefulness of combinatorial methods developed in information theory to investigation of the logarithmically asymptotically optimal (LAO) testing of statistical hypotheses was illustrated in the research of E. Haroutunian, P. Hakobyan, N. Grigoryan, A. Esayan, F. Hormozi-Nejad [40 - 44].

The research is dedicated to the problem of multiple hypothesis testing in terms of error exponents. The classical results for binary hypothesis testing are extended to a case with $M(\geq 2)$ distributions and $K(\geq 1)$ objects.

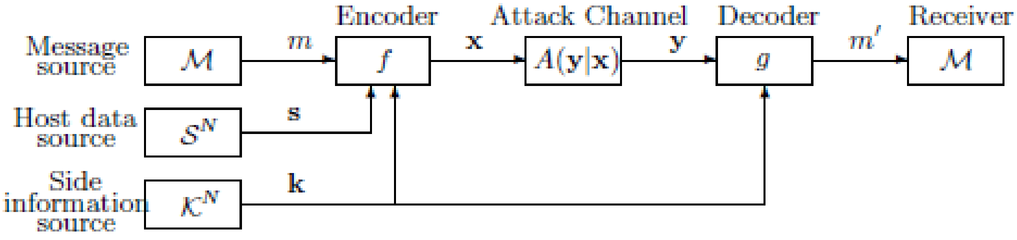
The multihypotheses optimal testing problem for the model of independent or related (stochastically, statistically and strictly dependent) objects also was considered. The main goal of the test given the observed sequence, is to decide which statistical distribution it represents. The above mentioned authors investigated the functional dependence of all possible pairs of the error probability exponents (reliabilities) of the optimal tests for the sequence of experiments concerning Markov chain and the case of many hypotheses for independent experiments. The dependence between reliabilities of the hypothesis two-stage optimal testing the model described by the Markov chain and the case for independent experiments was also investigated. So the reliability matrices of LAO procedures for some models were obtained. The hypothesis testing problem for arbitrarily varying source without and with multiterminal data compression was solved.

The identification problems for independent objects and for different objects are solved from the optimal testing point of view. The functional interconnection between the reliabilities of optimal identification is derived [45]. The problem of identification of an object distribution and the problems of r-identification and ranking also are discussed.

7 Major results in the field of information-theoretical security

The research has been carried out also in the field of information-theoretical security. Here we present the short description of main topics.

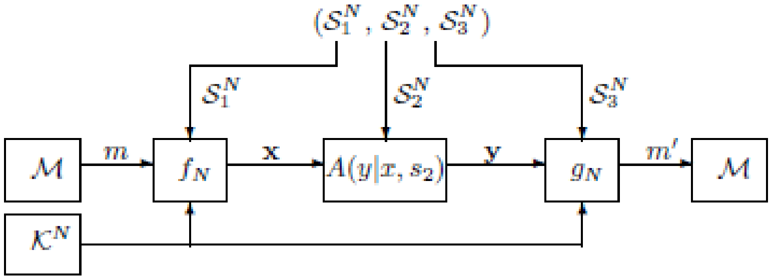
The first topic is the investigation of various **information-hiding systems** [46]. Many application areas, such as the copyright protection for digital media, watermarking, fingerprinting, steganography and data embedding have a certain generality, which can be formulated as information hiding problem. The message (watermark, fingerprint, etc.) needs to be embedded in the host data set (which can be the blocks from the audio, image and video data) and to be reliably transmitted to a receiver via channel which can be subject to random attacks. Side information, which can be cryptographic keys, properties of the host data, features of audio, image or video data or locations of watermarks, is available both to encoder and decoder. The encoding and decoding functions are known to the attacker, the side information is not. The information hider introduces certain distortion in the host data set for the data embedding. The attacker trying to change or remove this hidden information, introduces some other distortion.



The model of information hiding system

We investigate the information hiding *E*-capacity, which expresses the dependence of the information hiding rate from reliability and distortion levels for information hider and attacker [47].

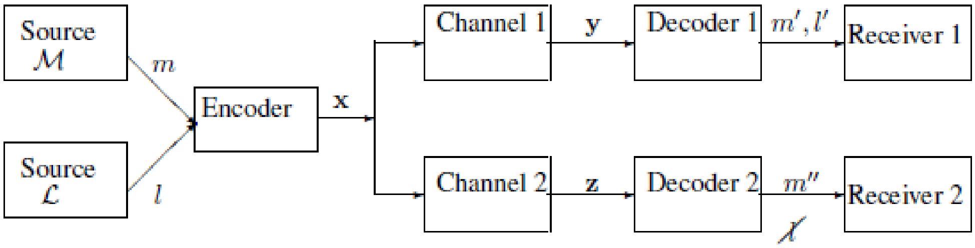
We have studied the **generalized model of channel with side information** [48], i.e. DMC with finite input and output alphabets and random state sequence (side information) partially known to the encoder, channel and decoder. The study includes family of channels with side information and information hiding coding problems as special cases. Information is to be reliably transmitted through the noisy channel selected by adversary. Reasoning from applications the actions of encoder and adversary are limited by distortion constraints. The encoder and decoder depend on a random variable which can be treated as cryptographic key. Two cases are considered, when the joint distribution of this RV and side information is given or this RV is independent from side information and its distribution can be chosen for the best code generation.



Channel with side information

We investigated the rate-reliability-distortion function for the mentioned model and derived the lower bound for it.

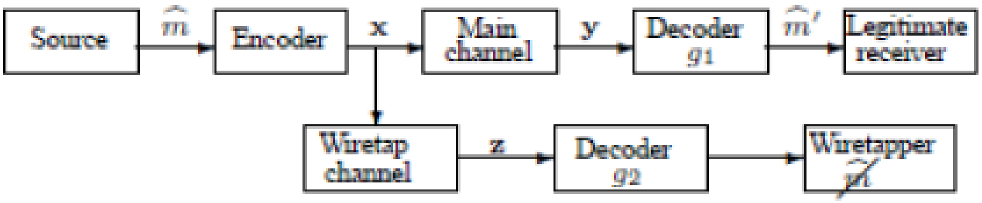
The next topic is the investigation of **broadcast channels with confidential messages** [49].



The model of broadcast channel with confidential messages

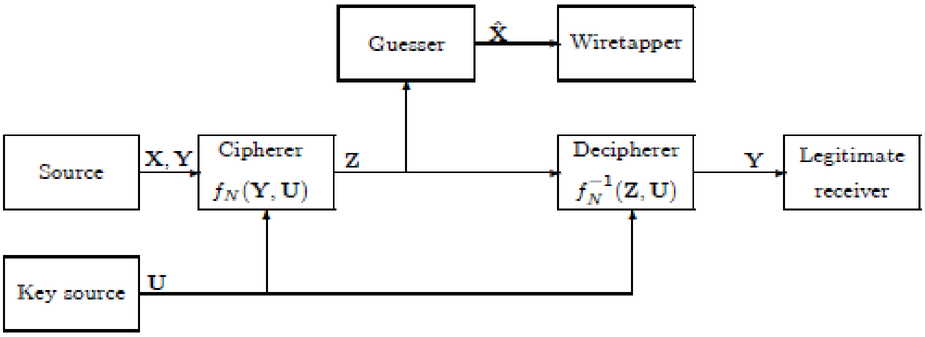
The information protection systems requiring the reliability and the confidentiality from eavesdropping are studied [50].

The object of study of **wiretap channel** is to maximize the rate of reliable communication from the source to the legitimate receiver, while the wiretapper learns as little as possible about source output. The E-capacity region of the wiretap channel is studied in [51].



The model of generalized wiretap channel

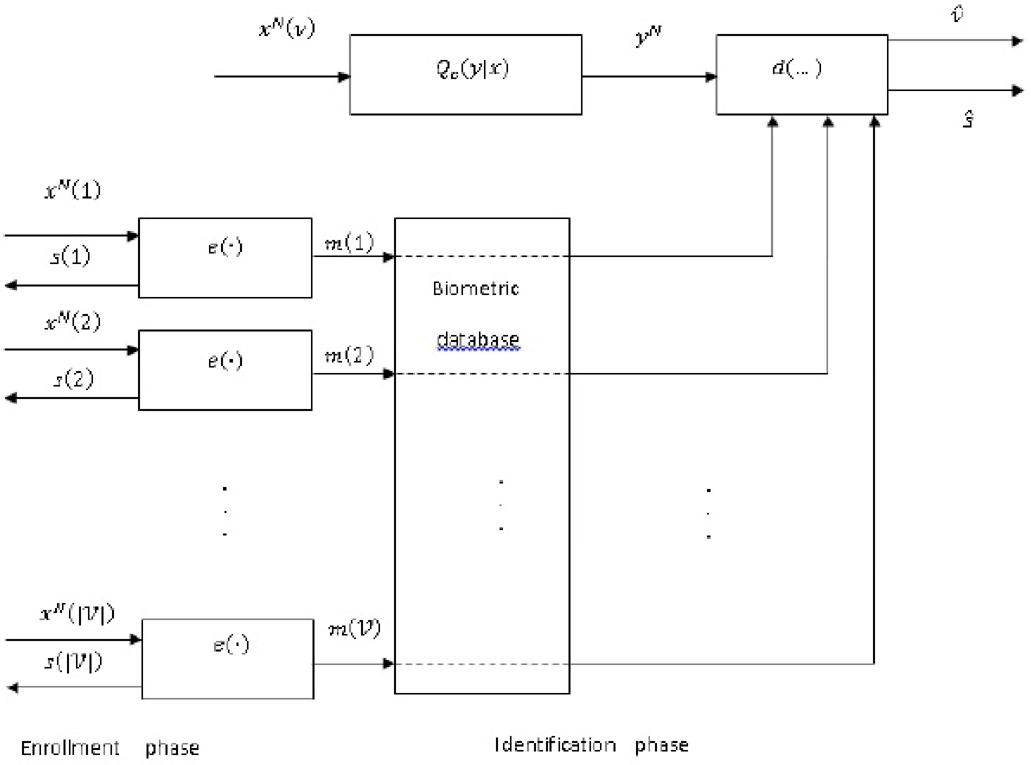
Another direction of investigations is the **Shannon cipher system with the guessing wiretapper** [52]. Various models of cipher systems are studied and the security level is estimated.



Shannon cipher system with the guessing wiretapper

As a security criterion the guessing exponent with reliability requirements is considered, when the wiretapper is allowed to have many chances to reconstruct the plaintext.

Information-theoretical investigations are also carried out for **biometric identification systems**. Person identification is one of the important scientific and practical tasks of current information security.



Identification system

This line of research aims to find the best possible security bounds of main characteristics of different systems. The interconnection of main characteristics (the key rate, secrecy rate, reliability, error probability, leakage rate) are investigated.

8 Module for R

The new package for R environment for estimation and computation of complex formulas of Information Theory has been developed [53]. An option for using of three types of parallelization inside the package has been developed, which allows change of parallelization type during computation process - parallelization through the local processor (CPU type), graphic card (GPU type) and cluster (MPI type). The package can be used by specialists in Information Theory for calculation of a number of sophisticated functions. It is available at http://packages.reviewed.r-project-0-mirror.com/AdvInfTheo_v1.0.5.tar.gz

9 New directions and open problems

Today we plan new investigations in the field of **Quantum Information theory**, which is developing very fast nowadays, because it gives a greater understanding of the foundations of quantum mechanics.

Secure distributed hypothesis testing is another field for our future investigations.

We are also interested in using information-theoretical techniques and concepts in **statistical learning** problems, where despite the practical success there are many open problems in finding fundamental limits and tradeoffs.

References

[1] C.E. Shannon, “A mathematical theory of communication”, *Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, 1948.

[2] C.E. Shannon, “Coding theorems for a discrete source with a fidelity criterion”, *IRE National Convention Record*, vol. 7, pp. 142-163, 1959.

[3] R.L. Dobrushin, “Information optimal coding theory”, (in Russian) *Cybernetics on the service of communism*, vol. 3, pp. 13-45, 1966.

[4] C.E. Shannon, “Probability of error for optimal codes in a Gaussian channel”, *Bell System Technical Journal*, vol. 38, no. 5, pp. 611-656, 1959.

[5] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[6] E.A. Haroutunian, “Upper estimate of transmission rate for memoryless channel with countable number of output signals under given error probability exponent”, (in Russian) *in 3rd All Union Conf. on Theory of Inform. Transmission and Coding, Uzhgorod, Publishing House of the Uzbek Academy of Science*, pp. 83-86, Tashkent, 1967.

[7] E.A. Haroutunian, “Estimates of the error probability exponent for a semicontinuous memoryless channel” (in Russian), *Problems of Information Transmission*, vol. 4, no. 4, pp. 37-48, 1968.

[8] E.A. Haroutunian, “On bounds for E-capacity of DMC”, *IEEE Transactions on Information Theory*, vol. 53, No. 11, pp. 4210-4220, 2007.

[9] 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. 2-3, pp. 27–263, 2007.

[10] K. Marton, “Error exponent for source coding with a fidelity criterion”, *IEEE Transactions on Information Theory,* vol. 20, no. 2, pp. 197-199, 1974.

[11] I. Csisza'r and J. Körner, *Information Theory. Coding theorems for Discrete Memoryless Systems*, New York: Academic Press, 1981.

[12] V.D. Kolesnik and G.Sh. Poltirev, *Information Theory Course,* (in Russian), Moscow, 1982.

[13] E.A. Haroutunian, “Lower bound for error probability in channels with feedback”, *Problems of Information Transmission,* (in Russian), vol. 13, no. 2, pp. 36-44, 1977.

[14] H. Palaiyanur and A. Sahai, “On Haroutunian’s exponent for parallel channels and an application to fixed-delay codes without feedback”, *IEEE Transactions on Information Theory,* vol. 61, no. 3, pp. 1298-1308, 2015.

[15] E. A. Haroutunian and M. E. Haroutunian, *Information Theory,* textbook in Armenian, YSU, 104p, 1987.

[16] M. E. Haroutunian, *Basics of Information Theory,* textbook in Armenian, ASEU, 140p, 2008.

[17] E.A. Haroutunian, “Combinatorial method of construction of the upper bound for E-capacity”, (in Russian), *Mezhvuz. Sbornic Nouch. Trudov, Matematika,* Yerevan, vol. 1, pp. 213-220, 1982.

[18] I. Csisza'r and J. Körner, “Graph decomposition: A new key to coding theorems”, *IEEE Transactions on Information Theory,* vol. 27, no. 1, pp. 5-12, 1981.

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

[20] I. Csisza'r. “The method of types.” *IEEE Transactions on Information Theory,* vol. 44, no. 6, pp. 2505-2523, 1998.

[21] E. Haroutunian, M. Haroutunian and A. Avetisyan, “Restricted two-way channel: Bounds for achievable rates region for given error probability exponents”, *in Proceedings of IEEE International Symposium on Information Theory,* Whistler, Canada, p. 135, 1995.

[22] R.F. Ahlswede, “The capacity region of a channel with two senders and two receivers”, *Annals of Probability,* vol. 2, no. 2, pp. 805-814, 1974.

[23] T.M. Cover, “Broadcast channels”, *IEEE Transactions on Information Theory,* vol. 18, no. 1, pp. 2-14, 1972.

[24] F.M.J. Willems, “The maximal-error and average-error capacity region of the broadcast channel are identical”, *Problems of Control and Information Theory,* vol. 19, no. 4, pp. 339-347, 1990.

[25] G. Dueck, “Maximal error capacity regions are smaller than average error capacity regions for multi-user channels”, *Problems of Control and Information Theory,* vol. 7, no. 1, pp. 11-19, 1978.

[26] J. Wolfowitz, “Simultaneous channels”, *Archive for Rational Mechanics and Analysis,* vol. 4, no. 4, pp. 371-386, 1960.

[27] S.I. Gelfand and M.S. Pinsker, “Coding for channel with random parameters”, *Problems of Control and Information Theory,* vol. 8, no. 1, pp. 19-31, 1980.

[28] M.E. Haroutunian, “Bounds of E-capacity for channel with random parameter”, *Problems of Information Transmission,* (in Russian), vol. 27, no. 1, pp. 14-23, 1991.

[29] M.E. Haroutunian, “Bounds of E-capacity for multiple-access channel with random parameter”, *Lecture Notes in Computer Science*, vol. 4123, "General

Theory of Information Transfer and Combinatorics", Springer Verlag, pp. 196-217, 2006.

[30] M.E. Haroutunian, “E-capacity of arbitrarily varying channel with informed encoder”, *Problems of Information Transmission,* (in Russian), vol. 26, no. 4, pp. 16-23, 1990.

[31] M.E. Haroutunian, “Information-theoretical investigation of discrete multi-terminal and varying channels”, *doctoral thesis*, Moscow, 2005.

[32] A.N. Harutyunyan and E.A. Haroutunian, “On properties of rate-reliability-distortion functions”, *IEEE Transactions on Information Theory,* vol. 50, no. 11, pp. 2768-2773, 2004.

[33] R.Sh. Maroutian, “Coding rate bounds of DMS”, *PhD thesis*, Yerevan, 1990.

[34] A.N. Harutyunyan, “Investigation of achievable interdependence between coding rates and reliability for several classes of sources”, *PhD thesis*, Yerevan, 1997.

[35] A.R. Ghazarian, “Multiterminal sources optimal coding rates depending on levels of reliability, distortion and secrecy”, *PhD thesis*, Yerevan, 1999.

[36] A.N. Harutyunyan and A.J. Han Vinck, “Error exponent in AVS coding”, *in Proceedings of IEEE International Symposium on Information Theory*, Seattle, WA, pp. 2166-2170, 2006.

[37] E.A. Haroutunian, A.N. Harutyunyan and A.R. Ghazarian, “On rate-reliability-distortion function for a robust descriptions system”, *IEEE Transactions on Information Theory,* vol. 46, no. 7, pp. 2690-2697, 2000.

[38] E.A. Haroutunian and A.N. Harutyunyan, “Successive refinement of information with reliability criterion,” *in Proceedings of IEEE International Symposium on Information Theory,* p. 205, Sorrento, Italy, 2000.

[39] A.N. Harutyunyan, “Notes on conditions for successive refinement of information,” *Lecture Notes in Computer Science,* vol. 4123, "General Theory of Information Transfer and Combinatorics", Springer Verlag, pp. 154-164, 2006.

[40] R.F. Ahlswede and E.A. Haroutunian, “On logarithmically asymptotically optimal testing of hypotheses and identification*”, Lecture Notes in Computer Science,* vol. 4123, "General Theory of Information Transfer and Combinatorics", Springer Verlag, pp. 462-478, 2006.

[41] P. Hakobyan, “Solution of hypotheses testing and identification reliability problems for models with many objects”, *PhD thesis*, Yerevan, 2010.

[42] N. Grigoryan, “Optimal hypothesis testing in some models with discrete sources”, *PhD thesis*, Yerevan, 2010.

[43] A.O. Yessayan, “Statistical hypothesis testing concerning multiple objects”, *PhD thesis*, Yerevan, 2011.

[44] F. Hormozi-Nejad, “Investigation of two-stage reliable hypotheses detection concerning models characterized with several families of distributions”, *PhD thesis*, Yerevan, 2013.

[45] E.A. Haroutunian and A.O. Yessayan, “On reliability approach to multiple hypotheses testing and to identification of probability distributions of two stochastically related objects”, *in Proceedings of IEEE International Symposium on Information Theory,* Saint Petersburg, Russia, pp. 2607-2611, 2011.

[46] S.A. Tonoyan, “Information-theoretical investigation of some information hiding systems”, *PhD thesis*, Yerevan, 2007.

[47] M.E. Haroutunian and S.A. Tonoyan, “Random coding bound of information

hiding E-capacity", *in Proceedings of IEEE International Symposium on Information Theory,* Chicago, USA, pp. 536, 2004.

[48] A. Muradyan, "Reliability investigation of sources and channels with compound states", *PhD thesis*, Yerevan, 2010.

[49] N. Afshar, "Investigation of E-capacity of broadcast channels with confidential messages", *PhD thesis*, Yerevan, 2013.

[50] N. Afshar, E. Haroutunian and M. Haroutunian, "Random Coding Bound for *E*-capacity Region of the Broadcast Channel With Confidential Messages", *Comunication in Information and Systems,* vol. 12, no. 2, pp. 131-156, 2012.

[51] N. Afshar, "On E-capacity region of the wiretap channel", *Electronics Communication and Computer Engineering,* vol. 4, no. 5, 2013.

[52] T. M. Margaryan, "Investigation of information protection in Shannon cipher system", *PhD thesis,* Yerevan, 2014.

[53] N. Pahlevanyan, "Information theoretical analysis of biometric generated secret key sharing system and development of new package for R environment", *PhD thesis,* Yerevan, 2016.

Secure Communication for Networked Systems

Yanling Chen

University of Duisburg-Essen, Germany

Abstract

Communication networks have had a transformative impact on our society as they have revolutionized almost all aspects of human interaction. The explosive growth of data traffic has led to an increased demand on improving the reliability, efficiency and security aspects of the systems. The talk will start with some conventional cryptographic techniques that are used in real-life applications in ensuing a secure communication. From a system design perspective, information theoretic secrecy is introduced since it is more desirable than the security based on the computational complexity. Then, we look into some elemental communication networks, for instance, point to point channel, broadcast channel, multiple channel and two-way communication channels with an adversary, where the goal is to characterize the fundamental limits on the secure transmission rate regions. In particular, as an example, we discuss in more details on the secrecy over the multiple access channel. Impact of using different secrecy criteria is illustrated and the code mechanism by channel prefixing is shown to be helpful in enlarging the corresponding secrecy rate regions. At the end, we show the collective secrecy over the multiple access channel can be regarded as a secret sharing scheme in the communication setting.

Applications of Coding Theory to Biometrics

Anahit R. Ghazaryan[1] and A.J. Han Vinck[2]

[1] School no. 21, Ministry of Defense of Russian Federation, Yerevan, Armenia

[2] Institute of Digital Signal Processing, University of Duisburg-Essen, Germany

Abstract

We describe different approaches of data processing for biometric verification over noisy data. The key issues are privacy protection and good verification performance. The developed methods can be applied to biometric data received from noisy measurements.

1 Introduction

One of the basic problems in data processing is verification with noisy observations. In this paper we consider the biometric verification problem when the processed data represent outcomes of measurements of biometrical parameters of people. The general biometric verification procedure can be presented in the following way. There is a database (DB) containing records associated with certain people. *In the case of authentication*, the verifier, having received data from a person and the name of the person, has to accept the claimed identity or not. *In the case of identification*, the verifier, having received data from a person, has to construct a list of names of people that are considered as candidates for the name of the person. Both the records and the received data are formed after processing the outcomes of biometric measurements, and these measurements are noisy. The basic problem is: *how to convert data to the records at the enrollment stage and specify the verification algorithms.* The verification algorithms and evaluation of their performance should be extended in such a way that the probability of

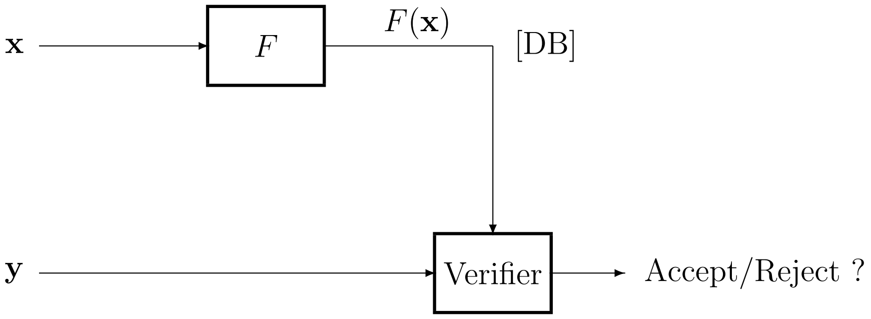


Figure 1: The structure of a general biometric authentication scheme.

successful passing through the verification stage with the acceptance decision will be very small for the non legitimate user. For the legitimate user the rejection decision should be low.

In this paper we present our main results on biometric authentication problem. Our results are developed in three directions. For the solution of the poblem we implement several coding schemes and construct codes having a low computational complexity.

2 Basic approaches

Let the outcomes of biometric measurements form the vector $\mathbf{x}$ at the enrollment stage and the vector $\mathbf{y}$ at the verification stage. Let both vectors have length n, and let the record stored in the DB be a vector $F(\mathbf{x})$ of length k, which can be understood as a password of the person whose biometric measurements are expressed by the vector $\mathbf{x}$. The transformation

$$\mathbf{x} \to F(\mathbf{x}) \tag{1}$$

will be referred to as the encoding of the vector $\mathbf{x}$. Without loss of generality, we can consider authentication as a procedure consisting of two steps. The verifier first maps

$$(F(\mathbf{x}), \mathbf{y}) \to \Delta(F(\mathbf{x}), \mathbf{y}) \in [0, 1],$$

and then accepts the identity claim if and only if the result belongs to the $[\delta_0, \delta_1]$ interval, where δ_0, δ_1 are parameters fixed in advance. The structure of a general biometric authentication scheme is given in Figure 1.

We consider three directions developed to solve the authentication problem.

P1: Direct authentication. The basic problem is: *how to convert outcomes of measurements to digital format and assign a distortion function?* We assume that parameter k is proportional to n and F is a deterministic function.

P2: Randomized block coding schemes. The basic problem is: *how to protect a verification scheme against attacks?* Let the parameter k be proportional to n and F is a function, which uses an extra randomness.

P3: Deterministic block coding schemes. The basic problem is: *how to compress data and organize an efficient verification?* We assume that the parameter k is much less than n and F is a deterministic function.

Our analysis of possible applications includes data received from the fingerprints and the DNA sequences.

2.1 Direct authentication

If the outcomes of the measurements characterize a person by some list of biometric parameters measured in different units, like kilograms, centimeters, coordinates of minutiae points of the plane, etc., one has to find an artificial space where components of the vectors $\mathbf{x}$ and $\mathbf{y}$ can be compared. We introduce a mathematical model of non-stationary random processes assuming that components of these vectors are values of random variables with specified probability distributions (PD's) $F_1, \ldots, F_n$, and different components are assumed to be independent. The value of the PD always belongs to the interval $[0,1]$. This observation allows us to unify parameters: the values of $x_1, \ldots, x_n$ are considered as arguments of the corresponding PD's and they are replaced by the values of the functions $F_1(x_1), \ldots, F_n(x_n)$. The encoding is defined as the result of uniform quantization of the components of the vector $(F_1(x_1), \ldots, F_n(x_n))$ in q levels, which can be expressed as $F(\mathbf{x}) = (\lfloor qF_1(x_1) \rfloor, \ldots, \lfloor qF_n(x_n) \rfloor)$. Similar transformations are applied to map the vector $\mathbf{y}$ to the vector $F(\mathbf{y}) = (\lfloor qF_1(y_1) \rfloor, \ldots, \lfloor qF_n(y_n) \rfloor)$.

An assumption about the independence of different biometric parameters that are measured to form the vectors $\mathbf{x}$ and $\mathbf{y}$ implies the independence of the pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. However, these pairs are not identically distributed. As $F_1, \ldots, F_n$ are the PD's, our transformations conserve the independence of the pairs of new random variables $(\lfloor qF_1(X_1) \rfloor, \lfloor qF_1(Y_1) \rfloor), \ldots, (\lfloor qF_n(X_n) \rfloor, \lfloor qF_n(Y_n) \rfloor)$ and add the property that these pairs are identically distributed. This point is important, because we want to introduce an additive cumulative distortion function that evaluates the contributions of different components to a number $\Delta(F(\mathbf{x}), \mathbf{y})$, which is used to make a decision about the independence of the processes $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$. Notice that searching for an additive distortion function is justified by the wish of reducing the analysis of the performance of the algorithm to the analysis of the sums of i.i.d. random variables. In this case, a standard probability theory technique brings the statement that the probability of error is an exponentially decreasing function of n, when we test the $(F_1, \ldots, F_n)$-independence. Hence we get an upper bound on the false acceptance rate that exponentially decreases with n for the biometric authentication problem.

We introduce the distortion function between the t-th components of the vectors $F(\mathbf{x})$ and $F(\mathbf{y})$ as the absolute value of their difference, $|\, \lfloor qF_t(x_t) \rfloor - \lfloor qF_t(y_t) \rfloor \,|$ and the cumulative distortion as the sum of these values over $t = 1, \ldots, n$. If the hypotheses that the pair (x_t, y_t) was generated by independent random processes X_t and Y_t having the same PD F_t, then both $\lfloor qF_t(X_t) \rfloor$ and $\lfloor qF_t(Y_t) \rfloor$ are independent discrete random variables uniformly distributed over the set $\{0, \ldots, q-1\}$ for all $t = 1, \ldots, n$. Otherwise, if the $(F_1, \ldots, F_n)$-independence hypothesis is not true, these distributions are different, and the verifier can distinguish between these two cases. The thresholds on the value of the cumulative distortion are fixed in a way to guarantee the false rejection rate at a certain level.

Some results of the analysis of direct block coding schemes are included in [1] and [2].

2.2 Randomized block coding schemes

To implement a biometric authentication scheme, the templates of a group of people are stored in the DB under the names of these people. These templates should be protected against attacks for discovery the biometrics and attacks for successful passing through the verification test. The randomized block coding schemes for biometric authentication are oriented to the protection of the stored data from attacks of these types. In this case, the biometric vector $\mathbf{x}$ is considered as a noise that corrupts a randomly chosen codeword

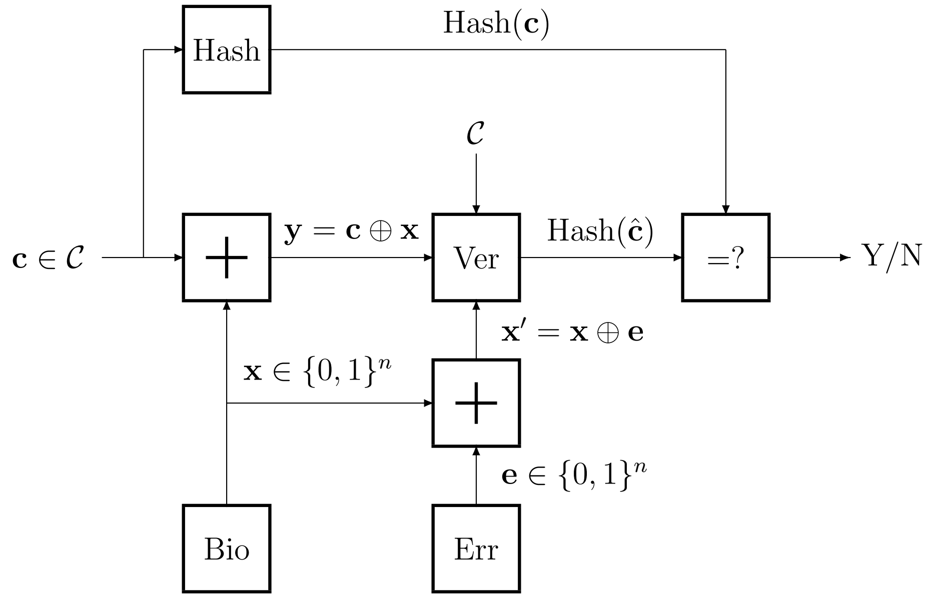

Hash
Hash($\mathbf{c}$)
$\mathcal{C}$
$\mathbf{c} \in \mathcal{C}$
+
$\mathbf{y} = \mathbf{c} \oplus \mathbf{x}$
Ver
Hash($\hat{\mathbf{c}}$)
=?
Y/N
$\mathbf{x}' = \mathbf{x} \oplus \mathbf{e}$
$\mathbf{x} \in \{0,1\}^n$
+
$\mathbf{e} \in \{0,1\}^n$
Bio
Err

Figure 2: Verification of a person using an additive block coding scheme.

of a binary block code. The known classes of schemes include *additive block coding schemes* and *permutation block coding schemes*.

An additive block coding treats the biometric vector $\mathbf{x}$ as an additive noise that corrupts information about the randomly chosen key $m \in \{0,\ldots,M-1\}$ that has to be delivered to the destination. The key is transformed to the key codeword $\mathbf{c} = \mathbf{c}(m)$ belonging to a binary block code $\mathcal{C}$ for M messages, and the vector $\mathbf{y} = \mathbf{c} \oplus \mathbf{x}$ is stored in the DB under the name of the person, i.e., $F(\mathbf{x}) = \mathbf{y}$ and $k = n$. Suppose that the vector $\mathbf{y}$ is used to verify the claimed identity of a person whose biometric vector is given as $\mathbf{x}' = \mathbf{x} \oplus \mathbf{e}$, where $\mathbf{e} \in \{0,1\}^n$ is the noise vector. If $\hat{\mathbf{c}} \in \mathcal{C}$ is the result of the decoding of the key codeword on the basis of the pair of vectors $(\mathbf{y}, \mathbf{x}')$, one can check whether the value of Hash($\mathbf{c}$), delivered from an auxiliary storage, is equal to Hash($\hat{\mathbf{c}}$) or not, where Hash is a fixed hash function. The claimed identity of a person is then either accepted or rejected. Verification of a person using an additive block coding scheme is illustrated in Figure 2.

A conventional approach to solving the authentication problem by using an additive block coding scheme was proposed Juels and Wattenberg [3], and

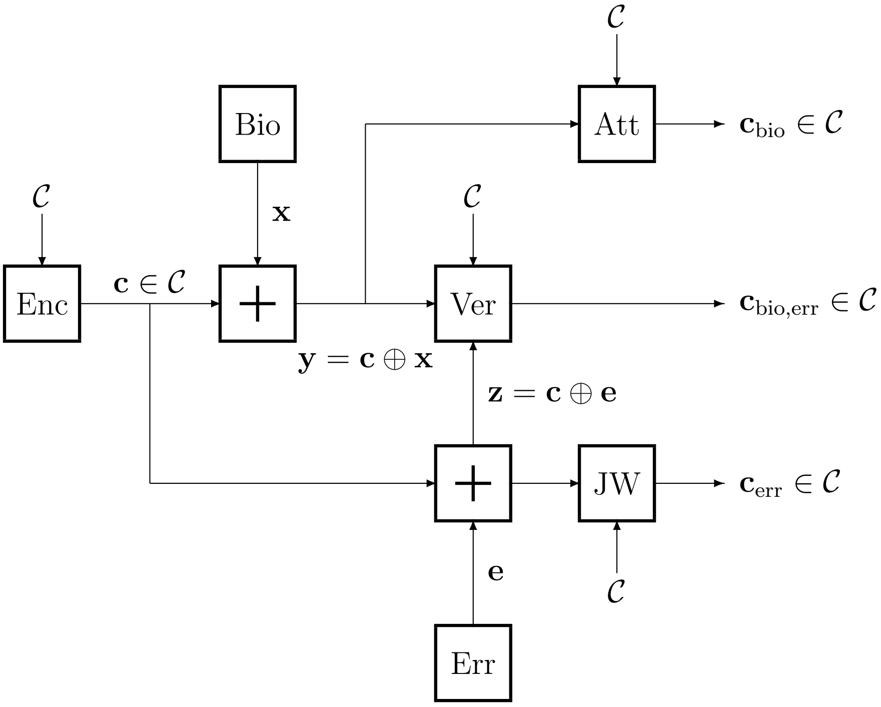

$\mathcal{C}$
Bio
Att
$\mathbf{c}_{\text{bio}} \in \mathcal{C}$
$\mathcal{C}$
$\mathbf{x}$
$\mathcal{C}$
Enc
$\mathbf{c} \in \mathcal{C}$
$+$
Ver
$\mathbf{c}_{\text{bio,err}} \in \mathcal{C}$
$\mathbf{y} = \mathbf{c} \oplus \mathbf{x}$
$\mathbf{z} = \mathbf{c} \oplus \mathbf{e}$
$+$
JW
$\mathbf{c}_{\text{err}} \in \mathcal{C}$
$\mathbf{e}$
$\mathcal{C}$
Err

Figure 3: Transmission of a key codeword over two parallel channels.

it is known as the JW algorithm. This algorithm is based on the observation:

$$\left.\begin{array}{l} \mathbf{y} = \mathbf{c} \oplus \mathbf{x} \\ \mathbf{x}' = \mathbf{x} \oplus \mathbf{e} \end{array}\right\} \Rightarrow \ \mathbf{c} \oplus \mathbf{e} = \mathbf{z},$$

where $\mathbf{z} = \mathbf{y} \oplus \mathbf{x}'$. The JW decoder tries to recover the vector $\mathbf{c}$ from the received vector $\mathbf{y} \oplus \mathbf{x}'$ using error–correcting capabilities of a code. In particular, any $\lfloor (d_{\mathcal{C}} - 1)/2 \rfloor$ errors, where $d_{\mathcal{C}}$ is the minimum distance of the code $\mathcal{C}$, will be corrected.

We observed that

$$\left.\begin{array}{rcl} \mathbf{y} & = & \mathbf{c} \oplus \mathbf{x} \\ \mathbf{x}' & = & \mathbf{x} \oplus \mathbf{e} \end{array}\right\} \Rightarrow \left\{\begin{array}{rcl} \mathbf{c} \oplus \mathbf{x} & = & \mathbf{y}, \\ \mathbf{c} \oplus \mathbf{e} & = & \mathbf{z}, \end{array}\right.$$

i.e., the verifier receives the pair of vectors $(\mathbf{c} \oplus \mathbf{x}, \mathbf{c} \oplus \mathbf{e})$, while the attacker receives only the first component of the pair and the JW decoder analyzes only the second component of the pair. Therefore the biometric authentication with an additive block coding scheme can be represented as transmission of the codeword **c** over two parallel channels (see Figure 3). An attacker and a conventional verifier can be viewed as decoders having access to the first and to the second line of the channel, respectively. We present general formulas for the probability of correct decoding and demonstrate an improvement in the performance of the decoding algorithm for both specific and random block codes as compared to the standard Juels–Wattenberg verification scheme [3]. We introduce a mathematical model for DNA measurements and present some numerical results illustrating the correction of errors for the DNA measurements of a legitimate user and protection of templates against attacks for successful passing the verification stage by an attacker.

The permutation block coding schemes are developed for the increase of the security of the system. They are based on the use of a binary code where all codewords have the same Hamming weight and assume that the biometric data are also represented by binary vectors of this weight. In the permutation block coding scheme, a randomly chosen permutation that transforms randomly chosen codeword **c** to the biometric vector **x** is used. By a proper assignment of the PD over the set of permutations that transform two binary vectors of the same weight to each other, it is possible to reach exactly the same secrecy of the coded system as the secrecy of the blind guessing the biometric vector, when the attacker does not have access to the database.

Some results of the analysis of randomized block coding schemes are included in [4]-[6].

2.3 Deterministic block coding schemes

We consider the data processing scheme where an input binary vector $\mathbf{x}$ of length n is mapped to a binary vector $\mathrm{Hash}(\mathbf{x})$ of length $k << n$, where Hash is a fixed hash function. The result is stored in the DB. Having received an observation binary vector $\mathbf{y}$ of length n and using the content of the DB, a verifier has to decide whether the observation vector can be considered as a corrupted version of the input vector $\mathbf{x}$ or not. The Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ serves as a measure of closeness of these vectors. The task is of interest for different applications. One of them is the biometric authentication: if $\mathbf{x}$ represents biometric measurements that were used to form the DB and $\mathbf{y}$ represents measurements of the same parameters of a person, who claims identity, then an authentication scheme has to accept the claim or not. An important requirement leading to the use of hashing is the point that an attacker should not discover the vector $\mathbf{x}$, since the biometrics, being compromised, is compromised forever.

A conventional implementation of hashing with noisy data is based on the analysis of the pair $(\mathrm{Hash}(\mathbf{x}), \mathrm{Hash}(\mathbf{y}))$ that can be efficient only under strong constraints on the available hash functions.Alternatively, we analyze the pair $(\mathrm{Hash}(\mathbf{x}), \mathbf{y})$ and present a combinatorial algorithm for transformation of binary biometric vector of lenght n to short codewords of length k. For the case $n = 10$ Kbytes, the code rate has the order of magnitude 10^{-4}, the decision about the closeness of strings has to be made if the Hamming distance is less than 4096, we demonstrate that the decoding error probabilities that can be attained by the algorithm have the order of magnitude 2^{-17}.

The evaluations of the compression factor, the false rejection/acceptance rates are derived and an illustration of a possible implementation of the verification algorithm for the DNA data is presented.

Some results of the analysis of deterministic block coding schemes are included in [7] and [8].

References

[1] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Testing the independence of two non–stationary random processes with applications to biometric authentication", *Proc. ISIT'2007*, Nice, France, June 24–29, pp. 2671–2675, 2007.

[2] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "An algorithm for biometric authentication based on the model of nonstationary random processes", *Lecture Notes in Computer Science: Advances in Biometrics*, no. 4642, pp. 319–327, 2007.

[3] A. Juels, M. Wattenberg, "A fuzzy commitment scheme," *Proc. ACM Conf. Computer and Communication Security*, pp. 28–36, 1999.

[4] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Additive block coding schemes for biometric authentication with the DNA data', *Proc. 1-st European Workshop on Biometrics and Identity Management*, Roskilde University, Denmark, pp. 164–173, Revised Selected Papers, Springer-Verlag Berlin Heidelberg 2008, 2008.

[5] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Block coding schemes designed for biometric authentication", Invited chapter in *Biometrics: Methods, Applications and Analysis*, NOVA Publishers, pp. 217–240, 2010.

[6] V. B. Balakirsky and A. J. Han Vinck, "Block coding schemes designed for biometric authentication", Invited chapter in *Advanced Biometric Technologies*, InTech Publisher, pp. 299–324, 2011.

[7] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, "Mathematical model for constructing passwords from biometrical data", *Security and Communication Networks*, Wiley, pp. 1–9, 2009.

[8] V. B. Balakirsky and A. J. Han Vinck, "A simple scheme for constructing fault-tolerant passwords from biometric data", *Eurasip Journal on Information Security*, no. 2010, 11 pages, 2010.

Sequential Hashing for Noisy Verification

Vladimir B. Balakirsky*and Anahit R. Ghazaryan[1]

[1] School no. 21, Ministry of Defense of Russian Federation, Yerevan, Armenia

Abstract

We present an efficient sequential data processing algorithm, which maps the input float-valued vector of length nL to a vector of length L, whose components belong to a finite set $\mathcal{Q}$. The constructed vector is interpreted as a password associated with input data. Given a noisy version of the input data, the verifier has to decide whether it is a corrupted version of the original data or not. The proposed algorithm provides the decoding error probability decreasing with n.

1 Introduction

The problem of construction of fault–tolerant passwords is one of directions in computer science, which attracts great attention (see [2]-[4], [6]). Another relevant direction is an extraction of secret key from noisy observations, addressed in many papers in information theory (see [5], [6]). In [1] we dealt with these directions and presented some variant of the verification scheme where the input float-valued vector of length nL is splitted into L blocks of length n and each block is mapped to one bit using block-by-block transformation. Our verification scheme [1] provided the decoding error probability decreasing with the length n of the block. In this paper we consider the setup of [1] when the input data consisting of nL floats. This string is partitioned in L blocks of length n: $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \in (\mathsf{R}^n)^L$, and each block is mapped by a fixed deterministic function f to some value from the finite

*Deceased

set $\mathcal{Q}$: $f(\mathbf{x}) = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_L)) \in \mathcal{Q}^L$. This transformation is considered as the encoding or hashing. Having received another string $\mathbf{y}$, consisting of nL floats, the verifier constructs the estimate of $f(\mathbf{x})$ as $\hat{f}(\mathbf{y}) \in \mathcal{Q}^L$. This transformation is considered as the decoding. The inequality $\hat{f}(\mathbf{y}) \neq f(\mathbf{x})$ is interpreted as the decoding error in the case when the string $\mathbf{y}$ is a corrupted version of the input string $\mathbf{x}$. In this paper we consider the basic setup for the data processing scheme when the verifier has access to the database and has to decide whether the data $\mathbf{y}$ received at the verification stage can be considered as a noisy version of the data $\mathbf{x}$ processed at the enrollment stage. Such a claim will be viewed as "the identity claim".

Given the input float-valued string $\mathbf{x}$, the decoding error probability, computed over the probabilistic ensemble, which describes the noise of the measurements, has to be small. On the other hand, the decoding error probability, computed over the ensemble where the received string is independent of the input string, has to be large. In this paper we construct combinatorial encoding (hashing) and decoding algorithms with using sequential strategy at the encoder and Bayes estimation at the decoder. Our algorithms provide decoding error probability decreasing with the length n of the block for the fixed length L of the password.

2 Notation and Basic Ideas

We will consider the situation when the input data are presented as a string consisting of nL floats. This string is partitioned in L blocks of length n and expressed as the vector

$$\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_L) \in (\mathsf{R}^n)^L,\ \mathbf{x}_\ell = (x_\ell^1, \dots, x_\ell^n),\ x_\ell^1, \dots, x_\ell^n \in \mathsf{R},\ l = \overline{1, L}. \quad (1)$$

Suppose that $\mathsf{Q} \geq 1$ is the given even integer and denote $\mathcal{Q} = \{0, \dots, Q-1\}$. Let $f, \hat{f} : \mathsf{R}^n \to \mathcal{Q}$ be two fixed deterministic functions that map float-valued vectors of length n to some value from the set $\mathcal{Q}$.

Let us map each block $\mathbf{x}_\ell$ of the string $\mathbf{x}$ presented in (1) to $f(\boldsymbol{x}_\ell) \in \mathcal{Q}$. The transformation $\boldsymbol{x}_\ell \to f(\boldsymbol{x}_\ell)$ is called encoding or hashing. An application of the block-by-block hashing to the string $\mathbf{x}$ brings the vector

$$f(\mathbf{x}) = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_L)) \in \mathcal{Q}^L,$$

which is considered as the password of the input data $\mathbf{x}$.

Denote by $\mathrm{ID}(\mathbf{x})$ a unique identifier of the person whose data are accumulated in the input string $\mathbf{x}$. Suppose that the input strings are stored in a database under the identifiers of the corresponding persons.
Let the verifier receives a string $\mathbf{y}$ consisting of nL floats and the claim that it is a noisy version of the input string $\mathbf{x}$. Then the verifier

- partitions the string $\mathbf{y}$ in L blocks of length n:

$$\mathbf{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L) \in (\mathsf{R}^n)^L,\ \mathbf{y}_\ell = (y_\ell^1, \ldots, y_\ell^n),\ y_\ell^1, \ldots, y_\ell^n \in \mathsf{R},\ l = \overline{1, L},$$

- constructs the estimate of the received string $\mathbf{y}$ using function $\hat{f}$:

$$\hat{f}(\mathbf{y}) = (\hat{f}(\boldsymbol{y}_1), \ldots, \hat{f}(\boldsymbol{y}_L)) \in \mathcal{Q}^L,$$

- constructs the vector $\varepsilon(\mathbf{y}) = (\varepsilon(\boldsymbol{y}_1), \ldots, \varepsilon(\boldsymbol{y}_L)) \in [0, 1/2]^L$ characterizing the reliability of the constructed estimate $\hat{f}(\mathbf{y})$,

- computes

$$\Lambda = \prod_{j=1}^{L} \begin{cases} 1 - \varepsilon(\boldsymbol{y}_j), & \text{if } \hat{f}(\boldsymbol{y}_j) = f(\boldsymbol{x}_j), \\ \varepsilon(\boldsymbol{y}_j), & \text{if } \hat{f}(\boldsymbol{y}_j) \neq f(\boldsymbol{x}_j). \end{cases}$$

If Λ is less than a fixed threshold value Λ^*, then the identity claim is rejected. Otherwise, the claim is accepted.

3 Probabilistic Ensembles

Suppose that $X^n = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are independent identically distributed continuous random variables generated by a memoryless source specified by the probability density function (PDF) $(\, P(x), x \in \mathsf{R}\,)$. Then the PDF, associated with the realization $X^n = \boldsymbol{x}$, is expressed as

$$P(\boldsymbol{x}) = \prod_{j=1}^{n} P(x_j). \tag{2}$$

Let $\boldsymbol{y}$ be a vector received as the outcome of the memoryless noisy observations of the input vector specified by the conditional PDF's

$$(\, V(y|x), y \in \mathsf{R}\,),\ x \in \mathsf{R}.$$

Then the conditional PDF, associated with the realization $Y^n = \boldsymbol{y}$, given the input vector $X^n = \boldsymbol{x}$, is equal to

$$V(\boldsymbol{y}|\boldsymbol{x}) = \prod_{j=1}^{n} V(y_j|x_j). \tag{3}$$

Let the function f be assigned in such a way that

$$\int_{\boldsymbol{x}\in\mathsf{R}^n} P(\boldsymbol{x})\chi\{f(\boldsymbol{x}) = z\} = \int_{\boldsymbol{x}\in\mathsf{R}^n} P(\boldsymbol{x})\chi\{f(\boldsymbol{x}) = \mathsf{Q} - 1 - z\}, \quad z \in \mathcal{Q}, \tag{4}$$

where χ denotes the indicator function: $\chi\{\mathcal{S}\} = 1$ if the statement $\mathcal{S}$ is true and $\chi\{\mathcal{S}\} = 0$ otherwise.
Let $(\,\Phi(z, \boldsymbol{y}),\ z \in \mathcal{Q},\ \boldsymbol{y} \in \mathsf{R}^n\,)$ denote the PDF associated with the pair of random variables $(Z = f(X^n), Y^n)$. Then

$$\Phi(z, \boldsymbol{y}) = \int_{\boldsymbol{x}\in\mathsf{R}^n} P(\boldsymbol{x})V(\boldsymbol{y}|\boldsymbol{x})\chi\{f(\boldsymbol{x}) = z\}. \tag{5}$$

Let us also denote the conditional PDF associated with the random variable Z, given the value $\boldsymbol{y}$ of the random variable Y^n, by

$$\Psi(z|\boldsymbol{y}) = \frac{\Phi(z, \boldsymbol{y})}{Q(\boldsymbol{y})}, \tag{6}$$

where

$$Q(\boldsymbol{y}) = \int_{\boldsymbol{x}\in\mathsf{R}^n} P(\boldsymbol{x})V(\boldsymbol{y}|\boldsymbol{x}).$$

4 Decoding Error Probability

Given a function f, satisfying ((4), the decoding error probability can be expressed as

$$\begin{aligned}\Lambda(\hat{f}) &= \Pr\{\hat{f}(Y^n) \neq f(X^n)\} \\ &= \int_{\boldsymbol{x},\boldsymbol{y}\in\mathsf{R}^n} P(\boldsymbol{x})V(\boldsymbol{y}|\boldsymbol{x})\chi\{\hat{f}(\boldsymbol{y}) \neq f(\boldsymbol{x})\}.\end{aligned}$$

Using (5) and the fact that

$$\chi\{\hat{f}(\boldsymbol{y}) \neq f(\boldsymbol{x})\} = \sum_{z\in\mathcal{Q}} \chi\{\hat{f}(\boldsymbol{y}) \neq z\}\chi\{f(\boldsymbol{x}) = z\}, \tag{7}$$

we can write

$$\begin{aligned}\Lambda(\hat{f}) &= \int_{\boldsymbol{y}\in \mathsf{R}^n} \sum_{z\in\mathcal{Q}} \chi\{\hat{f}(\boldsymbol{y}) \neq z\} \cdot \int_{\boldsymbol{x}\in \mathsf{R}^n} P(\boldsymbol{x})V(\boldsymbol{y}|\boldsymbol{x})\chi\{f(\boldsymbol{x}) = z\} \\ &= \int_{\boldsymbol{y}\in \mathsf{R}^n} \sum_{z\in\mathcal{Q}} \Phi(z,\boldsymbol{y})\chi\{\hat{f}(\boldsymbol{y}) \neq z\} \geq \Lambda^\star,\end{aligned}$$

where

$$\Lambda^\star = \int_{\boldsymbol{y}\in \mathsf{R}^n} \min_{z\in\mathcal{Q}} \Phi(z,\boldsymbol{y}). \tag{8}$$

The equality $\Lambda(\hat{f}) = \Lambda^\star$ is attained if and only if

$$\hat{f}(\boldsymbol{y}) = \arg\max_{z\in\mathcal{Q}} \Phi(z,\boldsymbol{y}) = \arg\max_{z\in\mathcal{Q}} \Psi(z|\boldsymbol{y}) \tag{9}$$

corresponds to the maximum *a posteriori* probability decoding.
Remark that (8) can be expressed as

$$\Lambda^\star = \int_{\boldsymbol{y}\in \mathsf{R}^n} Q(\boldsymbol{y})\varepsilon(\boldsymbol{y}), \tag{10}$$

where

$$\varepsilon(\boldsymbol{y}) = \frac{\min_{z\in\mathcal{Q}} \Phi(z,\boldsymbol{y})}{Q(\boldsymbol{y})} = \min_{z\in\mathcal{Q}} \Psi(z|\boldsymbol{y}). \tag{11}$$

Therefore the reliability of $\hat{f}(\boldsymbol{y})$ satisfying (9) is characterized by (11).

5 The encoding/decoding algorithms

Let $(T_1^0,\ldots,T_1^{\mathsf{Q}-1}),\ldots,(T_n^0,\ldots,T_n^{\mathsf{Q}-1})$ be a fixed sequence of floats constructed in such a way that

$$\int_{x\leq T_j^z} P(x) = \int_{x\geq T_j^{\mathsf{Q}-1-z}} P(x),\ z\in\mathcal{Q} \tag{12}$$

for all $j = 1,\ldots,n$ and

$$T_n^z = T_n^{\mathsf{Q}-1-z},\ z\in\mathcal{Q}. \tag{13}$$

For all $j = 1,\ldots,n$, denote

$$\boldsymbol{x}_j = (x_1,\ldots,x_j),\ \ \boldsymbol{y}_j = (y_1,\ldots,y_j)$$

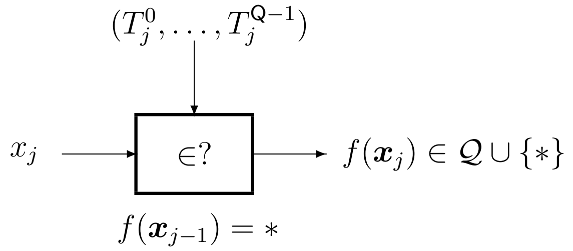

$(T_j^0, \ldots, T_j^{\mathsf{Q}-1})$
x_j
$\in ?$
$f(\boldsymbol{x}_j) \in \mathcal{Q} \cup \{*\}$
$f(\boldsymbol{x}_{j-1}) = *$

Figure 1: The structure of the j-th step of the hashing algorithm.

and notice that

$$\boldsymbol{x} = \boldsymbol{x}_n, \;\; \boldsymbol{y} = \boldsymbol{y}_n.$$

- **The encoding (hashing) algorithm**

H1: Set $j = 1$.

H2: Set

$$f(\boldsymbol{x}_j) = \begin{cases} 0, & \text{if } x_j \le T_j^0, \\ 1, & \text{if } T_j^0 \le x_j \le T_j^1, \\ \ldots & \\ Q-2, & \text{if } T_j^{Q-3} \le x_j \le T_j^{Q-2}, \\ *, & \text{if } T_j^{Q-2} \le x_j \le T_j^{Q-1}, \\ Q-1, & \text{if } x_j \ge T_j^{\mathsf{Q}-1}. \end{cases}$$

If $f(\boldsymbol{x}_j) = *$, then increase j by 1 and go to **H2**.

H3: Output

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_j).$$

The structure of the j-th step of the hashing algorithm is illustrated in Figure 1.

- **The decoding algorithm**

D1: Set $j = 1$ and $\Psi(z|\boldsymbol{y}_0) = 0$ for all $z \in \mathcal{Q}$.

D2: For all $j = 1, \ldots, n$, $z \in \mathcal{Q}$, compute

$$\Psi(z|\boldsymbol{y}_j) = \Psi(z|\boldsymbol{y}_{j-1}) + c^*(\boldsymbol{y}_{j-1})\Psi(z|y_j),$$

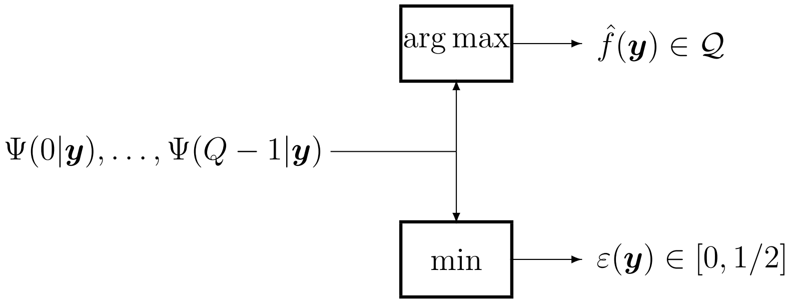

arg max
$\hat{f}(\boldsymbol{y}) \in \mathcal{Q}$
$\Psi(0|\boldsymbol{y}), \ldots, \Psi(Q-1|\boldsymbol{y})$
min
$\varepsilon(\boldsymbol{y}) \in [0, 1/2]$

Figure 2: The structure of the decoding algorithm.

where

$$\Psi(0|y_j) = \int_{x \le T_j^0} W(x|y_j),$$

$$\Psi(z|y_j) = \int_{T_j^{z-1} \le x \le T_j^z} W(y_j|x), \quad z \in \mathcal{Q} \setminus \{0, Q-1\}$$

$$\Psi(Q-1|y_j) = \int_{x \ge T_j^{Q-1}} W(x|y_j),$$

$$c^*(\boldsymbol{y}_{j-1}) = 1 - \sum_{z \in \mathcal{Q}} \Psi(z|\boldsymbol{y}_{j-1}).$$

D3: Output $\hat{f}(\boldsymbol{y})$ and $\varepsilon(\boldsymbol{y})$ using (9) and (11), respectively.

The structure of the decoding algorithm is illustrated in Figure 2.

The main features of the encoding/decoding algorithms are given in the following statement:

Proposition. *The decoding error probability* Λ *can vanish when* n *increase for a certain sequence of thresholds*

$$(T_1^0, \ldots, T_1^{\mathsf{Q}-1}), \ldots, (T_n^0, \ldots, T_n^{\mathsf{Q}-1})$$

satisfying (12) *and* (13).

The proposition leads to the following conclusion. The length of the password of the input string is fixed. However, the reliability of reconstruction of the vector using a noisy version of the string can be made arbitrary small by increasing the length of the string.

6 Conclusion

In this paper the sequential (by the encoder) data processing algorithm is proposed. The presented algorithm allows us to use the fact that there are more than one available observation. Our algorithm provides decoding error probability decreasing with the length n of the block for the fixed length L of the password. For the case $n = 2$ and Gaussian PDF's we can find the values of the thresholds satisfying (12) and (13). As a result, we conclude that the decoding error probability is decreased by about 75 percents as compared to the case $n = 1$.

References

[1] Balakirsky V. B., Ghazaryan A. R.(2011) Adaptive fault–tolerant hashing algorithms for noisy verification, *Proc. 8-th International Conference on Computer Science and Information Technologies*, Yerevan, Armenia.

[2] Balakirsky V. B. (2005) Hashing of Databases with the Use of Metric Properties of the Hamming Space, *The Computer Journal*, 48 (1), 4-16, doi:10.1093/comjnl/bxh059.

[3] Tuyls P., Scoric B., T. Kavenaar (2007) *Security with Noisy Data: Private Biometrics, Secure Key Storage and Anti–Counterfeiting.* Springer-Verlag, London.

[4] Frykholm N., Juels A. (2001) Error tolerant password recovery, *Proceedings of the 8th ACM Conference on Computer and Communication Security*, Philadelphia, Pa, USA, 1-9.

[5] Dodis Y., Reyzin L., and Smith A. (2004) Fuzzy extractors: how to generate strong keys from biometrics and other noisy data, *Lecture Notes in Computer Science*, Springer-Verlag, vol. 3027, 523-540.

[6] Juels A. and Wattenberg M. (1999) A fuzzy commitment scheme, *Proceedings of the 6th ACM Conference on Computer and Communication Security*, New York, NY, USA, 28 - 36.

Error Detection: The Past, the Present, and the Future

Yanling Chen

University of Duisburg-Essen, Germany

Abstract

This paper discusses the history and development of the check digit system from ancient time to today's big data era. Furthermore, new challenges in both theory and practice are proposed.

1 A bit history

The first error-detecting code was at least traced back to 135CE, by Jewish scribes working to produce copies of the Torah. As Spanish/Jewish scholar Maimonidies (better known as Rambam) said in the 12th century: *A Torah scroll missing even one letter is invalid.* Thus it was of critical importance that all new copies of the Torah and its associated writings be identical to the old copies. A system was developed, containing statistics and applying gematraia to entire pages, in order to provide a kind of check digit system. In this way, the accuracy of pages at a time could easily be checked, and thus of chapters, and, eventually, the entire Torah.

We human beings, tend to make mistakes of certain patterns, when we interact with data (often numbers or characters or their combinations). The most common errors made by human operators, as well their relative frequency of occurrence, are listed in Table 1, which is according to the statistical investigations by D. F. Beckley [B67] and J. Verhoeff [V69] (in a Dutch postal office 12,112 pairs, 6 digits).

The whole philosophy of error detection/correction is based on the assumption that the errors which a code is designed to detect/correct are very frequent compared with the errors not guarded against by it. In fact, the

Table 1: Error types and their frequencies [S00].

Error patterns	Description in symbol	Frequency in %	
		Verhoeff	Beckley
1) single error	$\cdots a \cdots \ \to \cdots b \cdots$	79.0	86
2) transposition	$\cdots ab \cdots \ \to \cdots ba \cdots$	10.2	8
3) jump transposition	$\cdots acb \cdots \ \to \cdots bca \cdots$	0.8	
4) twin error	$\cdots aa \cdots \ \to \cdots bb \cdots$	0.6	6
*) phonetic error ($a \geq 2$)	$\cdots a0 \cdots \ \to \cdots 1a \cdots$	0.5	
5) jump twin error	$\cdots aca \cdots \ \to \cdots bcb \cdots$	0.3	
other errors		9.1	

work of Verhoeff [V69] was also the first significant publication on check digit systems, which presented decimal codes known in the 1970s. In particular, Verhoeff [V69] also proposed a decimal code over the Dihedral group $\mathbb{D}_5$ together with an appropriate permutation. Surprisingly, his design still remains one of the best error-detecting schemes over decimal numbers.

In 1974, when Alan Haberman pioneered the use of a UPC barcode in his grocery business, he kicked off not only a revolution in inventory management and the supply chain but also made a step towards the metadata overlays of physical items and information. Several decades later, numerous identification systems with one or more check digits have been developed and used in different business applications, serving as an indispensable and ubiquitous role in our daily life. Beyond all doubt, a smart choice of the check digit system will ensure not only the data integrity, but also reduces the ARQ demands and thus boost the efficiency of the information flow.

2 State of Art

To check the integrity of the transmitted information, usually one or more check digits are appended to the information sequence such that they satisfy a relation defined by a check equation. If the check equation is not true then there is definitely an error present. If the check equation is true, however, one or more errors may still be present (with very small probability). Often this is sufficient since the operation can be repeated once an error is detected (known as ARQ, i.e., Automatic Repeat-reQuest). Unquestionably, the earlier the error is detected, the less it may cost in practice. Therefore, it is desirable to design check digit systems which can automatically detect all of the listed errors, or, if infeasible, at least those with higher occurrence frequencies. The objective is to minimize the probability of the undetected errors as listed in Table 1, thus reducing the ARQ demands and eventually improving the efficiency of the information flow. Most commonly, the systems in the literature and in practice are defined over alphabets endowed with a group structure, and they can be classified into the following categories.

(a) *Division method:* Such a system requires divisions of large integers. In general, it does not detect all the single errors, i.e., type 1).

(b) *Parity check method:* Such a system detects all the 1) single errors but not the 2) transposition errors.

(c) *Weighted parity check method:* This method imposes a weight on every digit position before conducting the modulo sum. It detects all the 1) single errors. With a carefully chosen weight sequence, it may detect most of the transposition errors of types 2)-3).

(d) *Polynomial method:* This is a special case of the weighted parity check method, in the manner that its weight sequence is a geometric progress. An example is the MOD 11-2 system as specified in ISO/IEC 7064 [ISO7064], and this specific example detects all the errors of types 1)-5). Other instances include the hexadecimal system in [N11] and its generalization in [C0].

(e) *Luhn formula:* This is a special system also known as the "IBM check", which is officially specified in [ISO7812, Annex B]. It detects all the errors of types 1) and 3).

(f) *Hybrid system:* A hybrid system MOD $m, m+1$ is specified in accordance with ISO/IEC 7064 [ISO7064]. It is hybrid since it involves both modular m and modular $m+1$. It detects all the single errors and most of other error types.

(g) *Verhoeff's method:* It is named after Verhoeff [V69] who built a decimal code based on the noncommutativity of the Dihedral group $\mathbb{D}_5$ operation and an appropriate permutation. The method is powerful and easy to implement. It was rediscovered and generalized by Gumm in [G85] to a code with alphabet of size $2s$, s odd. In general, it detects all the errors of types 1)-2).

Some concrete examples with more detailed information on check equation, group/quasigroup operation and alphabet size are listed in Table 2. For their performance on error detection, one can refer to Table 3.

Table 2: Check equations in different methods.

Check digit systems [2]	Check equations[1]	Operating Group	Alphabet size
(a) IBAN	$x_1x_2\cdots x_nx_{n+1} \equiv 1 \pmod{97}$	$\mathbb{Z}_{97}$	10
(b) Euro banknotes	$\sum_{i=1}^{n+1} x_i \equiv 0 \pmod{9}$	$\mathbb{Z}_9$	10
(c) U.S. Banking	$x_{n+1} \equiv \sum_{i=1}^{n} w_i \cdot x_i \pmod{10}$ $\mathbf{w}$: weight sequence $(7,3,9,7,3,9,...)$	$\mathbb{Z}_{10}$	10
(c) ISBN-10	$\sum_{i=1}^{10} i \cdot x_i \equiv 0 \pmod{11}$	$\mathbb{Z}_{11}$	10
(d) MOD 11-2	$\sum_{i=1}^{n+1} 2^i \cdot x_i \equiv 0 \pmod{11}$	$\mathbb{Z}_{11}$	10
(e) Credit cards	$\sum_{i=1}^{8} \sigma(x_{2i-1}) + x_{2i} \equiv 0 \pmod{10}$, σ : permutation $(0)(1,2,4,8,7,5)(3,6)(9)$	$\mathbb{Z}_{10}$	10
(f) MOD 37, 36	$(\cdots((((36+x_1)\|\|_{36} \cdot 2)\|_{37} + x_2)\|\|_{36} \cdot 2)\|_{37} + \cdots + x_{n+1})\|\|_{36} \equiv 1$ $x\|\|_{36} = x \pmod{36}$ and $x\|_{37} = x \pmod{37}$	$\mathbb{Z}_{36}, \mathbb{Z}_{37}$	36
(g) Deutsche Mark	$\delta(x_1) * \delta^2(x_2) * \cdots * \delta^n(x_n) * x_{n+1} = 0$ $*$: multiplication over $\mathbb{D}_5$ δ : permutation $(0,1,5,8,9,4,2,7)(3,6)$	$\mathbb{D}_5$	10

[1] In general, $x_1, x_2, \cdots, x_n$ denotes the information sequence and x_{n+1} is the check digit;

[2] Before calculation, the country code is converted into integers and necessary rearrangement is made.

Table 3: Undetected errors of error types 1)-5) in percentage.

Check digit system	Undetected errors[1] in %				
	1): single error	2): transposition	3): jump transposition	4): twin error	5): jump twin error
(c) U.S. Banking	0.0	11.11	11.11	44.4	36.6
(c) ISBN-10	0.0	0.0	0.0	11.11	0.0
(d) MOD 11-2	0.0	0.0	0.0	0.0	0.0
(e) (Decimal) Luhn formula	0.0	2.2	100.0	6.67	12.33
(e) (Hexadecimal) Luhn formula	0.0	0.833	100.0	4.167	6.667
(f) MOD 37, 36	0.0	0.159	1.905	1.900	3.642
(g) Verhoeff's method	0.0	0.0	5.78	4.45	5.78

[1] Calculations are based on the assumption that each character occurs in each position (of the information digits) with equal probability.

[2] Our design as alternatives for check digit systems over $\mathbb{Z}_{11}$, $\mathbb{Z}_{16}$ and $\mathbb{Z}_{36}$, respectively.

The use of check digit systems in real life applications is pervasive, penetrating every corner of the globe and of our daily lives. Some well-known examples include but not limited to credit card numbers, ISBN code for books, UPC code for products, VIN for Vehicles, IMEI for mobile phones [MEID05], and identity card number for people in most nations. In spite of advances in coding theory [MS77], most of codes used in practice are still based on simple arithmetic methods such as division, (weighted) modular check sum, Luhn formulae, which mostly operate over a group. A list of check digit systems in real life applications is provided in Table 4. As one can see, most of schemes could only detect all the 1) single errors, but fail to detect all the errors if belonging to types of 2)-5).

Table 4: Check digit systems in real life applications

Methods	Applications	Error Detection
(a) Division	Airline Tickets, Federal Express packages, UPS packages, Avis and National Rental Cars, International Bank Account Number (IBAN) ...	
(b) Parity check	Euro Banknotes, Visa Travellers Cheques, American Express Traveller's Checks, U.S. Postal Service Money Orders, US Zip Codes ...	1)
(c) Weighted parity check	US Banking system, International Standard Book Number (ISBN), International Standard Serial Number (ISSN), International Article Number (EAN), Universal Product Code (UPC), Vehicle Identification Number (VIN)...	1)
(d) Polynomial	Account numbers in Dresdener Bank and Sparkasse der Stadt Berlin West, Chinese citizen ID number (2nd generation)...	1)-5)
(e) Luhn's formula	Most credit cards (e.g.: Visa, Master Card and American Express), International Mobile Equipment Identifier (MEID) [MEID05], Canadian Social Insurance Numbers, National Provider Identifier (NPI), International Securities Identification Number (ISIN)...	1), 3)
(f) Hybrid system	European Blood Banks, Unique Identifier for People (U.S. NIH), Livestock Identification [?], International Standard Audiovisual Number (ISAN) [?, ?]...	1)
(g) Verhoeff's method	serial numbers of the former German Banknotes, i.e., Deutsche Mark SNOMED clinical term identifier Identity number for the residence in India...	1)-2)

3 New Challenges

3.1 In Theory

The well-developed error-control codes work over a finite field of a prime power order. However, the error detection codes desired for specified applications could have an alphabet size beyond this category. A typical example is the decimal codes which have alphabet size 10. Since there is no field of order 10, rich results in coding theory could not be applied here. This partially explains why it is difficult to design codes over an arbitrary alphabet.

One approach to reverse this unfavorable situation is to allow computation in a larger field. Its effectiveness has been demonstrated by some well-known and widely used schemes. Take the ISBN-10 code and MOD 11-2 code as examples, both of which are decimal codes whilst operate over $\mathbb{Z}_{11}$. That is, only decimal numbers (i.e., $0, \cdots, 9$) are allowed in the information digits. However, since the operation is over $\mathbb{Z}_{11}$, an extra X is needed to represent 10 in the check digit as ISBN-10 code does; or, alternatively, those information sequences resulting in a check digit 10 should not be allowed for use. As a gain by operating over a larger field, both codes perform better than those operating over a group of the alphabet size, which instances include the ISBN-13 code by the weighted parity check method modulo 10 and the Verhoeff's decimal code over the dihedral group $\mathbb{D}_5$.

Note that more check digits can be added so that the resulting codes have also capability of error correcting. In general, the problem differentiates itself from the classical error correcting codes (in a field) in threefold. First, it works over any alphabet, not necessarily a prime power; Secondly, the operation is not necessarily over a field (e.g., it could operate over quasigroups, which multiplication table is a Latin square). Thirdly, it deals with detecting/correcting atypical error patterns, for example, type 2) transposition error, which is treated as two independent substitution errors in classical coding theory. All these make the problem more general and difficult than the classical coding theory. Especially for the system over an alphabet of size $2s$, where s is an odd prime, the best known approach is still by Gumm in [G85] that detects all the errors of types 1)-2).

3.2 In practice

In spite of exciting theoretical achievements on the analysis and construction of the check digit system, which go beyond the decimal arithmetic (e.g., (a) division), make use of operation over groups (e.g., (b) parity check and (c) weighted parity check methods), multiplication over finite fields (e.g., (d) polynomial method), non-commutative operation over Dihedral groups (e.g.: (g) Verhoeff's method) [G85, V69], or even recent progress over quasigroups [D00, BIM105, BIM205]. However, as one can see from Table 4, most real life applications still favor the methods which enjoy the arithmetic simplicity, although their performance on error detection is far behind the theoretical advances.

An encouraging act is from the video community, which recently takes Niemenmaa's hexadecimal design [N11] that is capable of detecting all errors of types 1)-5), as a part of MISB standard 1204.1. And, an alphanumeric instance of [C0] is implemented to check the correctness of the Electric Vehicle

Contract IDs (short: CID; also known as eMA-ID or EVCO-ID). Note that the CID is described by the eMI^3 Group and standardized in ISO/IEC-15118 [ISO15118, Annex H]. While it is disappointing to observe that a good scheme in use could be actually replaced by a choice with worse performance just due to its simplicity. Such instances include the series number of Euro banknotes and the ISBN codes. More specifically for the banknotes, the Deutsche Mark (before 2002) used the Verhoeff's algorithm, which is easy to implement and still remains the best decimal code so far in terms of the error detection capability for being able to detect errors of types 1)-2); while the Euro banknotes (since 2002) use the sum modulo operation $\pmod 9$, i.e., (b) parity check method, which detects only all the 1) single errors.

In general, it is of great interest from both theoretical and practical points of view, to develop the code design methods which are universal regardless of the alphabet size and the number of information digits, as well as adaptive to further updates with additional information digits or/and check digits.

4 Acknowledgement

The work is supported by the German Research Foundation (DFG) under Grant CH 601/2-1.

References

[B67] D. F. Beckley, An optimum system with modulo 11, *Comp. Bull.*, **11**, pp. 213-215, 1967.

[V69] J. Verhoeff, Error detecting decimal codes, *Math. Centre Tracts*, **29**, Math. Centrum Amsterdam, 1969.

[MS77] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, 1977.

[G85] H. P. Gumm, A new class of check-digit methods for arbitrary number systems, *IEEE Trans. Info. Theory*, **31**, pp. 102-105, 1985.

[D00] H. M. Damm, Check digit systems over groups and anti-symmetric mappings, *Archiv der Mathematik*, **75**, pp. 413-421, 2000.

[S00] R. H. Schulz, On check digit systems using anti-symmetric mappings, *Numbers, Information and Complexity*, Kluwer Acad. Publ. Boston, pp. 295-310, 2000.

[BIM105] G. B. Belyavskaya, V. I. Izbash and G. L. Mullen, Check character systems using quasigroups: I, *Designs, Codes and Cryptography*, **37**, pp. 215-227, 2005.

[BIM205] G. B. Belyavskaya, V. I. Izbash and G. L. Mullen, Check character systems using quasigroups: II, *Designs, Codes and Cryptography*, **37**, pp. 405-419, 2005.

[N11] M. Niemenmaa, A check digit system for hexadecimal numbers, *Applic. Algebra in Eng., Comm. and Computing*, **22**, pp. 109-112, 2011.

[C0] Y. Chen, M. Niemenmaa, A. J. Han Vinck, A general check digit system based on finite groups, *Design, Codes and Cryptography*, **80**, no. 1, pp. 149-163, 2016.

[ISO7064] ISO/IEC 7064: 2003(E): Information technology - Security techniques - Check character systems.

[MEID05] 3GPP2 report S. R0048: 3G Mobile Equipment Identifier (MEID) - Stage 1. Jun. 2005.

[ISO7812] ISO/IEC 7812-1: 2006(E): Identification cards - Identification of issuers - Part 1: Numbering system.

[MIIS] MISB ST 1204.1: Motion Imagery Identification System (MIIS) - Core Identifier. Oct. 2013.

[ISO15118] ISO 15118-1:2013: Road vehicles - Vehicle to grid communication interface - Part 1: General information and use-case definition.

Rate-Reliability for Protected Biometric Identification System with Secret Generation

Mariam Haroutunian and Lilit Ter-Vardanyan

Institute for Informatics and Automation Problems of
Armenian National Academy of Sciences
Yerevan, Armenia,
Email: armar@ipia.sci.am, lilit@sci.am

Abstract

We investigate the region of all E-achievable secret-key, identification and privacy-leakage rate triples for protected biometric identification system with secret generation. This is the generalization of the achievable region studied by Ignatenko and Willems, ensuring when N increases the error probability exponential decrease with given exponent (reliability) E. In this paper the outer and inner bounds for this region are constructed. As a consequence from the main result we formulate the connection to previous results.

1 Introduction

Security concerns related to the use of biometric data in different secrecy systems were raised a long time ago. From the information-theoretical perspective the biometric secrecy systems were studied by O'Sullivan and Schmid [1] and Willems et al [2].

One of the main issues in biometric security is the **reliable identification** of persons based on their biometric data. Willems et al [2] investigated the fundamental properties of biometric identification system. It has been shown that it is not possible to identify reliably more persons than capacity which is an inherent characteristic of any identification system. They derived the capacity of such system

The problem of **generating secret keys** from biometric data is closely related to the concept of secret sharing, which was introduced by Maurer [3] and by Ahlswede and Csiszar [4]. The problem in biometric setting was considered by Ignatenko and Willems [5]. Unlike traditional secret key sharing, when the secret key is being generated and shared between terminals, in biometric secrecy systems a secret key is generated 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 an access.

Reliable biometric secrecy systems extract helper data from the biometric information at the time enrollment, as biometric measurements are typically noisy. These helper data contribute to reliable reconstruction of the secret key. In review of Ignatenko and Willems [6] various models of biometric secrecy systems are studied from an information-theoretical point of view.

In this paper we consider the model of identification with secret generation studied by Ignatenko and Willems [6, ch. 5.3]. In that system two terminals observe the enrollment and identification biometric sequences of a group of individuals. The first terminal forms a secret key for each enrolled individual and stores the corresponding helper data in a public database. These helper data on one hand facilitate reliable reconstruction of the secret key and on the other hand allow determination of the individuals identity for the second terminal, based on the presented biometric identification sequence. All helper data in the database are assumed to be public, hence the helper data should provide no information on the secrets and as little as possible information on biometric data.

These system should be able to identify as many individuals as possible while being able to assign as large as possible secret keys to each individual and minimize the privacy leakage. Privacy leakage is defined as the amount of information that the public data stored in the database contains about the biometric enrollment sequence.

The region of all achievable secret-key, identification and privacy leakage rate triples for this model is obtained by Ignotenko and Willems [6].

We investigate the interdependence between secret-key, identification and privacy leakage rate triples and error probability exponent. In difference from classic error exponent setting we consider the inverse dependence of rates and reliability. By analogy with notion of E-capacity or rate-reliability function introduced for DMC by E. Haroutunian [7, 8], we call this region

E-achievable triples region. This is the generalization of the achievable region studied by Ignatenko and Willems, ensuring when N increases the error probability exponential decrease with given exponent (reliability) E.
In survey of E. Haroutunian, M. Haroutunian and A. Harutunyan [8] various transmission systems are studied from reliability criteria point of view. This approach was developed also for biometric systems. In [9] the new concept of identification E-capacity for biometric identification system was investigated by authors, which is the generalization of capacity studied in [6]. A similar investigation was presented in [10] for the biometric generated secret key sharing system. Because of principal difficulty of finding the rate-reliability usually an upper and lower bounds of this function are constructed.
Error exponents are studied for a large number of secrecy contexts. Particulary, in [11] a tradeoff between the secret key rate and exponential bounds on the probability of key agreement failure and on the secrecy of the key generated from the excited distributed source is characterized.
In this paper we construct outer and inner bounds of the E-achievable triples region. As a consequence from the main result we formulate the connection to previous results.

2 Definitions

The following conventions are applied within the paper. Capital letters are used for random variables (RV) X, Y taking values in the finite sets $\mathcal{X}$, $\mathcal{Y}$, correspondingly, and lower case letters x, y for their realizations. Small bold letters are used for N-length vectors $\mathbf{x} = (x_1, x_2, ..., x_N) \in \mathcal{X}^{\mathcal{N}}$. The cardinality of the set $\mathcal{X}$ we denote by $|\mathcal{X}|$. The notation $|a|^+$ will be used for $\max(a, 0)$.

The model of biometric identification with secret generation system consists of enrollment and identification parts (Fig. 1). In an **enrollment phase** $|\mathcal{V}|$ individuals are observed. For each individual $v \in \{1, 2, \cdots, |\mathcal{V}|\}$ the biometric source produces a biometric enrollment sequence $\mathbf{x}(v) = \{x_1, x_2, \cdots, x_N\}$, where $x_n \in \mathcal{X}$, $n = \overline{1, N}$. All these sequences are supposed to be generated at random with a given probability distribution

$$Pr\{X^N = \mathbf{x}\} = Q_X^N(\mathbf{x}) = \prod_{n=1}^{N} Q_X(x_n), \quad \mathbf{x} \in \mathcal{X}^N.$$

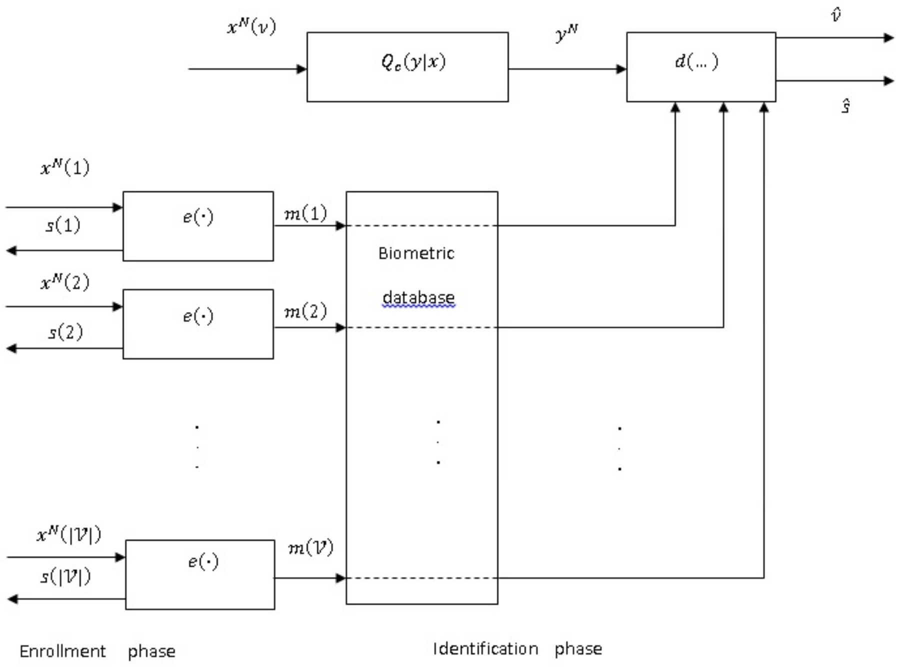


Figure 1: Model of protected biometric identification system with secret generation.

During the **enrollment procedure** the biometric sequence $\mathbf{x}(v)$ of individual $v \in \{1, 2, ..., |\mathcal{V}|\}$ is encoded into helper data (or protected template) $m(v) \in \{1, 2, ..., |\mathcal{M}|\}$ and a secret key $s(v) \in \{1, 2, ..., |\mathcal{S}|\}$, hence

$$e(\mathbf{x}(v)) = (m(v); s(v)); \quad \text{for} \;\; v \in \{1, 2, ..., |\mathcal{V}|\},$$

where e is the **encoder mapping**. The helper data $m(v)$ is then stored in a (public) database at position v. The generated secret key $s(v)$ is handed over to the individual.

The helper data that are stored in the database make reliable identification possible. The helper data should provide no information on the secret (secrecy leakage) and as little as possible information about biometric data (privacy leakage).

During the **identification procedure** a biometric identification sequence $\mathbf{y} = (y_1, y_2, \cdots, y_N)$, $y_n \in \mathcal{Y}$, $n = \overline{1, N}$ is observed. If individual v was observed, the sequence $\mathbf{y}(v)$ occurs with probability

$$Pr\{Y^N = \mathbf{y}|X^N = \mathbf{x}(v)\} = Q^N_{Y|X}(\mathbf{y}|\mathbf{x}) =$$

$$= \prod_{n=1}^{N} Q_{Y|X}(y_n|x_n), \quad \mathbf{y} \in \mathcal{Y}^N, \quad \mathbf{x} \in \mathcal{X}^N,$$

since the biometric channel $Q^N_{Y|X}(\mathbf{y}|\mathbf{x})$ is memoryless. We assume here that all individuals are equally likely to be observed for identification, hence

$$P_V\{V = v\} = \frac{1}{|\mathcal{V}|}; \quad \text{for all} \ \ v \in \{1, 2, ..., |\mathcal{V}|\}.$$

During identification, upon observing the biometric identification sequence $\mathbf{y}$ the decoder forms an estimate $\hat{v}$ of the identity of the observed individual as well as an estimate of his secret key $\widehat{s(v)}$,

$$d(\mathbf{y}, m(1), m(2), ..., m(|\mathcal{V}|)) = (\hat{v}, \widehat{s(v)}),$$

where d is the **decoder mapping**.
The estimate of the individual's identity $\hat{v}$ takes on values from the set of individuals, i.e. $\hat{v} \in \{1, 2, ..., |\mathcal{V}|\}$. Moreover, the decoder's estimate of the secret key $\widehat{s(v)}$ assumes values from the same alphabet as the secret key generated during enrollment, hence $\widehat{s(v)} \in \{1, 2, ..., |\mathcal{S}|\}$.
Now we are interested to find out what identification, secret-key and privacy-leakage rates can be realized by such an identification system with $\exp\{-NE\}$ probability of error, such that secret keys are close to uniform in the entropy sense and the helper data provide negligible information on the secret.
Definition (E-achievability). A secret-key rate R_S, identification rate R_I and privacy-leakage rate R_L triple (R_S, R_I, R_L) with $R_S \geq 0$ and $R_I \geq 0$ is called $E-$achievable in protected biometric identification setting with secret generation if for all $\delta > 0$, $E > 0$ and N large enough there exist encoders and decoders such that

$$Pr\{(\widehat{V}, \widehat{S(V)} \neq (V, S(V))\} \leq 2^{-N(E-\delta)}, \tag{1}$$

$$\frac{1}{N}\log|\mathcal{V}| \geq R_I - \delta,$$

$$\frac{1}{N}H(S) + \delta \geq \frac{1}{N}\log|\mathcal{S}| \geq R_S - \delta,$$

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

$$\frac{1}{N} I(X^N \wedge M) \leq R_L + \delta.$$

The region of all $E-achievable$ rate triples we denote by $\mathcal{R}(E, Q)$. Throughout this paper all logarithms and exponents are of base 2.

3 Formulation of the Result

We shall use the following probability distributions in the formulation of the result:

$$Q_{U|X} = \{Q_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

$$Q_{XU} = \{Q_{XU}(x, u) = Q_X(x) \cdot Q_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

$$Q_U = \{Q_U(u) = \sum_x Q_X(x) \cdot Q_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

$$P_X = \{P_X(x), x \in \mathcal{X}\},$$

$$P_{U|X} = \{P_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

$$P_{XU} = \{P_{XU}(x, u) = P_X(x) \cdot P_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

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

$$P_U = \{P_U(u) = \sum_x P_X(x) \cdot P_{U|X}(u|x), u \in \mathcal{U}, x \in \mathcal{X}\},$$

$$Q = \{Q(u, y) = \sum_x Q_X(x) \cdot Q_{U|X}(u|x) \cdot Q_{Y|X}(y|x),$$
$$u \in \mathcal{U}, y \in \mathcal{Y}, x \in \mathcal{X}\},$$

$$P = \{P(u, y) = \sum_x P_X(x) \cdot P_{U|X}(u|x) \cdot P_{Y|X}(y|x),$$
$$u \in \mathcal{U}, y \in \mathcal{Y}, x \in \mathcal{X}\}.$$

We refer to [7]-[10] and [12] for notions and notations of divergence $D(P||Q)$, mutual information $I_P(U \wedge Y)$, information-theoretic quantities. The proofs are based on the method of types [13]. We denote by $\mathcal{T}_{P_U}^N(U)$ the set of vector $\mathbf{u}$ of type P_U, by $\mathcal{T}_P^N(U, Y)$ the set of vector pairs $(\mathbf{u}, \mathbf{y})$ of type P. We use some known properties:

$$for \;\; \mathbf{u} \in \mathcal{T}_P^N(U), \;\; \mathbf{y} \in \mathcal{T}_P^N(Y|\mathbf{u}),$$

$$Q^N(\mathbf{u},\mathbf{y}) = \exp\{-N(H_P(U,Y) + D(P||Q))\}, \tag{2}$$

$$\begin{aligned}(N+1)^{-|\mathcal{U}|}\exp\{NH_P(U)\} \leq |\mathcal{T}_P^N(U)| \leq \\ \leq \exp\{NH_P(U)\},\end{aligned} \tag{3}$$

$$\begin{aligned}(N+1)^{-|\mathcal{U}||\mathcal{Y}|}\exp\{NH_P(Y|U\} \leq |\mathcal{T}_P^N(Y|\mathbf{u}))| \leq \\ \leq \exp\{NH_P(Y|U)\},\end{aligned} \tag{4}$$

$$Q^N(\mathcal{T}_P^N(U,Y)) \leq \exp\{-ND(P||Q)\}. \tag{5}$$

Our main result is stated in the following theorem.

Theorem. *For the biometric identification system with secret generation the region $\mathcal{R}(E,Q)$ is outer bounded by*

$$\begin{gathered}\mathcal{R}_{sp}(E,Q) = \{R_S, R_I, R_L) : R_I \geq 0, R_S \geq 0, \\ 0 \leq R_I + R_S \leq \min_{P:D(P||Q)\leq E} I_P(U \wedge Y), \\ R_L \geq \min_{P_{XU}:D(P_{XU}||Q_{XU})\leq E} I_P(U \wedge X) - \\ - \min_{P:D(P||Q)\leq E} I_P(U \wedge Y) + R_I, \\ \mathit{for\ \ some}\ \ Q_{U|X}\ \ \mathit{and}\ \ |\mathcal{U}| \leq |\mathcal{X}| + 1\},\end{gathered}$$

and inner bounded by

$$\begin{gathered}\mathcal{R}_r(E,Q) = \{R_S, R_I, R_L) : R_I \geq 0, R_S \geq 0, \\ 0 \leq R_I + R_S \leq \min_{P:D(P||Q)\leq E} |I_P(U \wedge Y) + D(P||Q) - E|^+, \qquad (6) \\ R_L \geq \min_{P_{XU}:D(P_{XU}||Q_{XU})\leq E} |I_{P_{XU}}(U \wedge X) + D(P_{XU}||Q_{XU}) + \\ + E|^+ - \min_{P:D(P||Q)\leq E} |I_P(U \wedge Y) + D(P||Q) - E|^+ + R_I\end{gathered}$$

for some $Q_{U|X}$ and $|\mathcal{U}| \leq |\mathcal{X}| + 1\}$.
The notation $\mathcal{R}_{sp}$ and $\mathcal{R}_r$ of bounds comes from the techniques (sphere packing and random coding) by which similar bounds were first obtained.
The regions characterize the trade-off between the three rates and reliability E. For given E if the number of individuals to be identified increases

then the secret keys get smaller. For greater E the sum of identification and secret-key rates is smaller.
From our result we can derive a number of previously established results.

Corollary 1. *When $E \to 0$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result obtained in [6, Theorem 5.1]:*

$$\begin{gathered}\mathcal{R}(Q) = \{(R_I, R_S, R_L) : R_I \geq 0, R_S \geq 0,\\ 0 \leq R_I + R_S \leq I_Q(U \wedge Y)\},\\ R_L \geq I_Q(U \wedge X) - I_Q(U \wedge Y) + R_I,\\ \text{for some } Q_u \text{ and } |\mathcal{U}| \leq |\mathcal{X}| + 1.\end{gathered}$$

Corollary 2. *When $E \to 0$ and $R_L = \infty$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result obtained in [5]:*

$$\mathcal{R}(Q) = \{(R_I, R_S) : 0 \leq R_I + R_S \leq I_Q(X \wedge Y)\}.$$

Corollary 3. *When $R_S = 0$, $R_L = \infty$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result from [9]:*

$$\mathcal{R}_{sp}^I(E,Q) = \min_{P:D(P||Q)\leq E} I_P(X \wedge Y),$$

$$\mathcal{R}_r^I(E,Q) = \min_{P:D(P||Q)\leq E} |I_P(X \wedge Y) + D(P||Q) - E|^+.$$

Corollary 4. *When $R_I = 0$, $R_L = \infty$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result from [10]:*

$$\mathcal{R}_{sp}^S(E,Q) = \min_{P:D(P||Q)\leq E} I_P(X \wedge Y),$$

$$\mathcal{R}_r^S(E,Q) = \min_{P:D(P||Q)\leq E} |I_P(X \wedge Y) + D(P||Q) - E|^+.$$

Corollary 5. *When $E \to 0$, $R_S = 0$, $R_L = \infty$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result obtained in [2]:*

$$\mathcal{R}(Q) = \{R_S : R_S \leq I_Q(X \wedge Y)\}.$$

Corollary 6. *When $E \to 0$, $R_I = 0$, $R_L = \infty$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result obtained in [6, Theorem 2.1]:*

$$\mathcal{R}(Q) = \{R_I : R_I \leq I_Q(X \wedge Y)\}.$$

Corollary 7. *When $E \to 0$, $R_I = 0$ we get the inner and outer bounds of $\mathcal{R}(E,Q)$ which coincide with the result obtained in [6, Theorem 3.1]:*

$$\mathcal{R}(E,Q) = \{(R_S, R_L) : 0 \leq R_S \leq I_Q(U \wedge Y),$$

$$R_L \geq I_Q(U \wedge X) - I_Q(U \wedge Y)$$

$$\text{for some } Q_{U|X} \text{ and } |\mathcal{U}| \leq |\mathcal{X}| + 1\}.$$

4 Proof of the Theorem

The outline of the proof. To prove the sphere packing bound we suppose that error probability satisfies the condition (1) and by properties of method of types bound $|\mathcal{S}||\mathcal{V}|$.
To prove the random coding bound we introduce the encoding function by the use of auxiliary vectors $\mathbf{u}$ and as a decoding function we consider the minimum divergence method. Then we estimate the encoding and decoding error probabilities and show that the error probability is small enough for (6).

The outer bound. Let $E > 0$ and the error probability satisfies the condition (1). Let us denote by $\mathbf{x}(m,s,v)$ the enrollment sequence $\mathbf{x}(v)$ of individual v that is encoded into $m(v); s(v)$. We also denote by $d^{-1}(s,v)$ the set of identification sequence that are decoded to (v,s):

$$d^{-1}(v,s) = \{\mathbf{y} : d(\mathbf{y}), m(1), ..., m(|\mathcal{V}|)\}.$$

Then (1) can be rewritten as

$$\sum_{\mathbf{x}} Q_X(\mathbf{x}(m,s,v)) \times Q_{Y|X}^N\{Y^N \setminus d^{-1}(s,v)|\mathbf{x}(m,s,v)\} \leq$$

$$\leq \exp\{-N(E-\delta)\}$$

for each s, v.
For some auxiliary alphabet $\mathcal{U}$ and probability $Q_{U|X}$

$$Q_X^N(\mathbf{x}) = \sum_{\mathbf{u}} Q_{XU}^N(\mathbf{x},\mathbf{u})$$

and we can write

$$\sum_{\mathbf{u}} \sum_{\mathbf{x}} Q_{XU}^N(\mathbf{x}(m,s,v), \mathbf{u}(m,s,v)) \times$$

$$\times Q^N_{Y|X}\{Y^N \setminus d^{-1}(s,v)|\mathbf{x}(m,s,v)\} \leq$$
$$\leq \exp\{-N(E-\delta)\}$$

or

$$\sum_{\mathbf{u}} Q^N_U(\mathbf{u}(m,s,v)) \cdot Q^N\{Y^N \setminus d^{-1}(s,v)|\mathbf{u}(m,s,v)\} \leq$$
$$\leq \exp\{-N(E-\delta)\}.$$

For any type P we have

$$\sum_{\mathbf{u}(m,s,v)\in\mathcal{T}^N_P(U)} Q^N_U(\mathbf{u}(m,s,v))\times$$
$$\times Q^N\{\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v)) \setminus d^{-1}(s,v)|\mathbf{u}(m,s,v)\} \leq$$
$$\leq \exp\{-N(E-\delta)\}.$$

From (2) we have that the probability $Q^N(\mathbf{u},\mathbf{y})$ is constant for various $\mathbf{u}$ and $\mathbf{y}$ of fixed type P, hence we derive

$$\sum_{\mathbf{u}(m,s,v)\in\mathcal{T}^N_P(U)} \{|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v))|-$$
$$-|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v)) \cap d^{-1}(s,v)|\} \cdot Q^N(\mathbf{u},\mathbf{y}) \leq$$
$$\leq \exp\{-N(E-\delta)\}.$$

According to (2) and (3) we obtain

$$\begin{gathered}\{|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v))|- \\ -|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v)) \cap d^{-1}(s,v)|\}\times \\ \times\exp\{-N(D(P||Q)+H_P(Y,U)-H_P(U))\} \leq \\ \leq \exp\{-N(E-\delta)\}\end{gathered} \tag{7}$$

which holds for all s, v, hence

$$\sum_{s,v}\{|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v))|-$$
$$-|\mathcal{T}^N_P(Y|\mathbf{u}(m,s,v)) \cap d^{-1}(s,v)|\}\times$$
$$\times\exp\{-N(D(P||Q)+H_P(Y,U)-H_P(U))\} \leq$$

$$\leq |\mathcal{S}| \cdot |\mathcal{V}| \cdot \exp\{-N(E-\delta)\}.$$

From the definition of decoding function d it follows that the sets $d^{-1}(s,v)$ are disjoint, therefore

$$\sum_{s,v} |\mathcal{T}_P^N(Y|\mathbf{u}(m,s,v)) \cap d^{-1}(s,v)|\} \leq |\mathcal{T}_P^N(Y)|. \tag{8}$$

Then from (6) and (7) we have

$$\sum_{s,v} |\mathcal{T}_P^N(Y|\mathbf{u}(m,s,v))| -$$

$$-\frac{|\mathcal{S}| \cdot |\mathcal{V}| \cdot \exp\{-N(E-\delta)\}}{\exp\{-N(D(P||Q)+H_P(Y|U))\}} \leq |\mathcal{T}_P^N(Y)|.$$

Hence from (4) we obtain

$$|\mathcal{S}| \cdot |\mathcal{V}| \leq \frac{\exp\{NI_P(Y \wedge U)\}}{(N+1)^{-|\mathcal{U}||\mathcal{Y}|} - \exp\{ND(P||Q) - E + \delta\}}.$$

The right-hand side of this inequality can be minimized by the choice of type P keeping the denominator positive, which for large N holds when $D(P||Q) \leq E - \delta$.

The proof of privacy leakage repeats the main steps of the proof from [6].

The inner bound. Let us fix the auxiliary alphabet $\mathcal{U}$ and the conditional probability $Q_{U|X}$.

Code Construction. Let us generate $|\mathcal{J}|$ auxiliary vectors $\mathbf{u}$ at random according to Q_U. For each $\mathbf{u}(j)$, $j \in \{1,2,...,|\mathcal{J}|\}$ encoder generates uniformly at random a helper label $m(j) \in \{1,2,...,|\mathcal{M}|\}$ and a secret-key label $s(j) \in \{1,2,...,|\mathcal{S}|\}$.

The encoder observes the biometric source sequence $\mathbf{x}(v)$ of individual v and then finds the index j such that

$$(\mathbf{x}(j), \mathbf{u}(v)) \in T_{P_{XU}}^N(X,U)$$

with minimal $D(P_{XU}||Q_{XU})$.

For that j the helper data $m(j)$ is stored at location v in the database and the secret $s(j)$ is handed over to the individual. If there is another index

$j' \neq j$ such that $s(j') \neq s(j)$ and $m(j') \neq m(j)$, the encoder declares an error and j gets an arbitrary value from $\{1, 2, ..., |\mathcal{J}|\}$ (event A).

The decoder observes biometric sequence $\mathbf{y}$ and checks all the records $v \in \{1, 2, ...|v|\}$ in the database. The decoder determines a unique individual $\hat{v}$ whose record contains $m(\hat{v}) = m(\hat{j})$ for which there exists a unique index $\hat{j}$ such that $(\mathbf{u}(\hat{j}), \mathbf{y}) \in T_P(U, Y)$ with minimal $D(P||Q)$. The decoder outputs the estimate $\hat{v}$ and a secret-key estimate $\hat{s(v)} = \hat{s(j)}$. Error occurs if $(\hat{v}, \hat{s(v)}) \neq (v, s(v))$ (event B).

Error probability. It can be shown that for

$$\frac{1}{N} \log |J| \leq \min_{P_{XU}: D(P_{XU}||Q_{XU}) \leq E} |I_{P_{XU}}(U \wedge X) +$$

$$+ D(P_{XU}||Q_{XU}) - E|^+$$

the probability of the event A is small enough

$$Pr\{A\} \leq \exp\{-N(E - \varepsilon_1)\}.$$

The proof is similar for the probability of the event B, so we bring the detailed proof only for B.

From (5) we obtain that for any P such that $D(P||Q) > E$ the probability of that type is small enough

$$Q^N(\mathcal{T}_P^N(U, Y)) \leq \exp\{-NE\}. \tag{9}$$

For any P such that $D(P||Q) \leq E$ the error can occur if given m, v, s there exists a pair $(\hat{v}, \widehat{s(v)}) \neq (v, s(v))$, such that for some $\hat{P}$

$$(\mathbf{u}(m, s, v), \mathbf{y}) \in \mathcal{T}_P^N(U, Y), (\hat{\mathbf{u}}(\hat{m}, \hat{s}, \hat{v}), \mathbf{y}_m) \in \mathcal{T}_{\hat{P}}^N(U, Y)$$

and

$$D(\hat{P}||Q) \leq D(P||Q).$$

The mathematical expectation of this event can be upper bounded by the following expression:

$$\sum_{P, \hat{P}: D(\hat{P}||Q) \leq D(P||Q)} \sum_{(\hat{v}, \hat{s(v)}) \neq (v, s(v))} \sum_{\mathbf{u} \in \mathcal{T}_P^N(U)} \tag{10}$$

$$\sum_{\mathbf{y}\in\mathcal{T}_P^N(Y)} Q^{\mathrm{N}}(\mathbf{u}(m,s,v),\mathbf{y})\times$$
$$\times Pr\{(\mathbf{u}(m,s,v),\mathbf{y})\in\mathcal{T}_P^N(U,Y)\}\times$$
$$\times Pr\{(\hat{\mathbf{u}}(\hat{m},\hat{s},\hat{v}),\mathbf{y})\in\mathcal{T}_{\hat{P}}^N(U,Y)\}.$$

From

$$R_I+R_S\leq\min_{P:D(P||Q)\leq E}|I_p(U\wedge Y)+D(P||Q)-E|^+$$

follows that for any $\hat{P}$

$$|\mathcal{S}|\cdot|\mathcal{V}|\leq\exp\{N(I_{\hat{P}}(U\wedge Y)+D(\hat{P}||Q)-E-\delta_1)\}.$$

From (2) and (3) we obtain that (10) is not greater than

$$\sum_{P,\hat{P}:D(\hat{P}||Q)\leq D(P||Q)}\exp\{N(I_{\hat{P}}(U\wedge Y)+D(\hat{P}||Q)-E-\delta_1)\}\times$$
$$\times\exp\{-N(H_P(U,Y)+D(P||Q))\}\times$$
$$\times\exp\{N(H_P(U)+H_P(Y))\}\times$$
$$\times\exp\{-N(I_P(U\wedge Y)-\delta_2)\}$$
$$\times\exp\{-N(I_{\hat{P}}(U\wedge Y)-\delta_3)\}=$$
$$\sum_{P,\hat{P}:D(\hat{P}||Q)\leq D(P||Q)}\exp\{N(D(\hat{P}||Q)-E-(\delta_1+\delta_2+\delta_3))\}\times$$
$$\times\exp\{-N(D(P||Q)\leq\exp\{-N(E-\delta')\}. \tag{11}$$

From the (9) and (11) we state that

$$Pr\{B\}\leq\exp\{-N(E-\varepsilon_2)\}$$

and hence

$$Pr\{(\hat{V},S(\hat{V}))\neq(V,S(V))\}=Pr\{A\}+Pr\{B\}\leq\exp\{-N(E-\delta)\}.$$

The proof of **uniformity and secrecy leakage** is similar to the proof from [6].

The proof of **privacy leakage** follows immediately from

$$\frac{1}{N}I(X^N\wedge M)\leq\frac{1}{N}H(M)\leq\frac{1}{N}\log|M|=$$

$$= \frac{1}{N} \log \frac{|J|}{|S|} \leq$$

$$\leq \min_{P_{XU}: D(P_{XU}||Q_{XU}) \leq E} |I_{P_{XU}}(U \wedge X) + D(P_{XU}||Q_{XU}) - E|^{+} +$$

$$+ R_I - \min_{P: D(P||Q) \leq E} |I_P(U \wedge Y) + D(P||Q) - E|^{+}.$$

The bound on cardinality of U is proved in [6] using the Fenchel-Eggleston strengthening the Caratheodary lemma.

5 Conclusion

We have considered biometric identification system with protected templates in the case of secret-key generation. The region of all $E-$achievable secret-key, identification and privacy-leakage rates triples is investigated by constructing outer and inner bounds. For small E these bounds coincide and as a consequence some previously established results follow.

References

[1] J. A. OSullivan and N. A. Schmid, “Performance prediction methodology for biometric systems using a large deviations approach”, *IEEE Trans. on Signal Proc.*, vol. 52, no. 10, pp. 3036-3045, 2004.

[2] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, “On the capacity of a biometric identification system”, *International Symposium on Information Theory*, Yokohama, Japan, p. 82, 2003.

[3] U. Maurer, ”Secret key agreement by public discusion from common information”, *IEEE Trans. Inform. Theory*, vol.39, no. 3, pp. 733-742, May 1993.

[4] R. Ahlswede and I. Csiszár, “Common randomness in information theory and cryptography - Part I : Secret sharing”, *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1121 - 1132, July 1993.

[5] Ignatenko, T. and Willems, F.M.J. “Secret-key and identification rates for biometric identification systems with protected templates”. *Proceedings of the 31st Symposium on Information Theory*, pp. 121-128, 2010.

[6] T. Ignatenko and F. 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.

[7] E. Haroutunian. ”On bounds for *E*-capacity of DMC.”, *IEEE Trans. Inform. Theory*, vol. 53, no. 11, pp. 4210–4220, 2007.

[8] 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. 2,3, pp. 97–263, 2007.

[9] M. Haroutunian, A. Muradyan and L. Ter-Vardanyan. ”Upper and lower bounds of biometric identification *E*-capacity.”, *Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science.*, vol. 37, pp. 7–16, 2012.

[10] M. Haroutunian, and N. Pahlevanyan. ”Information theoretical analysis of biometric generated secret key haring model.”, *Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science.*, vol. 42, pp. 10–17, 2014.

[11] T. H. Chou, S. C. Draper and A. M. Sayeed. ”Key generation using external sourse excitation: capacity, reliabilty and secrecy exponent.”, *IEEE Trans. Inform. Theory*, vol. 5, no. 4, pp. 2455–2474, 2012.

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

[13] I. Csiszar. ”The method of types.” *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

Binary Multimedia Wrap Approaches to Protection and Verification over Noisy Data

Vladimir B. Balakirsky*, Anahit R. Ghazaryan[1], and A.J. Han Vinck[2]

[1] School no. 21, Ministry of Defense of Russian Federation, Yerevan, Armenia
[2] Institute of Digital Signal Processing, University of Duisburg-Essen, Germany

Abstract

We describe approaches that can be included into a theoretical base of privacy protection and verification over noisy data.

1 Introduction

The content of the database containing biometric data can be used for different purposes. These purposes include different variations of processing noisy data, like authentication and identification of a user on the basis of his biometric observations, etc. In each case, the key issues are privacy protection and good verification performance. In [3] we designed algorithms when input data are mapped to ternary vectors (the value is either significant of type 0, or significant of type 1, or non-significant). In [4] we presented verification algorithms for processing data represented by float-valued vectors of fixed length. The result of processing at the enrollment stage is expressed as a pair of vectors. The first vector is published, while the second vector is supposed to be encoded on the basis of the first vector and a noisy version

*Deceased

of the input vector. An efficiency of the presented scheme depended on the parameters, extracted from the input vector received at the enrollment stage. In this paper we deal with the setup when data, received at the enrollment stage, are represented by the float-valued vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ of fixed length n, which have to be mapped to a binary string $(\boldsymbol{b}, \boldsymbol{s})$ of length $2n$, called representative of the vector $\boldsymbol{x}$. The first half of this string, called the wrapped version of the data, is published, while the second half, called the message carried by the data, is hidden by a one-way hash function. The hidden part has to decoded after noisy version of the vector $\boldsymbol{x}$ is received at the verification stage. In other words, we presented a lossy secret sharing scheme [2]. Notice that constructed scheme is relevant to the combinatorial construction of colored hypergraphs where the vector $\boldsymbol{s}$ specifies the vertex and the vector $\boldsymbol{b}$ specifies the color (see, for example [1]). The decoder has to identify the vertex having a given color.

The constructed encoding and decoding are combinatorial algorithms, and the performance of data processing scheme essentially depends on matching of the encoding/decoding rules and the probabilistic description of input data and noise of the observation channel. The designed scheme is oriented to reliable delivery of input vectors components, having large magnitudes, to the verifier. Components having this property can be referred to as significant components. An important ingredient of our construction is the so–called F-transformation of the input data, where F specifies the probability distribution of the data. As a result, input data are bijectively mapped to the $(-1, +1)$ interval in such a way that the created probability distribution is uniform. Moreover, significance of component, understood as the fact that the value of the component deviates from the average value, is preserved in a sense that significant components are transmitted over the observation channel with high reliabilities. The statement is true at least for an additive white Gaussian noise observation channel.

2 Notation and basic ideas

We suppose that input multimedia data are represented by a float-valued vector $\boldsymbol{r} = (r_1, \ldots, r_n) \in \mathsf{R}^n$, where n is even, and the identifier $\mathrm{ID}(\boldsymbol{r})$. Let us fix the mapping

$$\Big(\mathrm{ID}(\boldsymbol{r}), \boldsymbol{r}\Big) \rightarrow \Big(\mathrm{ID}(\boldsymbol{r}), \mathrm{wrp}(\boldsymbol{r}), \mathrm{Hash}(\mathrm{mes}(\boldsymbol{r}))\Big),$$

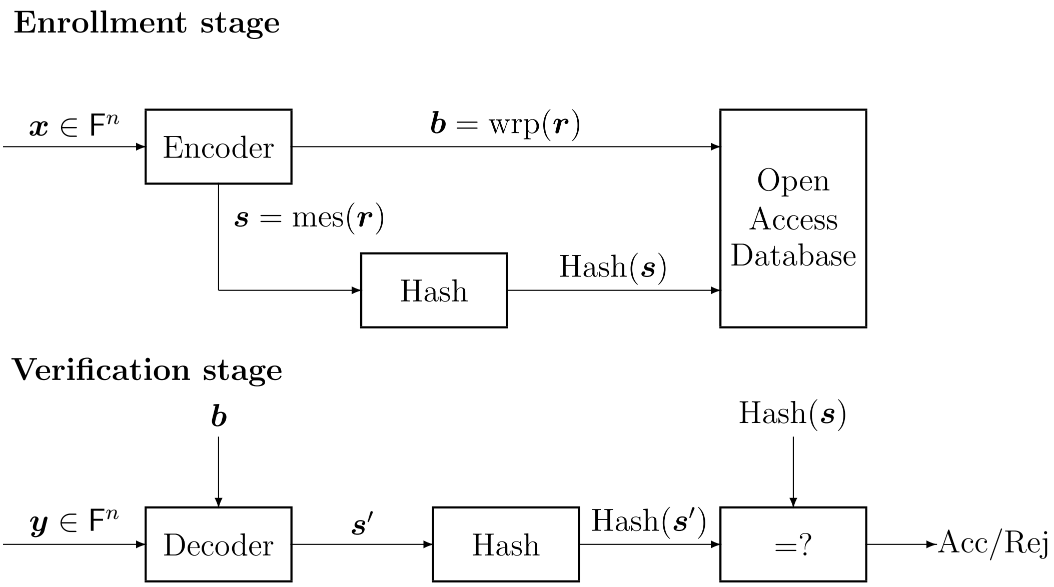

Enrollment stage
$\boldsymbol{x} \in \mathsf{F}^n$
Encoder
$\boldsymbol{b} = \mathrm{wrp}(\boldsymbol{r})$
Open Access Database
$\boldsymbol{s} = \mathrm{mes}(\boldsymbol{r})$
Hash
$\mathrm{Hash}(\boldsymbol{s})$
Verification stage
$\boldsymbol{b}$
$\mathrm{Hash}(\boldsymbol{s})$
$\boldsymbol{y} \in \mathsf{F}^n$
Decoder
$\boldsymbol{s}'$
Hash
$\mathrm{Hash}(\boldsymbol{s}')$
=?
Acc/Rej

Figure 1: General structure of data processing scheme.

which is realized at the enrollment stage. The result of this mapping is stored in an open access database. The vector $\mathrm{wrp}(\boldsymbol{r})$, called the wrapped version of the vector $\boldsymbol{r}$, is a binary vector of length n. The vector $\mathrm{mes}(\boldsymbol{r})$, called the message carried by the vector $\boldsymbol{r}$, is a binary vector of length n and weight $n/2$. The function Hash is a cryptographic ”one–way” function having the property that one can easily find the value of the function for the given argument, but the inversion (finding the argument for the known value of the function) is practically impossible.
Let $\boldsymbol{r}' = (r_1', \ldots, r_n') \in \mathsf{R}^n$ be the vector observed by the verifier. Then the mapping $\Big(\mathrm{ID}(\boldsymbol{r}), \boldsymbol{r}'\Big) \rightarrow \mathrm{Hash}(\boldsymbol{s}')$ called the decoding, is realized at the verification stage. We assume that submission of the identifier $\mathrm{ID}(\boldsymbol{r})$ to the verifier at the verification stage gives him access to the pair

$$\Big(\boldsymbol{b}, \mathrm{Hash}(\boldsymbol{s})\Big) = \Big(\mathrm{wrp}(\boldsymbol{r}), \mathrm{Hash}(\mathrm{mes}(\boldsymbol{r}))\Big).$$

The decision rule is defined as

$$\mathrm{Decision} = \begin{cases} \mathrm{Acc}, & \text{if } \mathrm{Hash}(\boldsymbol{s}') = \mathrm{Hash}(\boldsymbol{s}) \\ \mathrm{Rej}, & \text{if } \mathrm{Hash}(\boldsymbol{s}') \neq \mathrm{Hash}(\boldsymbol{s}). \end{cases}$$

Therefore the message can be viewed as the information that has to be reliably decoded from the noisy version $\boldsymbol{r}'$ of the vector $\boldsymbol{r}$, given the vector $\boldsymbol{b}$. The value of Hash($\boldsymbol{s}$) is treated as an encrypted version of the message.

The enrollment and the verification stages contain the so–called F-transformation of the vectors $\boldsymbol{r}$ and $\boldsymbol{r}'$, to the vectors $\boldsymbol{x}$, $\boldsymbol{y}$, respectively. This transformation is defined as follows. Suppose that $\boldsymbol{r}$ is generated by a stationary memoryless source having the probability distribution (PD) F and the probability density function (PDF) P,

$$F = \Big(\, F(r) = \Pr_{\text{data}}\{\, R < r \,\},\ r \in \mathsf{R} \Big),\ \ P = \Big(\, P(r) = \frac{d}{dr}F(r),\ r \in \mathsf{R} \Big).$$

For all $t = 1, \ldots, n$, let us introduce the F-transformations of $r_t, r_t' \in \mathsf{R}$ to the set $\mathsf{F} = (-1, +1)$ by

$$(x_t, y_t) = \Big(\, 2F(r_t) - 1,\ 2F(r_t') - 1 \Big). \tag{1}$$

The F-transformations of the vectors $\boldsymbol{r}, \boldsymbol{r}' \in \mathsf{R}^n$, expressed as

$$(\boldsymbol{x}, \boldsymbol{y}) = \Big(\, 2F(\boldsymbol{r}) - 1,\ 2F(\boldsymbol{r}') - 1 \Big)$$

will be understood as the results of n applications of the mapping (1).

Data processing scheme under our consideration is given in Figure 1.

3 The encoding/decoding algorithms

Let us denote

$$\begin{aligned} |\boldsymbol{x}| &= (|x_1|, \ldots, |x_n|), \\ \operatorname{sgn}(\boldsymbol{x}) &= (\operatorname{sgn}(x_1), \ldots, \operatorname{sgn}(x_n)), \\ \overline{\operatorname{sgn}}(\boldsymbol{x}) &= (\overline{\operatorname{sgn}}(x_1), \ldots, \overline{\operatorname{sgn}}(x_n)), \end{aligned}$$

where $|x_t|$ is the magnitude of the t-th component of the vector $\boldsymbol{x}$:

$$|x_t| = \begin{cases} +x_t, & \text{if } x_t \geq 0 \\ -x_t, & \text{if } x_t < 0, \end{cases}$$

$\operatorname{sgn}(x_t)$ is the sign of the t-th component of the vector $\boldsymbol{x}$:

$$\operatorname{sgn}(x_t) = \begin{cases} 1, & \text{if } x_t \geq 0 \\ 0, & \text{if } x_t < 0 \end{cases}$$

$\overline{0} = 1$ and $\overline{1} = 0$. Denote also

$$|\boldsymbol{y}| = (|y_1|, \ldots, |y_n|), \;\; \mathrm{sgn}(\boldsymbol{y}) = (\mathrm{sgn}(y_1), \ldots, \mathrm{sgn}(y_n)).$$

For a formal convenience, let us assume that the magnitudes of all components of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are different, i.e.,

$$t \neq t' \;\Rightarrow\; |x_t| \neq |x_{t'}| \tag{2}$$

and

$$t \neq t' \;\Rightarrow\; |y_t| \neq |y_{t'}|. \tag{3}$$

Notice that the conditions

$$\left\{ \begin{array}{rcl} \mathcal{S} & \subset & \{1, \ldots, n\} \\ |\mathcal{S}| & = & n/2 \\ \min_{t \in \mathcal{S}} |x_t| & > & \max_{t \notin \mathcal{S}} |x_t| \end{array} \right. \tag{4}$$

uniquely specify the set $\mathcal{S}$ consisting of $n/2$ indices of the maximum magnitudes of the vector $\boldsymbol{x}$, as it follows from (2). Furthermore, by (3), the conditions

$$\left\{ \begin{array}{rcl} \mathcal{S}' & \subset & \{1, \ldots, n\} \\ |\mathcal{S}'| & = & n/2 \\ t \in \mathcal{S}' & \Rightarrow & \mathrm{sgn}(y_t) = b_t \\ \min_{t \in \mathcal{S}'} |y_t| & > & \max_{t \notin \mathcal{S}' :\, \mathrm{sgn}(y_t) = b_t} |y_t| \end{array} \right. \tag{5}$$

uniquely specify at most one set $\mathcal{S}'$ consisting of $n/2$ indices of the maximum magnitudes of the vector $\boldsymbol{y}$ located at positions where their signs coincide with the corresponding signs stored in the vector $\boldsymbol{b}$. If there is no set $\mathcal{S}'$ satisfying (5), then the verifier makes the rejection decision.

We will present the procedures above in a more detailed form using the permutations that sort all components of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

• **The BMW encoding**. *Construct the permutation* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ *that sorts components of the vector* $|\boldsymbol{x}|$ *in the decreasing order, i.e.,*

$$\{\pi_1, \ldots, \pi_n\} = \{1, \ldots, n\} \textit{ and } |x_{\pi_1}| > \cdots > |x_{\pi_n}|.$$

Assign $\mathcal{S} = \{\pi_1, \ldots, \pi_{n/2}\}$ *and construct the pair of vectors* $(\boldsymbol{b}, \boldsymbol{s})$ *whose components are defined by*

$$b_t = \left\{ \begin{array}{ll} \mathrm{sgn}(x_t), & \textit{if } t \in \mathcal{S} \\ \overline{\mathrm{sgn}}(x_t), & \textit{if } t \notin \mathcal{S}, \end{array} \right.$$

$$s_t = \begin{cases} 1, & \textit{if } t \in \mathcal{S} \\ 0, & \textit{if } t \notin \mathcal{S}. \end{cases}$$

The t-th component of the vector $\boldsymbol{x}$ will be referred to as a significant component if $t \in \mathcal{S}$, and as a non-significant component if $t \notin \mathcal{S}$.

The BMW decoding, described below, can be presented as a generalized version of the encoding.

• **The BMW decoding**. *Construct the permutation* $\boldsymbol{\pi}' = (\pi'_1, \dots, \pi'_n)$ *that sorts components of the vector* $|\boldsymbol{y}|$ *in the decreasing order, i.e.,*

$$\{\pi'_1, \dots, \pi'_n\} = \{1, \dots, n\} \textit{ and } |y_{\pi'_1}| > \cdots > |y_{\pi'_n}|.$$

Construct the set $\mathcal{S}'$ *by the following rules:*

(a) set $j = 1$;

(b) if $\operatorname{sgn}(y_{\pi'_j}) = b_{\pi'_j}$, *then include* π'_j *into the set* $\mathcal{S}'$;

(c) if $|\mathcal{S}'| < n/2$ *and* $j < n$, *then increase* j *by* 1 *and go to (b);*

(d) output the set $\mathcal{S}'$.

Construct the vector $\boldsymbol{s}'$ *whose components are defined as follows:*

$$s'_t = \begin{cases} 1, & \textit{if } t \in \mathcal{S}' \\ 0, & \textit{if } t \notin \mathcal{S}'. \end{cases}$$

Example. Let $n = 8$ and the observation channel is an additive white Gaussian noise channel with the variance 1. Suppose that

$$\boldsymbol{r} = (+1.0, -1.2, +1.7, +0.3, -2.4, +0.1, -0.7, -1.3),$$

$$\boldsymbol{r}' = (+1.35, -0.76, +2.05, -0.1, -1.3, -0.06, -0.94, -1.65).$$

Then

$$\boldsymbol{x} = (+.68, -.77, +.91, +.24, -.98, +.08, -.52, -.81),$$

$$\boldsymbol{y} = (+.82, -.55, +.96, -.08, -.80, -.05, -.65, -.90),$$

since using (1) we have $x_t = \operatorname{erf}\Big(\frac{r_t}{\sqrt{2}}\Big)$, $y_t = \operatorname{erf}\Big(\frac{r'_t}{\sqrt{2}}\Big)$, $t = 1, \dots, n$, where $\operatorname{erf}(\cdot)$ is the erf-function:

$$\operatorname{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r \exp\{-\tilde{r}^2\}\, d\tilde{r}, \quad r \in \mathsf{R}.$$

Table 1: Example of the encoding and the decoding for $n = 8$.

Encoding								
$\boldsymbol{x}$	$+.68$	$-.77$	$+.91$	$+.24$	$-.98$	$+.08$	$-.52$	$-.81$
$\boldsymbol{b}$	0	0	1	0	0	0	1	0
$\boldsymbol{s}$	0	1	1	0	1	0	0	1
Decoding								
$\boldsymbol{y}$	$+.82$	$-.55$	$+.96$	$-.08$	$-.80$	$-.05$	$-.65$	$-.90$
$\boldsymbol{b}$	0	0	1	0	0	0	1	0
$\boldsymbol{s}'$	0	1	1	0	1	0	0	1

Table 2: Permutations $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ for the example in Table 1. The indices, specifying positions of ones in the vectors $\boldsymbol{s}$ and $\boldsymbol{s}'$, are given in the bold font.

$\boldsymbol{x}$	$+.68$	$-\,\mathbf{.77}$	$+\mathbf{.91}$	$+.24$	$-\mathbf{.98}$	$+.08$	$-.52$	$-\mathbf{.81}$
$\boldsymbol{b}$	0	0	1	0	0	0	1	0
$\boldsymbol{y}$	$+.82$	$-\mathbf{.55}$	$+\mathbf{.96}$	$-.08$	$-\mathbf{.80}$	$-.05$	$-.65$	$-\mathbf{.90}$
$\boldsymbol{\pi}$	**5**	**3**	**8**	**2**	1	7	4	6
$\boldsymbol{\pi}'$	**3**	**8**	1	**5**	7	**2**	4	6

The vectors constructed at the enrollment and verification stages are given in Table 1. Table 2 contains permutations $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ that sort the magnitudes of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in the decreasing order. Remark that $\mathcal{S} = \{5,3,8,2\}$, $\mathcal{S}' = \{3,8,5,2\}$. As a result, the decoding is correct, since $\{5,3,8,2\} = \{3,8,5,2\}$.

The main line of our considerations is as follows. We develop necessary tools for the analysis of the basic verification scheme (the set of components of the input vector is divided into two parts) under basic assumptions (the input data are generated by a stationary memoryless source with the known probability distribution). At the same time, we show the points where the scheme and the assumptions can be modified to reach the conditions of practical situations. These considerations are oriented to the applications for finite lengths, and we present the methods of computing the exact values of the decoding error probability.

4 Conclusion

We consider the combinatorial algorithm, where the magnitudes and the signs of the constructed vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, are taken into account. The spplitting of data in magnitudes and signs is introduced to satisfy the cryptogaraphy requirements.

• The F-transformation allows us to separate the probabilistic and combinatorial arguments, and it simultaneously preserves the property that significant components are transmitted over a more reliable observation channel. More precisely, if data are generated by a stationary memoryless source and F is the probability distribution of the source, then the probability distribution over the $(-1, +1)$ set is uniform.

• The possibilities of an attacker, who wants to submit an artificial vector to pass through the verification with the acceptance decision, are very limited, when the decision is made on the basis of the encrypted version of the binary vector, specifying positions of significant components. The possibilities of an attacker in guessing of the input data using the information leakage, coming from data stored in the database, are also restricted.

References

[1] R. Ahlswede, ”Coloring hypergraphs: A new approach to multi–user source coding”, *Journal of Combinatorics, Information and System Sciences*, no. 4, pp. 76–115, 1979.

[2] R. Ahlswede, I. Csiszár, ”Common randomness in information theory and cryptography, Part I: Secret sharing”, *IEEE Trans. Inform. Theory*, no. 39, 1121–1132, 1993.

[3] V. B. Balakirsky, A. J. Han Vinck, ”Biometric authentication based on significant parameters”, *Lecture Notes in Computer Science: Biometrics and ID management*, no. 6583, pp. 13–24, 2011.

[4] V. B. Balakirsky, A. J. Han Vinck,”Algorithms for processing biometric data oriented to privacy protection and preservation of significant parameters”, *New Trends and Developments in Biometrics*, no. 13, pp. 305–333, 2012.

White-Box Encryption Algorithm based on SAFER+

Gurgen Khachatrian and Martun Karapetyan

American University of Armenia
Yerevan, Armenia

Abstract

White-box cryptosystems aim to be executed in untrusted environments where the attacker sees not only the input and output of the encryption routine but also every intermediate computation that happens along the way. The attacker can not only see the contents of the memory of the execution device, but also alter the execution at will. In 2002, Chow, Eisen, Johnson and van Oorschot presented the first white-box implementation of AES algorithm [2], which was shown to be insecure against the BGE attack presented by Billet, Gilbert and Ech-Chatbi in 2004 [4]. In 2010, another white-box AES implementation was presented by Karroumi [9], which was supposed to withstand the BGE attack. In 2013, De Mulder, Roelse, and Preneel showed, that Karroumis and Chows implementations are equivalent [14], I.E. the BGE attack can be successfully applied to both. In this paper a design and security analysis of a white-box encryption algorithm based on SAFER+ block cipher is presented which is shown to be secure against the BGE attack [4].

1 Introduction

In the black-box encryption model an encryption/decryption operation is executed in an environment, where the attacker can see the inputs and outputs of the encryption routine but has no access to any intermediate value generated during the execution. In some execution environments the attacker may gain access to the whole memory of the device, which will allow him/her to see all the intermediate values generated during the execution of the algorithm, including the secret key itself. White-box algorithms are designed to perform the same encryption operations in such a way, that key extraction is not possible even in such untrusted environments. Look-up tables are used to perform the cryptographic operations, and the key extraction is impossible given those tables. Chow, Eisen, Johnson and van Oorschot presented the first white-box implementation of AES in 2002 [2], which was broken in [4]. Attempts to securely implement AES white-box encryption were further continued in [9] (broken in [14]), [12] (broken in [13]). The BGE attack [4] is the main attack successfully applied to the AES white-box implementations. In this paper we represent a white-box encryption based on SAFER+ algorithm and show that it is secure against the BGE attack. The paper is organized as follows: In the sections 2 and 3, black and white-box encryption rationale of SAFER+ is represented. Section 4 is devoted to the detailed security analysis of SAFER+ white-box encryption. Section 5 includes results of the computational speed and memory requirements of SAFER+ white-box encryption described. The paper ends with the conclusion.

2 SAFER+ black-box encryption

The underlying symmetric encryption algorithm for the white-box encryption is based on SAFER+ algorithm [1] with 128-bit key running 6 rounds plus and extra key addition at the end of 6^{th} round. After 6 rounds of encryption, an extra 16-byte key is simply XOR-ed with the output of the 6^{th} round.

The round encryption procedure is graphically shown in the Figure 1. Below are the meanings of the functions used during SAFER+ round.

a) *xp* is based on exponential function $45^x = Y\ mod\ 257$.

b) *log* is based on logarithmic function $log_{45}(x) = Y\ mod\ 257$.

c) *xor* stands for regular XOR of X and Y

d) *add* stands for the sum between X and Y with mod 256.

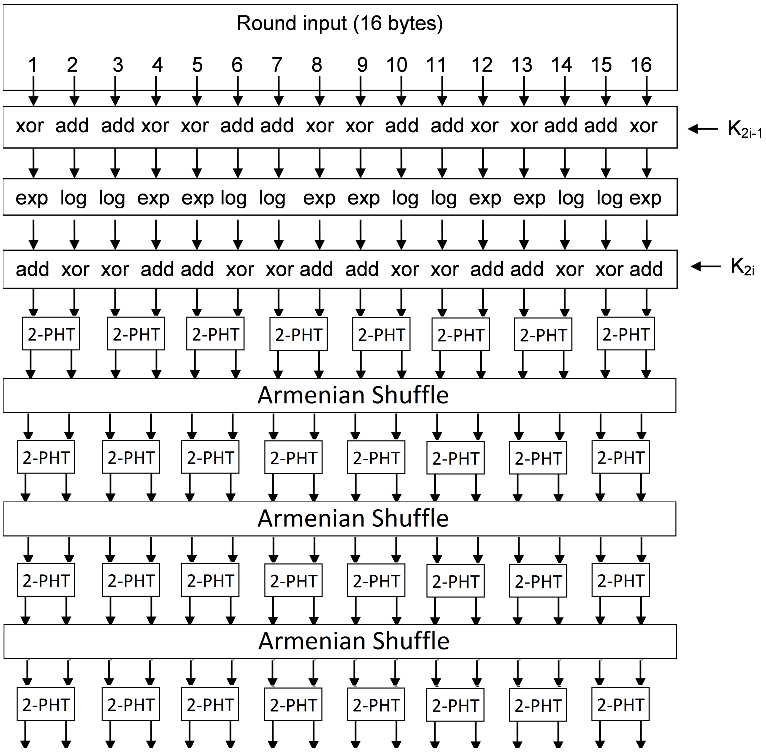

Round input (16 bytes)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
xor add add xor xor add add xor xor add add xor xor add add xor
K_{2i-1}
exp log log exp exp log log exp exp log log exp exp log log exp
add xor xor add add xor xor add add xor xor add add xor xor add
K_{2i}
2-PHT
Armenian Shuffle
2-PHT
Armenian Shuffle
2-PHT
Armenian Shuffle
2-PHT

Figure 1: Round structure of SAFER+

3 White-box encryption design rationale

Our White-box encryption is based on SAFER+ algorithm described in [1]. We're merging several steps of encryption rounds into look-up tables and obfuscate the table's inputs and outputs with input/output encodings (bijections) similar to those first introduced by Chow in [2]. The round confusion operations, namely key additions and non-linear layer evaluation, can be combined into 16 tables which map 1 byte of input to 1 byte of output in following manner:

for $r = 1$

$$T_i^1(x) := exp(x^1_{k_{i_1}}) + k^1_{i_2} \quad i \in A \tag{1}$$

$$T_i^1(x) := log(x + k^1_{i_1}) \oplus k^1_{i_2} \quad i \in B \tag{2}$$

for $2 \leq r \leq 6$

$$T_i^r(x) := exp(x \oplus k^r_{i_1}) + k^r_{i_2} \quad i \in A \tag{3}$$

$$T_i^r(x) := log(x + k^r_{i_1}) \oplus k^r_{i_2} \quad i \in B \tag{4}$$

for r=7

$$T_i^r(x) := x \oplus k_i^{13} \quad 1 \leq i \leq 16 \tag{5}$$

Though, for security reasons, we are randomly dividing the output into 2 numbers that sum up to the real value. So taking this into account, we need 16 tables which map 1 byte to 2 bytes for round 1, 2 bytes to 2 bytes for rounds 2 to 6, and 2 bytes to 1 byte for round 7 in the following manner. Let's denote the i^{th} mapping for input value x for round r=1 as $T^1_{i_1}(x)$ and $T^1_{i_2}(x)$, i^{th} pair of mappings of the r^{th} round, for rounds r=2..6, for input values of x_1 and x_2 by $T^r_{i_1}(x_1, x_2)$ and $T^r_{i_2}(x_1, x_2)$ and for round r=7, i^{th} mapping for input values x_1 and x_2 by $T^7_i(x_1, x_2)$. We split the 16 bytes indexes for each round into two sets A= 1,4,5,8,9,12,13,16 and B=2,3,6,7,10,11,14,15 according to the order in which exponential (exp) and logarithmic (log) functions are applied. Taking into account also the input and output permutations IP and OP; we can make the following definitions:

for $r = 1$

$$T^1_{i_1}(x) := exp(IP_i(x) \oplus k^1_{i_1}) + k^1_{i_2} + R^1_i(x) \quad i \in A \tag{6}$$

$$T^1_{i_1}(x) := log(IP_i(x) + k^1_{i_1}) \oplus k^1_{i_2} + R^1_i(x) \quad i \in B \tag{7}$$

$$T^1_{i_2}(x) := -R^1_i(x) \quad 1 \leq i \leq 16 \tag{8}$$

for $2 \leq r \leq 6$

$$T^r_{i_1}(x_1, x_2) := exp((x_1 + x_2 - S_i) \oplus k^r_{i_1}) + k^r_{i_2} + R^r_i(x_1, x_2) \quad i \in A \tag{9}$$

$$T^r_{i_1}(x_1, x_2) := log(x_1 + x_2 - S_i + k^r_{i_1}) \oplus k^r_{i_2} + R^r_i(x_1, x_2) \quad i \in B \tag{10}$$

$$T^r_{i_2}(x_1, x_2) := -R^r_i(x_1, x_2) \quad 1 \leq i \leq 16 \tag{11}$$

for r=7

$$T^r_i(x_1, x_2) := OP_i((x_1 + x_2 - S_i) \oplus k^{13}_i \quad 1 \leq i \leq 16 \tag{12}$$

Where $R^r_i(x_1, x_2)$ is a random function for each round, byte and input values, S_i is a compensation for random values added in the 2-PHT boxes described later. For each round $1 \leq r \leq 6$ the confusion step is followed by layers of 2-PHT boxes and Armenian shuffles. Look-up tables will be provided for applying 2-PHT operation on each pair of outputs.

3.1 Structure of SAFER+ white-box algorithm round

The structure of SAFER+ round using encryption boxes is shown in Figure 1. Two types of boxes are used: E-Boxes and 2-PHT boxes. E-boxes are accounting for the round confusion operations, while 2-PHT boxes are similar to 2-PHT boxes of regular SAFER+ algorithm described in [1], except they operate on 2 pairs of bytes. Armenian shuffles also operate on 2 bytes, I.E. the permutation is applied on byte pairs, instead of single bytes.

3.2 E-boxes

E-Boxes are look-up tables for getting values of $T^1_{i_1}(x)$ and $T^1_{i_2}(x)$ for round $r = 1$, values of $T^r_{i_1}(x_1, x_2)$ and $T^r_{i_2}(x_1, x_2)$ for all x_1 and x_2 for rounds $2 \leq r \leq 6$ and values of $T^7_i(x_1, x_2)$ for round $r = 7$. We apply random encodings

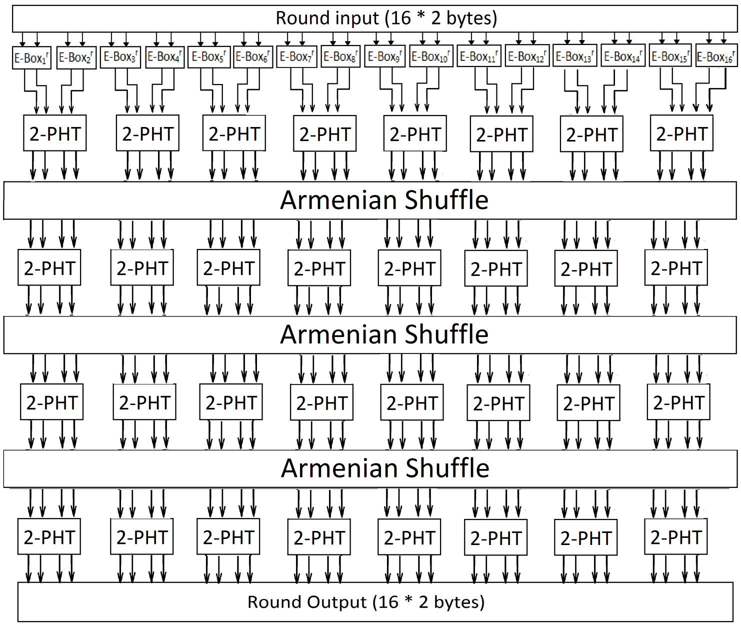


Figure 2: Round structure of SAFER+ White-box

to the E-boxes outputs, and the inputs for all the rounds are also encoded. 32 random output encodings (permutations) $f_1^r, f_2^r, ..., f_{32}^r$ are generated at white-box table generation phase per round for rounds $1 \leq r \leq 6$, where $f_i^r : Z_{256} \mapsto Z_{256}$, and thirty two input encodings $g_1^r, g_2^r, ..., g_{32}^r$ are generated for rounds $2 \leq r \leq 6$ $g_i^r : Z_{256} \mapsto Z_{256}$. We apply reverse permutations of g_{2*i-1}^r and g_{2*i}^r on the inputs of the $E - Box_i^r$ and output encodings f_{2*i-1}^r and f_{2*i}^r on the outputs. So we get the following formulas for the outputs of E-boxes:

for $r = 1$

$$P_{i_1}^1(x) := f_{2*i-1}^r(T_{i_1}^1(x)) \tag{13}$$

$$P_{i_2}^1(x) := f_{2*i}^r(T_{i_2}^1(x)) \tag{14}$$

for $2 \leq r \leq 6$

$$P_{i_1}^r(x_1, x_2) := f_{2*i-1}^r(T_{i_1}^r({g_{2*i-1}^r}^{-1}(x_1), {g_{2*i}^r}^{-1}(x_2))) \tag{15}$$

$$P_{i_2}^r(x_1, x_2) := f_{2*i}^r(T_{i_1}^r({g_{2*i-1}^r}^{-1}(x_1), {g_{2*i}^r}^{-1}(x_2))) \tag{16}$$

for $r = 7$

$$P_i^1(x_1, x_2) := T_i^7({g_{2*i-1}^r}^{-1}(x_1), {g_{2*i}^r}^{-1}(x_2)) \tag{17}$$

3.3 2-PHT boxes

2-PHT boxes are boxes that multiply the input with matrix $H = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, I.E. given $a, b \in Z_{256}$ on inputs, will produce values of $2 * a + b$ and $a + b$ on the outputs. In our white-box implementation 2-PHT boxes must operate on 4 bytes of input, because each input is divided into 2 values and encoded. One can easily notice that this kind of a box can be built with 2 standard 2-PHT boxes as shown in figure 3 and that the sum of pairs of the outputs $(2 * a_1 + b_1) + (2 * a_2 + b_2) = 2 * (a_1 + a_2) + (b_1 + b_2) and (a_1 + b_1) + (a_2 + b_2) = (a_1 + b_1) + (a_2 + b_2)$, I.E. 2-PHT operation can be successfully applied on 2 pairs of numbers. Random input and output encodings (permutations) are used on all the inputs and outputs on all the 2-PHT boxes. Input encodings of the first layer of 2-PHT boxes must match the output encoding of the corresponding E-boxes and the output encoding of the last layer of 2-PHT boxes must match the input encodings of the corresponding E-boxes. If input encodings f_1, f_2 and output encodings g_1, g_2 are used for a particular 2-PHT box, the outputs for input values of x_1, x_2 will be

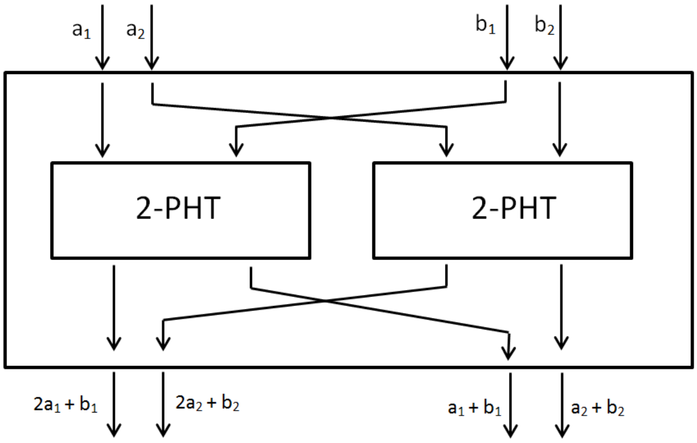


Figure 3: 2-PHT operation for 2 pairs of bytes

$$PHT_1(x_1, x_2) = g_1(2 * f_1^{-1}(x_1) + 2 * f_2^{-1}(x_2) + S_1) \tag{18}$$

and

$$PHT_2(x_1, x_2) = g_2(f_1^{-1}(x_1) + f_2^{-1}(x_2) + S_2) \tag{19}$$

Where S_1 and S_2 are 2 different random values which are generated for each 2-PHT box. The effect of these values is canceled out by S_i values used in the E-boxes.

Proposition 1. A white box encryption result described above is equivalent to the black box encryption result of SAFER+ accurate to the input and output permutations applied to the white box input and output.

Proof: The proof is straightforward. We have to compare two functions the one which makes a black box encryption based on the formulas 1-5 and formulas implementing white-box algorithm described in by formulas 13 - 17 and 18. One can easily notice that

$$T_{i_1}^r(x_1, x_2) + T_{i_2}^r(x_1, x_2) = exp((x_1 + x_2) \oplus k_{i_1}^r) + k_{i_2}^r = T_i^r((x_1 + x_2)i \in A \tag{20}$$

$$T_{i_1}^r(x_1, x_2) + T_{i_2}^r(x_1, x_2) = log(x_1 + x_2 + k_{i_1}^r) \oplus k_{i_2}^r = T_i^r((x_1 + x_2)i \in B \quad (21)$$

So E-boxes correctly apply the SAFER+ confusion step for input value $x_1 + x_2$. All the output permutations of E-boxes match the input permutations of the corresponding 2-PHT boxes, and the output encodings of the last layer of 2-PHT boxes matches the input encodings of E-boxes, so it's obvious that all the encodings cancel each other. So the only difference between ordinary black-box SAFER+ encryption procedure and white-box encryption procedure described above are the input and output encodings IP_i and OP_i applied accordingly to the input of the E-boxes of round 1 and the output of E-boxes of the final round.

3.4 Key generation schedule for SAFER+ white-box encryption

By the key schedule for SAFER+ White-box encryption we mean the key schedule for generating keys for corresponding regular black box encryption. SAFER+ key schedule details can be found in [1]. We modify the key schedule in order to eliminate any possibility of using the key schedule algorithm in key recovery attacks against the described white-box implementation. We use the same key schedule algorithm for generating the 2^{nd} key of the first round, after which each next key pair will be the SHA-256 hash value of the key pair of the previous round.

4 SAFER+ white-box encryption security analysis

It was shown in [1] that SAFER+ algorithm is secure against differential and linear cryptanalysis attacks after 6 rounds of encryption. Black box encryption described in current document is similar to SAFER+, except it applied input and output encodings for round 1 input and last round's output. Clearly these encodings will not reduce the security of black box encryption from the point of view of differential and linear cryptanalysis.

4.1 Security against BGE attack

An algebraic attack against White-box AES implementation was developed by Billet, Gilbert and Ech-Chatbi called BGE attack [4]. It is shown in [16] that BGE attack can be applied to all SLT (substitution-linear transformation network) ciphers which match the following definition.

Definition 1. A cipher is called Substitution-Linear Transformation (SLT) cipher if it can be specified as follows: It consists of R rounds where $R > 0$. A single round r is a bijective function $F^r_{SLT}(x_1, x_2, ..., x_s)$ on $GF(2^n)$ where $n = ms$ and $x_i \in GF(2^m)$. This function starts with the XOR-ing an n-bit length round key $k^r = (k^r_1, k^r_2, ..., k^r_s)$ with the input $x_1, x_2, ..., x_s$. That is, a value $y_i = x_i \oplus k_i$ is computed. Next the round computes $z_i = S^r_i(y_i)$ for all y_i where the (non-linear) invertible S-boxes $S^r_1, S^r_2, .., S^r_s$ are part of the cipher. These two steps achieve confusion. The diffusion is realized by multiplying the outcome of the S-boxes with an invertible matrix $M(r)$ over $GF(2^m)$. The BGE attack in general proceeds in three steps:

1) Transform the non-linear Output encodings $Q^r_{(i,j)}$ to and unknown $GF(2)$-affine transformation. (i.e. to recover the unknown non-linear part accurate to some unknown affine transformation).

2) Fully determine the $Q^r_{(i,j)}$ using the algebraic analysis of the known form of round function.

3) Obtain round keys for two consecutive rounds and recover the symmetric secret key using the reversibility of underlying key-schedule.

In our case the random input-output permutations of White-boxes f^r_i and g^r_i as well as the input/output encodings of all the 2-PHT boxes can be recovered up to an unknown affine transformation by applying the first step of BGE attack on the 2-PHT boxes. However the techniques used for determining the exact affine approximations will not work in our case. The 2 byte input/output structure of our E-boxes is significantly different from SLT ciphers and makes it impossible for the attacker to use the same technique to recover the permutations and round secret keys. Any attack targeted on only one output of an E-box will fail because of totally random values of $R^1_i(x)$, so the attacker must try to analyze both outputs $T^r_{i_1}(x_1, x_2)$ and $T^r_{i_2}(x_1, x_2)$ together, which makes it impossible to apply the BGE attack.

5 Speed and memory requirements

SAFER+ White-box encryption described in this document uses 16 E-boxes for round 1 of size 256 * 2 = 512 bytes, 80 E-box tables for rounds 2 to 6 each of them having size of 256 x 256 x 2 bytes=128 KB and another 16 E-boxes for the final transformation which have size equal to 256 x 256 bytes = 64 KB. Each of 64 2-PHT boxes has size of 256 x 256 x 2 bytes = 128 KB. So the memory required for all White-boxes will be approximately 18.5 MB. In overall the encryption operation looks up each E-box once, and each 2-PHT table r-1=6 times, so the encryption requires 304 table look-up operations.

6 Conclusion

In this paper we have presented a complete specification of a novel White-box encryption algorithm based on SAFER+ block cipher. We have also presented the results confirming that given White-box implementation is secure against the BGE attack that was successfully applied to the white-box AES implementations. Further cryptanalyses are required to proof resistance against other known attacks.

References

[1] Massey, G.Khachatrian, M.Kuregian, "Nomination of SAFER+ as a Candidate Algorithm for Advanced Encryption Standard (AES)" - *Represented at the first AES conference*, Ventura, USA, August 20-25, (1998)

[2] S. Chow, P. Eisen, H. Johnson, P.C. van Oorschot, "White-Box Cryptography and an AES Implementation", *In 9th Annual Workshop on Selected Areas in Cryptography (SAC 2002)*, Aug.15-16 2002.

[3] S. Chow, P. Eisen, H. Johnson, P.C. van Oorschot, "A White-box DES Implementation for DRM Applications", *In Proceedings of 2nd ACM Workshop on Digital Rights Management (DRM)*, 2002, *volume 2696 of Lecture Notes in Computer Science*, pp. 1-15.

[4] Olivier Billet, Henri Gilbert, Charaf Ech-Chatbi, "Cryptanalysis of a White-box AES Implementation", *In Selected Areas in Cryptography 2004 (SAC 2004)*, pages 227-240, 2004.

[5] E. Biham, A. Shamir, "Differential cryptanalysis against DES-like cryptosystems" *pp. 212-241 in Advances in Cryptology CRYPTO-90*, *Lecture notes in Computer Science N 537, Springer 1990.*

[6] M. Matsui, "Linear Cryptanalysis Method for DES cipher", *Advances in Cryptology-Eurocrypt 93*, pp. 386-397, *Lecture Notes in Computer Science, Springer N 765*, Springer 1994.

[7] J. L. Massey, G. H. Khachatrian and M. K. Kuregian, "Nomination of SAFER++ as Candidate Algorithm for the New European Schemes for Signatures, Integrity, and Encryption (NESSIE)", *Submission document from Cylink Corporation*, 2000.

[8] T.Lepoint, M. Rivain, "Another Nail in the coffin of Whitebox AES implementations", 2013.

[9] Mohamed Karroumi, "Protecting white-box AES with Dual Ciphers", *In Kyung-Hyune Rhee and DaeHun Nyang, editors, Information Security and Cryptology - ICISC 2010*, *volume 6829 of Lecture Notes in Computer Science*, pages 278291. *Springer Berlin Heidelberg*, 2011.

[10] Julien Bringer, Herve Chabanne, and Emmanuelle Dottax, "White-box cryptography: Another attempt", 2006.

[11] Yoni Mulder, Brecht Wyseur, and Bart Preneel, "Cryptanalysis of a perturbated white-box AES implementation", *In Guang Gong and KishanChand Gupta, editors, Progress in Cryptology - INDOCRYPT*, 2010, *volume 6498 of Lecture Notes in Computer Science*, pages 292310.

[12] Yaying Xiao and Xuejia Lai, "A secure implementation of white-box AES", *In Computer Science and its Applications(CSA09)*, 2009.

[13] Yoni Mulder, Peter Roelse, and Bart Preneel, "Cryptanalysis of the Xiao Lai white-box AES implementation", *In Lars R. Knudsen and Huapeng Wu, editors, Selected Areas in Cryptography*, volume7707 of Lecture Notes in Computer Science, pages 3449. *SpringerBerlin Heidelberg*, 2013.

[14] Yoni De Mulder, Peter Roelse, and Bart Preneel, "Revisiting the BGE Attack on a White-Box AES Implementation", *IACR Cryptology ePrint Archive*, 2013.

Experiences in Building an Enterprise Data Analytics

Ashot N. Harutyunyan, Arnak V. Poghosyan, and
Naira M. Grigoryan
VMware
Email: {aharutyunyan;apoghosyan,ngrigoryan}@vmware.com

Abstract

The Information Age made data easily accessible and omnipresent, currently with features of big volume, high velocity, and large variety, never seen before. For sciences, that is an unbelievable opportunity to explain the world better. Moreover, in the post-Information Age, businesses make any attempt to collect data and deeply benefit from it to achieve highly innovative technologies in terms of automation, performance, and efficiency. We share our experiences in building an enterprise data analytics for managing modern cloud computing infrastructures, as well as make parallels with information theory problems.

1 Introduction

In the era of Big Data, technologies are made of data. They increasingly tend to design smart applications with data-driven intelligence to profoundly benefit from the advantage of measured/monitored storm of data overwhelming human capabilities to process it and retrieve actionable knowledge. Therefore, industries are vastly investing in building relevant data analytics platforms and solutions to adequately address the challenges of the new era. Those challenges imply research for novel data scientific and machine learning approaches for real-time and proactive view into various systems.

As a provider of software-defined data centers and cloud computing infrastructures through virtualization, VMware dominates in the market of management of those systems by measuring and leveraging data. To have the full and proactive view of the systems real-time in terms of performance (health), capacities (IT resources), configuration and compliance, company's cloud management solutions [1,2] monitor both sources of IT data: structured and unstructured, respectively. The goal is to effectively and efficiently manage hundreds of thousands of IT objects with millions of different parameters (metrics) over time, terabytes of logs per day, and millions of events for anomaly identification and/or prevention. Building a generic data analytics platform to target such a goal in a context-independent way is a hard problem. Our experiences in providing a real-time performance analytics for data centers are summarized in a system (see [3], [6,7], and [8]) of several modules encompassing (Fig. 1)

- *behavioral analysis* for time series data and *extreme value analysis*: typical vs. atypical behavior to judge about anomalies based on data categorization, change point and periodicity detections;
- *abnormality degree* estimation for an outlying process to measure its severity or form an anomaly event;
- *ranking of events* in terms of their impact factor and *problem root causing*;
- *principal feature analysis and event reduction*;
- *data compression*;
- *prediction of alterations* in the system (allows sparing computational resources needed to run expensive behavioral pattern extraction procedures)

and other building blocks.

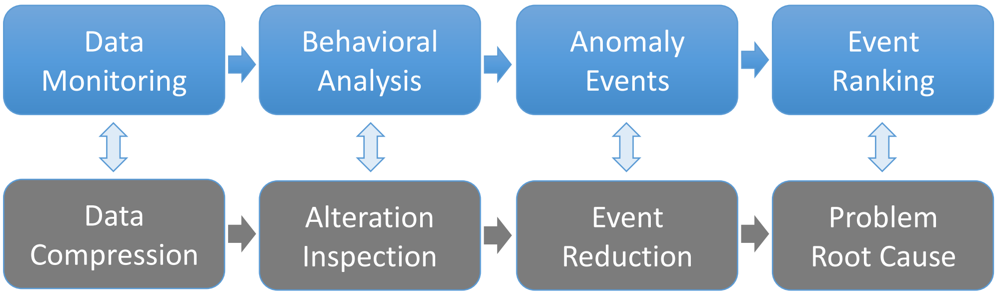


Fig. 1. Building blocks of a multi-layer analytics.

2 The Role of Information Theory

Concepts of information theory and its measures help in tackling problems we are working on, such as pattern and anomaly detection in logs [4,5], extreme value analysis applying a maximum entropy principle [8], identification of problem root causes [3], and feedback-enhanced analytics [7].

We observe inspirational parallels with information theory especially when dealing with data compression problems. One of our solutions in this domain applies correlations within data sets to compress metrics in a "meaningful" and efficient way. It parallels with Wyner-Ziv coding (Fig. 2). Specifically, the method is based on finding the principle (independent) features of a metric group (in other words, its basis) and compress the rest of dependent metrics using the corresponding linear combination coefficients. Moreover, a uni-variate (for a single metric) lossy compression method subject to fidelity criteria we are developing currently parallels with rate-distortion theory. The main idea behind the approach is to design a data compression model subject to the application needs such as anomaly detection, preserving relevant "interesting" patterns with high resolution and losing accuracy in other patterns.

Below we would like to enclose an example of using entropy measure to estimate confidence of users in beliefs they utilize as our data-driven recommendations and adjust those beliefs accordingly (see [7]), as well as show how it applies as a generic tool for ranking items based on user ratings as additional entertaining examples.

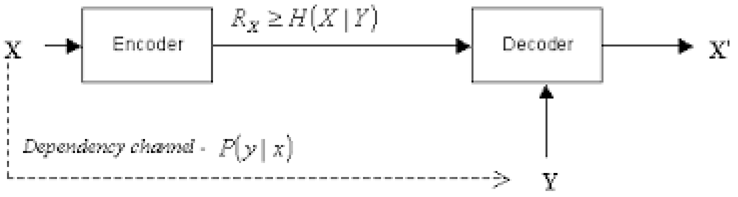


Fig. 2. Wyner-Ziv source coding.

Anomaly Detection in Logs with User Feedback. Here the basic data-agnostic analysis concerns the anomaly detection problem in log files via Dynamic Normalcy Graphs (DNG) [5]. We show how the general approach on feedback-based belief evaluation in [7] specializes into a specific solution for this correlation-based abnormality analysis and recommendation generation. We assume that the main notations and concepts of the work [7] can be used without detailed explanations. In summary, the goal is to enhance the efficiency of the DNG as a causation tool via processing of user feedback statistics on correlation breakage alarms. That can be performed if we evaluate the confidence for each correlation (belief) in DNG from that statistics and apply it in computation of abnormality degree of data stream. In some sense, it is an update of the conditional probabilities in DNG.

Formally, DNG is a collection of beliefs each representing the conditional probability from node j to i, so we denote the corresponding beliefs by $B_{i,j} = P(i \mid j)$. Let the user be asked to answer to the question if he/she is satisfied (and to what degree) by abnormality recommendation regarding a missing event type which is related to belief $B_{i,j}$. And let the users provide feedback taking value from $[0,1]$ (assume $l = 3$ and quantized feedback are from the intervals $[0,0.25)$, $[0.25,0.75)$, and $[0.75,1]$) at each time t_k when facing the breakage in $B_{i,j}$. So the feedback for $B_{i,j}$ is a series of ratings:

$$F(B_{i,j}) \equiv \left\{f(t_k, B_{i,j})\right\}_{k=1}^{K} \equiv \left\{f_k(B_{i,j})\right\}_{k=1}^{K}$$

where 1 is full satisfaction and 0 is complete dissatisfaction.

Based on $F(B_{i,j})$ we want to make a convergence evaluation in user opinion and output a confidence $C(B_{i,j})$ which can be incorporated into the existing abnormality analysis to tune its performance (an optimization of false positive alarms). This confidence will support the degree of validity of the initial belief $B_{i,j}$ and lead to a new DNG-based mismatch calculation. In other words, an updated $P'(i \mid j)$ can be obtained which is a combination of original DNG and one obtained from the feedback processing. All the postulates formulated in Section II of [7] are valid here. If there is a convergence to some degree of user satisfaction in the recommended beliefs, then the basic conditional probabilities are updated for further usage in anomaly detection with their confidences. The results in the original probabilities can increase or decrease as new beliefs about the system are incorporated. Correspondingly, their role in abnormality (mismatch) computation may change.

The notations in Section II of [7] easily apply to the beliefs $B_{i,j}$ of the DNG. According to the same reference, in case of comparably large entropy

$$H(\bar{S}(B_{i,j})) \in \left(1 - \frac{2}{3}\log_3 2, 1\right]$$

there is no convergence in feedback and hence the system has no update. The value $1-(2/3)\log_3 2$ corresponds, for instance, to the following scenario: $h_1 = 0$, $h_2 = 1/3$, $h_3 = 2/3$. If entropy is smaller than $1-(2/3)\log_3 2$ we determine the needed confidence $C(B_{i,j})$ by checking at which interval is the bias in the uncertainty. The bias is determined by the mode of the histogram $h_{\max} = \max\{h_1, h_2, h_3\}$.

Let $m(h_{\max})$ be the weighted average of the values $S(f_k(B_{i,j}))$ calculated by the technique of [7] (Section II) and corresponding to the mode $h_{\max}$. Then the confidence of $B_{i,j}$ is determined by the entropy:

$$C(B_{i,j}) = 1 - H(\bar{S}(B_{i,j}))$$

This means, for instance, that if the entropy is high, then the confidence in user feedback is zero:

$$C(B_{i,j}) = 0, \text{ if } H(\bar{S}(B_{i,j})) \geq 1 - \frac{2}{3}\log_3 2 .$$

If $C(B_{i,j})$ is a positive, then we define a feedback-based belief $B^f_{i,j}$ or $P_f(i \mid j)$ as

$$P_f(i \mid j) \equiv m(h_{\max}).$$

Now we can combine the basic data agnostic and the feedback-based beliefs on our DNG to have a new belief as

$$P'(i \mid j) = \alpha P(i \mid j) + (1-\alpha) P_f(i \mid j)$$

where

$$\alpha = 1, \text{ if } C(B_{i,j}) = 0,$$

entropy is big and

$$\alpha = C(B_{i,j}), \text{ if } C(B_{i,j}) > 0$$

entropy is low, there is a convergence.

In this way the basic DNG transforms to a new correlation structure.

To further experiment with the entropy-based confidence we applied the prototype algorithm to two public databases on consumer ratings for books and movies. In these cases, we deal with equally ranked prior recommendations for items. In other words, all the books and films included in those data sets have initial beliefs ranked to 1 until the user feedback modifies that basis assumption. So, what the program outputs on both databases is a feedback-based ranked list of items.

The next subsections demonstrate some results obtained while experimenting with the mentioned data sources, respectively. Note also that we produce two categories of item lists, one for items with converged user opinion (positive confidence) and the other with uncertain results. For the latter we do not show any rank, although it can be performed if we relax the requirement on the confidence.

In both experiments the ratings interval quantization level is $l = 3$.

Results for Book Ratings. The first experiment was performed on the book ratings data from http://www.informatik.uni-freiburg.de/~cziegler/BX/. We fed our algorithm with 600,000 ratings by 278,859 users on 271,379 books. For this data portion, we included into our analysis only the set of books that have been rated at least 30 times.

Note that the ratings timestamps are not available in the dataset. Therefore, the ratings temporal weighting is not applied in this case.

Table 1 illustrates the first three highest rank books according to our algorithm and another three famous works (“The Little Prince”, “Animal Farm”, and “Lolita”) from 20th century that are of high rank but comparably low in overall feedback confidence. Moreover, those classics exhibit a larger uncertainty in the user feedback (perhaps also due to the fact that the first three works led to popular films).

Table 2 shows works by popular authors that exhibit severe disparity in reader opinions. This means that the uncertainty in reader’s rating is so high that they are within the most disagreeable items in our analysis, although being historically significant and impactful works.

Table 1. Several books with their ranks.

Title	**Author**	**Year**	**Conf.**	**Rank**
The Return of the King	Tolkien	1955	0.81	0.99
Harry Potter and the Goblet of Fire	Rowling	2000	0.83	0.99
Charlotte's Web	White	1952	0.68	0.99
The Little Prince	de Saint-Exupéry	1943	0.61	0.98
Lolita	Nabokov	1955	0.55	0.96
Animal Farm	Orwell	1945	0.54	0.96

Table 2. Several books with zero confidences.

Title	**Author**	**Year**	**Conf.**
Call of the Wild	London	1903	0
Jonathan Livingston Seagull	Bach	1970	0
The Catcher in the Rye	Salinger	1951	0
The reader	Schlink	1995	0

Results for Movie Ratings. The second data set processed by our approach was from http://www.grouplens.org/node/73, namely the 100k-MovieLens rating database (20,000 ratings by 459 users on 1410 movies). We included into our analysis only the set of those movies that have been rated at least 15 times.

Table 3 displays some of the highest ranked movies (with rather distinct confidences) that are of different eras. For comparison, the well-known IMDB rating (varying from 1 to 10) for listed films is also included. Note that the entropy-based rank coincides with IMDB rating for Godfather.

Table 3. Several films with their ranks.

Title	**Release**	**Conf.**	**Rank**	**IMDB Rating**
Taxi Driver	1996	0.85	1	8.5
Three Colors: Red	1994	1	1	8
12 Angry Men	1957	0.65	1	8.9
Casablanca	1942	0.79	1	8.7
Pinocchio	1940	0.67	1	7.6
The Wizard of Oz	1939	0.46	0.99	8.2
Amadeus	1984	0.61	0.97	8.4
Godfather	1972	0.61	**0.92**	**9.2**
Schindler's List	1993	0.59	0.92	8.9

The movies in Table 4 are examples of high IMDB rated films, including Academy award-winning ones, with wildly varying audience opinions, since the users' ratings show high uncertainty.

Table 4. Films with zero confidences.

Title	**Release**	**Conf.**	**IMDB Rating**
Mighty Aphrodite	1995	0	7
The Lion King	1994	0	8.4
The Fifth Element	1997	0	7.6
Men in Black	1997	0	7.2
Toy Story	1995	0	8.3
Twelve Monkeys	1995	0	8.1
Seven	1995	0	8.7

References

[1] VMware vRealize Operations Manager, http://www.vmware.com/products/vrealize-operations.html.

[2] VMware vRealize Log Insight, http://www.vmware.com/products/vrealize-log-insight.html.

[3] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "An anomaly event correlation engine: Identifying root causes, bottlenecks, and black swans in IT environments," *VMware Technical Journal*, vol. 2, no. 1, pp. 35-45, 2013.

[4] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Pattern detection in unstructured data: An experience for a virtualized IT environment," IFIP/IEEE International Symposium on Integrated Network Management, Ghent, Belgium, May 27-31, pp. 1048-1053, 2013.

[5] A.N. Harutyunyan, A.V. Poghosyan, N.M. Grigoryan, and M.A. Marvasti, "Abnormality analysis of streamed log data," Proc. *IFIP/IEEE Network Operations and Management Symposium*, May 5-9, Krakow, Poland, pp. 1-7, 2014.

[6] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "An enterprise dynamic thresholding system", Proc. USENIX 11th International Conference on Autonomic Computing, June 18-20, Philadelphia, PA, pp. 129-135, 2014.

[7] M.A. Marvasti, A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Ranking and updating beliefs based on user feedback: Industrial use cases," Proc. 12th *IEEE International Conference on Autonomic Computing*, July 07-10, Grenoble, France, pp. 227-203, 2015.

[8] A.V. Poghosyan, A.N. Harutyunyan, and N.M. Grigoryan, "Managing cloud infrastructures by a multi-layer data analytics," Proc. *IEEE International Conference on Autonomic Computing*, July 18-22, Wuerzburg, Germany, pp. 351-356, 2016.