Skolkovo Institute of Science and Technology

SIGNATURE CODES FOR MULTIPLE ACCESS CHANNELS, DIGITAL FINGERPRINTING CODES AND SYMMETRIC GROUP TESTING

Doctoral Thesis

by

ELENA EGOROVA

DOCTORAL PROGRAM IN COMPUTATIONAL AND DATA SCIENCE AND ENGINEERING

Supervisor
doctor of physical and mathematical sciences
Kabatiansky Grigory

Moscow - 2020

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Candidate (Elena Egorova)

Supervisor (Kabatiansky Grigory)

Abstract

The history of code division for *multiple access channels* (MAC, for short) started with the famous paper of Claude Shannon — "Two-way communication channels " (1961) [1], where it was shown that code division can outperform time division. During next decades, a lot of theoretical results were obtained, including the exact characterization of the *capacity region* for many different models of MAC. Despite that the *ordinary capacity* is widely known, on the contrary, the *zero-error capacity* is mainly unknown for the same MAC models. For instance, the zero-error capacity region is unknown for the binary adder channel with just two users, which is one of the simplest MAC model. Even less is known about the zero-error capacity for MAC with partial users activity, when not more than t users among total M users are active at each time slot. The last problem was originated in ALOHA type communication systems, i.e., in systems with random users access to a joint communication channel. It is very useful for such systems to know in advance which users will be active during a next communication session. The corresponding codes with zero-error probability are called *signature codes*. This thesis is devoted to the investigation of signature codes for MAC, digital fingerprinting codes, especially multimedia ones, group testing as well as their interplay.

The first chapter describes in details main objects of the thesis, namely, different models of MACs and codes for them, digital fingerprinting codes and especially multimedia fingerprinting codes, different variations of the group testing problem and the relationships between all these objects.

The second chapter is devoted to the A-channel. Based on the established relationship between signature codes for A-channel and separating codes we derive new upper and lower bounds on the rate of the best signature codes for A-channel. First construction of signature codes with efficient, i.e., polynomial complexity in the code length, coding and decoding procedures is proposed. Moreover, the case of A-channel with adversarial noise is considered and upper and lower bounds on the rate of signature codes resistant to such type of noise are proved.

In the third chapter we investigate signature codes for nonbinary B-channel and weighted adder channel. New upper and lower bounds on the rate of sig-

nature codes for nonbinary B-channel are proved. The proof of lower bound is based on the random coding method, and the proof of the upper bound is based on the entropy method. Explicit constructions of signature codes for weighted binary adder channel, which can be considered as some variation of B-channel, are proposed for noiseless and for adversarial noise cases.

In the last chapter we consider three possible applications of signature codes. The first application, known as multimedia digital fingerprinting (MDF) codes, is motivated by the digital right management (DRM). The idea of this technique is to construct the set of fingerprints that are embedded in the original digital content and that have the following property – if at most t users collude to produce a forged copy, then a distributor can identify at least one of these users. This property is often called *identifying parent property* or IPP, for short. It is shown that different reformulations of this problem for MDF codes are equivalent to signature codes for A-channel and for weighted binary adder channel. Based on this relationship we derive new results about MDF codes.

The second application devoted to newly introduced class of IPP codes, namely, constant weight IPP codes, which is a class of codes with identifying parent property and "simple decoding", i.e., tracing at least one guilty user by finding the nearest in the Hamming distance codeword. A new notion of *nonbinary* traceability codes (or set systems) introduced and an analogue of Gilbert-Varshamov bound for these codes is proved.

The last application is known as the symmetric group testing (SGT) problem that comes from the general group testing problem. The problem is to find the set of defective elements in a base set by conducting tests in which answers provide more information than classical group testing, namely, an answer says 0 if no defective samples in a given testing set, says 1 if all chosen samples are defective, and says $*$ - otherwise. In the case of non-adaptive search this problem is equivalent to the construction of signature codes for A-channel, what allows to derive new results on the minimal number of tests.

This Thesis contains not only some new results from so different areas of research as signature codes for multiple access channels, digital fingerprinting codes and symmetric group testing, but also a new approach which allows us to investigate these areas *uniformly*.

Publications

1. Egorova, E., Fernandez, M., Kabatiansky, G., (2020). On non-binary traceability set systems. Designs, Codes and Cryptography, accepted for publication, DOI: 10.1007/s10623-020-00749-4 (Q2 Scopus)

2. Egorova, E., Fernandez, M., Kabatiansky, G., Lee, M. H. (2019). Signature codes for weighted noisy adder channel, multimedia fingerprinting and compressed sensing. Designs, Codes and Cryptography, 87(2-3), pp. 455-462, DOI: 10.1007/s10623-018-0551-9 (Q2 Scopus)

3. Egorova, E. E., Potapova, V. S. (2018). Compositional restricted multiple access channel. Problems of Information Transmission, 2(54), pp. 116-123. DOI: 10.1134/S0032946018020023 (WoS)

4. Egorova, E., Fernandez, M., Kabatiansky, G., Lee, M. H. (2016). Signature codes for the A-channel and collusion-secure multimedia fingerprinting codes. In 2016 IEEE International Symposium on Information Theory (ISIT), pp. 3043-3047.

5. Kabatiansky, G., Fernandez, M., Egorova, E. (2016). Multimedia fingerprinting codes resistant against colluders and noise. In 2016 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 79-83.

6. Egorova, E., Potapova, V. (2016). Signature codes for a special class of multiple access channel. Proceedings XV International Symposium Problems of Redundancy in Information and Control Systems, pp. 38-42.

7. Egorova, E. (2019). Symmetric group testing. Proceedings XVI International Symposium on Problems of Redundancy in Information and Control Systems, pp. 38-42. (Scopus)

Contents

List of abbreviations

A-channel	Multi-frequency channel without intensity information
B-channel	Multi-frequency channel with intensity information
BAC	Binary adder channel
DRM	Digital right management
IPP	Identifiable parent property
MAC	Multiple access channel
MDF	Multimedia digital fingerprinting
OR-channel	Disjunctive channel
SGT	Symmetric group testing
wBAC	Weighted binary adder channel

Introduction

This PhD thesis was started from my simple observation that already known families of superimposed codes [2] can be used as digital fingerprinting codes for multimedia (MDF codes, for short), resistant to coalition's attacks, and the fact that these codes have rate R separated from zero, was unknown before my very first paper [3]. A bit later I found out that this observation wasn't totally new, see [4]. Then I improved this approach and derived that MDF codes are, in fact, t-signature codes for A-channel, where t is the maximal possible size of malicious coalitions. Moreover, it was proved in [5] that t-signature codes for A-channel can be used as MDF codes which can find a malicious coalition not only in the case of the so-called "averaging" attack, as it was considered in all previous papers, but also in the case of general linear attack. It became a bridge between the theory of signature codes and its applications to investigation of MDF codes [5]. The corresponding results can be found in chapters 1, 2 and 4.

On the contrary, a relationship between signature codes for MACs and non-adaptive group testing was well known for many years. Nevertheless, I found a new (despite rather obvious) relationship between symmetric group testing and signature codes for A-channel, and this approach is used in Chapter 4.

In the binary case the adder channel and B-channels coincide. In non-binary case they are rather different and asymptotic lower and upper bounds for B-channel were unknown. Note that B-channel provides the maximal possible information about the inputs of the channel and therefore its rate should be maximal among all q-ary MACs. V.Potapova and I independently found the corresponding lower and upper bounds [6], [7]. The corresponding results are described in chapter 3.

This research was not a one-way road, only from signature codes for MACs towards the fingerprinting codes and symmetric group testing. Detailed analysis of multimedia fingerprinting codes (MDF) has shown some weaknesses of quantized model and the corresponding coalition attacks. We consider a continuous model of MDF codes what led us to introducing a weighted adder channel, which is a generalization of the classical adder channel, and to constructing of more effective

families of codes protecting multimedia content from illegal redistribution. These results became parts of chapters 1 and 4.

Problem statement and its history

The problem of simultaneously transmitting many messages over a single communication medium is known as multiplexing and its history actually is as old as the communications itself. As popular examples, one may consider AM or FM radio, television or telephony, and nowadays — mobile communications. All these types of communication use some form of multiplexing, i.e., the transmission of multiple signals over a common channel in such a way that at a receiver side these signals can be separated with small or no interference among them. The most popular and studied forms of multiplexing are known as frequency division multiplexing (FDM), time division multiplexing (TDM), space division multiplexing (SDM), code division multiplexing (CDM) and orthogonal frequency-division multiplexing (OFDM).

The general idea of multiplexing schemes is to construct such set of signals that if these signals are mixed together to form a composite signal then the constituting parts (individual signals) can be recovered from the composite form. Mechanisms of forming a mixed signal and recovering the parts constitute the main difference of multiplexing schemes of different types.

The most studied example of code division multiplexing system with a single receiver is known as *the multiple access channel* (MAC). The study of MAC was initiated in the famous paper "Two-way Communication Channel" by Claude Shannon [1]. Let us start from a short description of MAC's mathematical model. MAC consists of M independent information sources (or users) and a common channel for information transmission to the single receiver, which is the same for all users. The i-th user has m_i messages to be transmitted, which are encoded by some code $C^{(i)}$ of the same cardinality m_i and length n over some finite set X, which is called the input alphabet. We assume that different users have the same alphabet X. We assume also that all M users maintain bit and word synchronization. Period of communications is split on sessions and during a session each user sends one of its messages by sending the corresponding codeword.

There is the following classification of MAC models:

1. Discrete or continuous time: for discrete time it is assumed that the session is divided in the finite number of time slots, and the signal is defined for

each time slot; for continuous time MACs inputs and output are considered as the functions of time $X(t), Y(t)$ with $t \in \mathbb{R}$.

2. Noiseless or noisy: for noiseless cases it is assumed that the transmission passes without any disturbance; for noisy case the real output might be different from what it was intended to be.

3. With or without feedback: for channels with feedback all senders receive a feedback information from the receiver and usually it is assumed that feedback is error-free.

4. Full or partial activity: when each user transmits a message and when only part of users has messages (information) to be transmitted. In the last case it could be that users has only single message to be transmitted, namely, to show that a given user is active. The corresponding codes with zero-error probability are called *signature codes*, and constitute the main object of this thesis.

5. Memoryless or with memory: for memoryless the output of the channel depends only on the inputs of the current session, and does not depend on previous input-output pairs as it is for channels with memory.

We shall consider *deterministic* memoryless multiple access channels with discrete time, without feedback, mainly with partial activity and noiseless as well as with noise (when we have a solution since noisy case is surely more difficult). MAC is defined by its input and the output alphabets X and Y correspondingly, where for a given input symbols $x_1, \ldots, x_M \in X$ the output of the channel equals $y = s(x_1, \ldots, x_M) \in Y$ and $s(.)$ is some map $X^M \to Y$, which is the main "ingredient" of MAC. We assume that the channel is memoryless and, hence, for input codevectors

$$\mathbf{c}^{(1)} = (c_1^{(1)}, \ldots, c_n^{(1)}), \ldots, \mathbf{c}^{(M)} = (c_1^{(M)}, \ldots, c_n^{(M)})$$

the output of the channel is

$$\mathbf{y} = S(\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(M)}) = (y_1, \ldots, y_n),$$

where $y_i = s(c_i^{(1)}, \ldots, c_i^{(M)})$.

The following transfer functions s generates the most popular models of deterministic MACs.

Examples.

(a) *Binary adder channel (Figure 1)*: $X = \{0, 1\}$, $Y = \{0, 1, 2, \ldots\} := \mathbb{N}_0$ and

$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} x_i, \tag{1}$$

where sum is the ordinary sum of integers and $\mathbb{N}_0$ denotes the set of all non-negative integers.

(b) *OR-channel*: $X = Y = \{0, 1\}$ and

$$s(x_1, \ldots, x_M) = \bigvee_{i=1}^{M} x_i \tag{2}$$

(c) *A-channel (Figure 2)*: $X = \{0, 1, \ldots, q-1\}$, $Y = \{0, 1\}^q$

$$s(x_1, \ldots, x_M) = (y_0, ..., y_{q-1}), \tag{3}$$

where $y_i = 0$ if $|\{j : x_j = i\}| = 0$ and $y_i = 1$ — otherwise

(d) *B-channel (Figure 2)*: $X = \{0, 1, \ldots, q-1\}$, $Y = \{\mathbf{y} = (y_0, y_1, \ldots, y_{q-1})\}$ and

$$s(x_1, \ldots, x_M) = (y_0, ..., y_{q-1}), \text{ where } y_i = |\{j : x_j = i\}| \tag{4}$$

(e) *Adder by mod 2 channel*: $X = \{0, 1\}$, $Y = \{0, 1\}$ and

$$s(x_1, \ldots, x_M) = x_1 \oplus x_2 \oplus \ldots \oplus x_M, \tag{5}$$

where $\oplus$ denotes the sum by mod 2.

(f) *Nonbinary adder channel*: $X = \{0, 1, \ldots, q-1\}$, $Y = \mathbb{N}_0$ and

$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} x_i, \tag{6}$$

where sum is the ordinary sum of integers.

(g) *Weighted binary adder channel*: $X = \{0, 1\}$, $Y = \mathbb{R}$ and

$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} \alpha_i x_i, \tag{7}$$

where α_i are some *nonzero* real numbers (called "weights", despite that α_i can be negative) and the sum is the sum of real numbers.

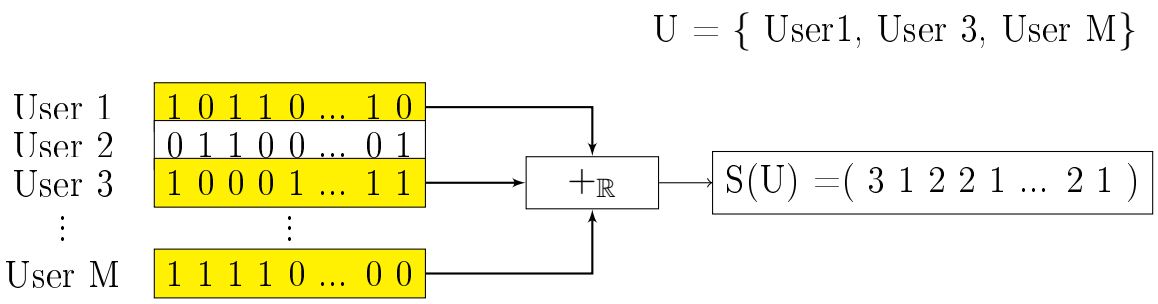

Figure 1: Model of the binary adder channel with user 1, user 3 and user M as active users

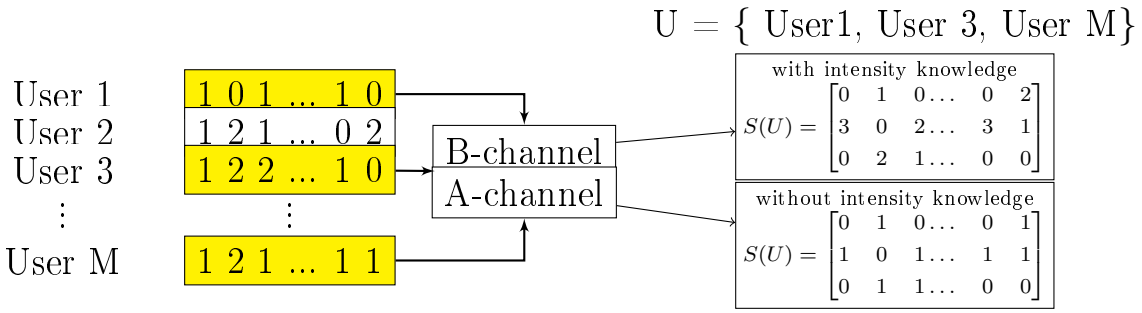

Figure 2: Model of A-channel and B-channel with user 1, user 3 and user M as active users

Classical information theory provides the exact solution for the ordinary capacity, i.e., when the probability of wrong decoding tends to zero, of general deterministic MAC model in the case of all-active users and blocklength tends to infinity, see [8, 9].

A lot of efforts were devoted to evaluation of the *zero-error capacity* for rather natural and simple models of MAC. Codes $C_1, \ldots, C_M$ are called *uniquely decodable* or *M-user code with zero-error probability* if the receiver can uniquely recover transmitted messages of all users by observing the corresponding output of MAC. Formally, it means that for any codevectors $\mathbf{c}^{(1)}, \hat{\mathbf{c}}^{(1)} \in C_1, \ldots, \mathbf{c}^{(M)}, \hat{\mathbf{c}}^{(M)} \in C_M$ if $S(\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(M)}) = S(\hat{\mathbf{c}}^{(1)}, \ldots, \hat{\mathbf{c}}^{(M)})$ then $\mathbf{c}^{(1)} = \hat{\mathbf{c}}^{(1)}, \ldots, \mathbf{c}^{(M)} = \hat{\mathbf{c}}^{(M)}$.

The rate of M-user code is, by the definition, M-dimensional vector $\mathbf{R} = (R_1, \ldots, R_M)$, where $R_j = n^{-1} \log |C_j|$ (here and below all logarithms are binary, except of cases when we shall use other base of the logarithm). Finding the *zero-error capacity region* means to find all achievable (when codes' length increase to infinity) rates $\mathbf{R}$ for a given MAC. The zero-error capacity region is mostly

unknown, in particular, for first four MACs described above in examples. In particular, it is unknown even for the *binary adder channel* (BAC) in the simplest case of two users, and it is one of the most intriguing topics of MAC theory today, see [10] for the latest results.

In this thesis we consider a particular case of M-user code with zero-error probability when at most t users are active, i.e., only they send information while other users are silent. Moreover, we will be interested in the case when all users have only one message to transmit, i.e., i-th user sends vector $\mathbf{c}_i \in X^n$ for sending message that he is active, and doesn't send a vector if he is inactive. Then, the property of being uniquely-decodable transforms to the so-called signature property defined as follows.

Definition. A code $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\} \subset X^n$ of the cardinality M and length n over the alphabet X is called a (t, M)-*signature code* or simply a t-*signature code* for a given MAC with the transfer function S if for any two different subsets $J, J' \subset \{1, \ldots, M\}$ such that $|J| \leq t$ and $|J'| \leq t$ the corresponding outputs are different also, i.e., $S(J) \neq S(J')$.

Signature codes are very closely related to some old and new problems in group testing and digital fingerprinting. These relationships will be described in details in Ch.1 and Ch.4. In particular, there are known relationships between some combinatorial problems of non-adaptive group testing (search of counterfeit coins among M coins on a spring scale, see [11, 12]), and signature codes for the binary adder channel with M users, what was observed in [13]. Another well known relationship is between signature codes for OR-channel and the classical group testing, introduced in 1943 [14]. There is a variation of the classical group testing called symmetric group testing which is not so well studied. We consider this problem by establishing its relation to signature codes for A-channel and derive new upper bound on the rate signature codes for A-channel as well as new lower bound on the minimal number of tests in symmetric group testing, see Ch. 1, 2 and 4.

A big part of this thesis is devoted to signature codes for two deterministic multiple-access channels — A-channel and B-channel. These channel models were proposed in [15] and are known as multi-frequency channels with and without intensity information, or as A-channel and B-channel, see examples 3 and 4 above, respectively. Let us explain how these MAC models are arisen in practice.

Consider the set $F = \{f_0, f_1, \ldots, f_{q-1}\}$ of q frequencies which can be employed by users for information transmission with the following restriction: each active user at each time slot employs only one frequency. The difference between

these two channel models occurs in the channel's output. Namely, the output of A-channel is the set of frequencies occurred as inputs of the channel at a given time slot, or the corresponding binary vector of length q. Whereas for the B-channel the output at a given time slot shows how many times each frequency was used for transmission, i.e., the output is an integer vector of length q with all coordinates being non-negative and their sum equals to $t' \leq t$ - the number of active users. Hence, replacing frequencies f_i on their q-ary "numbers" i leads us to MAC of examples 3 and 4 correspondingly.

Note, that for binary case B-channel and adder channel coincide. For non-binary case they are very different and obtaining new lower and upper bounds of signature codes for these channels constitutes probably the main part of the thesis.

It was mentioned above that my investigation of signature codes for different types of MACs has been started from an attempt to construct digital fingerprinting codes for multimedia which are better than known ones. Note that known at that time multimedia fingerprinting codes have rates which tends to zero when the code length increases. It's an obvious drawback which was overcame in my fist papers by usage of a new approach based on the relationship between signature codes for MAC and digital fingerprinting codes, see the corresponding explanation in Ch.1, and new results are given in Ch. 2 and 4.

Goals

Goals of the PhD thesis are the following:

- Derive new upper and lower bounds on the rate of signature codes for A-channel, B-channel and their modifications;
- Develop new constructions of the corresponding signature codes and their efficient decoding algorithms;
- Study these channels with noisy output;
- Investigate possible applications of constructed codes towards digital fingerprinting codes and symmetric group testing.

Scientific novelty of the work

All results obtained in this thesis are new. The following main results were obtained:

1. New lower and upper bounds on the rate of signature codes for B-channel.

2. A new explicit construction of signature codes for A-channel with efficient (polynomial time) decoding algorithm.

3. A new upper bound on the rate of signature codes for A-channel with adversarial noise.

4. A new class of signature codes for weighted adder channel with explicit construction and performance better than for known codes.

5. A new class of digital fingerprinting codes with simplified tracing traitors based on minimum distance decoding.

6. New classes of multimedia fingerprinting codes with significantly better performance than all previously known multimedia fingerprinting codes.

7. New upper bound for the minimal number of tests in symmetric group testing with adversarial noise.

Research Methodology

Throughout this thesis classical "probability method" (due to Paul Erdos) also known as "random coding" (due to Claude Shannon) is used in order to prove the existence of asymptotically good codes. Also some analytic methods and combinatorial coding theory methods were used.

Practical and theoretical significance of the work

The results of the thesis are mostly theoretical with description of possible practical applications. They may be useful to specialists working in information theory, combinatorial coding theory and cryptography.

Organization of the thesis

The thesis consists of an introduction, four chapters, conclusion and list of references.

In the first chapter we describe different models of MACs and known results about error-free codes for these MACs, and pay especial attention to relationships among signature codes for MACs, non-adaptive group testing and digital fingerprinting codes. In section 1.1 we describe the so-called "Zoo of MACs", and

establish a partial order on MACs, what allows us later to derive some results "for free". In section 1.2 we describe relationships between error-free codes for some MACs and corresponding non-adaptive group testing problem. In particular, we derive a new relationship between symmetric group testing and signature codes for A-channel and prove a new upper bound on the minimal number of tests in symmetric group testing. Last section is devoted to digital fingerprinting codes, especially to multimedia digital fingerprinting codes, and how to construct such codes from signature codes for the corresponding MACs (mainly, A-channel and weighted adder channel).

In the second chapter we introduce the general idea of coding for multiple access A-channel, in particular, the notion of signature codes. We establish the connection between signature codes for A-channel and separating codes. Recall that a q-ary code C is called (s,t)-*separating code* if for any two disjoint sets $V, U \subset C$ such that $|V| \leq s,\ |U| \leq t$ there exists at least one coordinate which separates them, i.e., there exists coordinate k s.t. $V_k \cap U_k = \emptyset$, see [16–19]. It is easy to see that any $(1,t)$-separating code is a t-signature code for A-channel, and a t-signature code for A-channel is, at the same time, a $(1,t-1)$-separating code. Then known results on separating codes imply the following upper and lower bounds on the rate of the maximal possible rate R_t^A of binary t-signature codes for A-channel, when the code length n tends to infinity and t is large but fixed:

$$\Theta(t^{-2}) \leq R_t^A \leq O\left(\frac{\log t}{t^2}\right).$$

In section 2.3 we provide the first construction of signature codes for A-channel with polynomial in code length complexity of decoding. We "pay" for this property by decreasing the code rate to t^{-3} instead of order t^{-2} for general signature codes for A-channel, but in return we decrease decoding complexity from exponential to polynomial. The proposed construction is based on the concatenation technique where binary $(1,t)$-separating codes are taken as inner codes and codes with large minimal code distance as outer codes (algebraic-geometry codes, Reed-Solomon codes, in particular), see [5].

In section 2.4 we consider A-channel with adversarial noise, i.e., when the channel output might be erroneous but in no more than L positions and these positions could be chosen in the worst for users way. The following bounds on the rate of binary t-signature codes that are able to correct up to L errors are proved [20, 21]. Firstly, the following upper bound is proved [21]:

$$R_t(\delta) \leq \frac{1}{t-1} R(\delta),$$

where $\delta = \frac{2L+1}{n}$ and $R(\delta)$ denotes the asymptotic maximal possible rate of a code in the Hamming space with relative distance δ. On the other hand, for any $\delta < \delta_{crit} = t^{-1}(1-t^{-1})^t$ the following lower bound holds

$$R_t(\delta) \geq \frac{2\log_2 e}{t}(\delta_{crit} - \delta)^2,$$

where $\delta_{crit} < (et)^{-1}$.

In the third chapter the signature codes for B-channel as well as for its modification, known as weighted adder channel, are considered. Denote by $M_q^B(n,t)$ the maximum possible size of a q-ary t-signature code for B-channel of length n, and by $R_q^B(n,t) = n^{-1}\log_q(M_q^B(n,t))$, the maximum possible rate of such a code. To simplify notations let us denote $R_q^B(t) := \lim_{n\to\infty} R_q(n,t)$ (in Ch. 3 we use correct notations since the limit may not exists). The main result of the chapter is the proof that for t large enough we have the following upper and lower bounds on the asymptotic rate $R_q^B(t)$ of the best t-signature codes for B-channel [6, 7]:

$$(q-1)\frac{\log_q t}{4t} - \frac{c_1}{4t} \leq R_q^B(t) \leq 2\left((q-1)\frac{\log_q t}{4t} + \frac{c_2}{4t}\right),$$

where $c_1 = c_1(q) = q(1-\log_q(2\pi)) + 2\log_q e$ and $c_2 = c_2(q) = q(-1+4\log_q e)$.

Note that the binary adder channel and binary B-channel coincides. Therefore we consider in this chaper also the following modification of the binary adder channel, which we call weighted binary adder channel, or wBAC for short. The output of the wBAC is a linear combination (*with unknown coeffcients*) of the code vectors that correspond to active users. For such model the following lower and upper bounds on the rate of corresponding signature codes was proved [22]

$$\frac{\log t}{t}(1+o(1)) \geq R \geq \frac{1+o(1)}{t}.$$

In the fourth chapter three different applications of signature codes are presented. The first application concerns the digital right management area and is called *multimedia digital fingerprinting codes* (MDF codes). Namely, it is shown that the problem of constructing the set of fingerprints that could protect the digital content from illegal redistribution under collusion attacks can be reformulated in two ways: quantized and continuous. For the quantized reformulation the construction of fingerprinting codes is equivalent to the problem of constructing signature codes for the A-channel. Such equivalence allows us to prove better results than all previously known results for the multimedia digital fingerprinting codes. Namely, previously known constructions do not give "good codes", since the rate of corresponding codes tends to zero with the growth of the code length. The relationship with separating codes allows to prove the existence of good codes

with non-vanishing rate [5]. Moreover, the proposed construction with efficient decoding also can be used for the MDF codes, and this construction also improves previously known results, see [5].

As for another reformulation, i.e., without quantization, it turned out that the problem is equivalent to the generalization of binary B-channel, namely, weighted binary adder channel for which the allowed weights of linear combinations are such that $\lambda_j \in (0, 1]$ and $\sum_{j \in J} \lambda_j = 1$ where J is a set of active users, or equivalently, users from a coalition. It was proved [22] that the rate of corresponding codes is of order at least $\Theta(t^{-1})$ which significantly improves previously known results as well as the new results for quantized version.

Another application that is considered in the fourth chapter is also related to digital right management problem, namely, to the broadcast encryption scheme. The traceability property for the most general broadcasting scheme is considered, i.e., the scheme which allows to identify at least one user from the malicious coalition by minimal distant decoding. The analogue of Gilbert-Varshamov bound is proved for the corresponding codes, and some numerical results are provided.

The last application that is considered in the fourth chapter is the symmetric group testing (SGT). The ordinary formulation of group testing is well studied problem whereas the symmetric group testing model is rather unexplored. SGT differs from ordinary group testing in the question-answer model. In SGT the response on a test equals 0 iff no defective elements belong to the tested subset, equals 1 iff all elements of tested subset are defective, and equals $\{0, 1\}$ otherwise, i.e., the output is in the ternary alphabet. It is easy to check that the problem of non-adaptive search of t defective elements in a symmetric model is equivalent to the problem of constructing t-signature codes for A-channel. So, the results about upper and lower bounds for noisly and noiseless cases as well as construction can be also applied for t-SGT codes. This equivalence, as well as the received results are new for the SGT codes.

Approbation of the thesis

The results of the thesis were reported by the author at the following research seminars:

1. Presentations "Digital fingerprinting codes 1-2" (29 February and 11 April, 2016) at Kolmogorov's Seminar on Descriptive and Computational Complexity of the Department of Mathematical Logic and Theory of Algorithms at Faculty of Mechanics and Mathematics of Moscow State University.

2. Presentation "Compositional multiple access channel with partial activity" at All-Moscow seminar on information and coding theory, IITP RAS, 11th October 2016.

3. Presentation "On one generalization of cover-free codes" at "All Moscow" seminar on coding theory, IITP RAS, 19th March 2019.

4. Presentation "On one generalization of cover-free codes" at joint research seminar on coding theory of Skoltech and MIEM HSE, 27th March 2019.

The results of the thesis were reported at the following conferences:

1. Problems in Theoretical Computer Science, December 18-20, 2015, Higher School of Economics, Moscow, Russia, session talk "Digital fingerprinting codes for multimedia";

2. 2016 IEEE International Symposium on Information Theory, July 10-15, 2016. Barcelona, Spain, session talk "Signature codes for the A-channel and collusion-secure multimedia fingerprinting codes";

3. Nexus of Information and Computation Theories, Secrecy and Privacy Theme, Marseille, France, poster "On multimedia digital fingerprinting codes ";

4. The Alan Turing Contest in Theoretical Computer Science and Discrete Mathematics for Russian-speaking students (TuCo 2016), 1st of July, 2016, St. Petersburg Academic University, St. Petersburg, Russia, session talk "Signature codes for a special form of multiple access channel", Third Prize;

5. XV International Symposium Problems of Redundancy in Information and Control Systems, 26-29 September 2016, Saint-Petersburg, Russia, session talk "Signature codes for a special class of multiple access channel";

6. IEEE International Workshop on Information Forensics and Security (WIFS), 4-7 December 2016, Abu-Dhabi, UAE, session talk "Multimedia fingerprinting codes resistant against colluders and noise";

7. Arithmétique, Géométrie, Cryptographie et Théorie des Codes, 19-23 June 2017, CIRM - Luminy, Marseille, France, session talk "Zero-error coding for multiple-access channels as a new test bed for AG-codes";

8. The Tenth International Workshop on Coding and Cryptography 2017, September 18-22, 2017, Saint-Petersburg, Russia, session talk "Multimedia fingerprinting with noise via signature codes for weighted noisy adder channels and compressed sensing";

9. Munich International Workshop on Coding and Cryptography (MWCC) 2018, April 10-11, 2018, talk "On group testing with noise";

10. The 2nd RUSSIAN-HUNGARIAN COMBINATORIAL WORKSHOP, June 27-29, 2018, Budapest. Session talk "Symmetric group testing with noise";

11. IEEE International Symposium on Information Theory 2019, 7-12 July, Paris, France, session talk "A Construction of Traceability Set Systems with Polynomial Tracing Algorithm".

12. Munich International Workshop MIT-TUM-DLR 24-25 Feb 2020. Session talk "Signature codes for different multiple-access channels - what do we know and what we don't? "

Acknowledgments

Firstly, I would love to thank my family, and, especially, my mom who 16 years ago has chosen one of the best mathematical schools in Moscow for me.
Secondly, I would like to express my sincere gratitude to my adviser Prof. Grigory Kabatiansky for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Chapter 1

Coding for multiple access channels, non-adaptive group testing and digital fingerprinting

This chapter provides state of the art of signature codes for different models of multiple access channels, introduces a notion of partial order on MACs, and establishes relationships among three main objects of this thesis, namely, signature codes for MACs, digital fingerprinting codes and non-adaptive group testing.

1.1 Zoo of MACs and error-free codes for them

We consider *deterministic memoryless multiple-access channel* (MAC), i.e., for the given input symbols $x_1, \ldots, x_M$ the output of the channel equals to $y = s(x_1, \ldots, x_M) \in Y-$ the output alphabet.
The following transfer functions s are the most studied. In the parentheses we give the labels of the channels which will be used for short references.

1. *Binary adder channel* (label *BAC*): $X = \{0, 1\}$, $Y = \{0, 1, 2, \ldots\} = \mathbb{N}_0$ and
$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} x_i, \tag{1.1.1}$$
where $\sum$ is the ordinary sum of integers.

2. *OR-channel* (label *OR*): $X = Y = \{0, 1\}$ and
$$s(x_1, \ldots, x_M) = \bigvee_{i=1}^{M} x_i. \tag{1.1.2}$$

3. *A-channel* (label A): $X = \{0, 1, \ldots, q-1\}$, $Y = \{0,1\}^q$ and $s(x_1, \ldots, x_M) = (y_0, ..., y_{q-1})$, where $y_i = 0$ if $|\{j : x_j = i\}| = 0$ and $y_i = 1$ - otherwise

4. *B-channel* (label B): $X = \{0, 1, \ldots, q-1\}$, $Y \subset \mathbb{Z}^q$ and $s(x_1, \ldots, x_M) = (y_0, ..., y_{q-1})$, where $y_i = |\{j : x_j = i\}|$

5. *Adder by mod 2 channel* (label $\Sigma\oplus$): $X = \{0,1\}$, $Y = \{0,1\}$ and

$$s(x_1, \ldots, x_M) = x_1 \oplus x_2 \oplus \ldots x_M, \tag{1.1.3}$$

where $\oplus$ denotes the sum by mod 2.

6. *Non-binary adder channel* : $X = \{0, 1, \ldots, q-1\}$, $Y = \mathbb{N}_0$ and

$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} x_i, \tag{1.1.4}$$

where sum is the ordinary sum of integers.

7. *Weighted binary adder channel* (label wBAC): $X = \{0,1\}$, $Y = \mathbb{N}_0$ and

$$s(x_1, \ldots, x_M) = \sum_{i=1}^{M} \alpha_i x_i, \tag{1.1.5}$$

where α_i are some real numbers (called "weights", despite that α_i can be negative) and the sum is the sum of real numbers.

We will be interested in error-free coding for MAC when every user wants only to inform the receiver (the output of the MAC) if he/she is active or not. If i-th user is active than he/she sends vector c_i, and sends nothing — otherwise.

Definition 1.1.1. The set $C = \{c_1, \ldots, c_M\}$ of q-ary vectors is called a (t, M)-*signature code* or a t-*signature code* if for any two different subsets $U, V \subset C$ both of the cardinality at most t the corresponding outputs of MAC are different too.

Let as usual define the code rate $R(C) = n^{-1} \log_q M$ and let $R^\alpha(t, n)$ denote the maximal possible rate of t-signature code of length n for a given MAC, and α is its label. Let informally define

$$R_t^\alpha = \lim_{n \to \infty} R^\alpha(t, n)$$

Below we give an overview of known results on R_t^α for aforementioned MACs.

1.1.1 Partial order on MACs

Let $\mathcal{A}$ and $\mathcal{B}$ be two MACs with the same input alphabet X, output alphabets Y_A and Y_B and the corresponding transfer functions S_A and S_B. We introduce the following partial order on the set of all MACs by saying that MAC $\mathcal{A}$ is "smaller" than MAC $\mathcal{B}$ and denoted as $\mathcal{A} \preceq \mathcal{B}$ if there is a function $f : Y_B \to Y_A$ such that $f(S_B(U)) = S_A(U)$ for any t-subset $U \subset X$. Saying informally, it means that the output of the channel $\mathcal{A}$ provides less information than the output of the the channel $\mathcal{B}$.

For example, the following orderings hold:

- OR-channel $\preceq$ A-channel;
- Adder by mod 2 channel $\preceq$ BAC;
- B-channel is the maximal element of the MACs poset.

Proposition 1.1.1. *If $\mathcal{A} \preceq \mathcal{B}$ and a code $C \subset X^n$ is a t-signature code for MAC $\mathcal{A}$ then it is a t-signature code for MAC $\mathcal{B}$.*

Proof. Indeed, let C be a t-signature code for MAC $\mathcal{A}$ but not for MAC $\mathcal{B}$. The last property means that there are two different t-subsets $U, V \subset C, |U|, |V| \leq t$ such that $S_{\mathcal{B}}(U) = S_{\mathcal{B}}(V)$. But then $f(S_{\mathcal{B}}(U)) = f(S_{\mathcal{B}}(V))$, i.e., $S_{\mathcal{A}}(U) = S_{\mathcal{A}}(V)$ what contradicts to the property C is a t-signature code for MAC $\mathcal{A}$. □

The established relationship helps to establish the bound on the rate of signature codes using different partially ordered channels, i.e. if $\mathcal{A} \preceq \mathcal{B}$ then $R_t^{\mathcal{A}} \leq R_t^{\mathcal{B}}$.

1.1.2 Codes for the binary adder channel

Let us start from the remark that error-free codes for binary adder channel were first discovered in group testing under the name of non-adaptive search of counterfeit coins on a spring scale. Let us recall this problem.

There is a set of M coins and it is known that some of them are counterfeit, and the weights of genuine and counterfeit coins are known. And there is an exact spring scale which gives us the exact weight of a chosen subset of coins, hence allows to find out how many false coins are in a given weighted (tested) subset. The problem is to propose a non-adaptive strategy with the minimal possible number of weightings that allows to find all counterfeit coins. Let us enumerate coins as $\{1, \ldots, M\}$, and let $\mathbf{x} = (x_1, \ldots, x_M)$ be a binary vector, where $x_i = 1$ if the i-th coin is counterfeit, and $x_i = 0$ if the i-th coin is genuine one. Let $\mathcal{T}_j \subset \{1, \ldots, M\}$ be the set of coins which are measured by j-th weighting. Denote by $H_{\mathcal{C}}$ the $n \times M$ binary matrix which rows are characteristic vectors of sets

$\mathcal{T}_1, \ldots, \mathcal{T}_n$. This matrix is called *search matrix*. Then the corresponding non-adaptive strategy of n weightings can detect all counterfeit coins iff for any two different binary vectors $\mathbf{x}, \mathbf{y} \in \{0,1\}^M$ the following holds:

$$H_{\mathcal{C}}\mathbf{x}^T \neq H_{\mathcal{C}}\mathbf{y}^T. \tag{1.1.6}$$

Now consider a binary M-signature code $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\} \subset \{0,1\}^n$ for BAC. It means that sums

$$S(J) = \sum_{j \in J} \mathbf{c}_j \tag{1.1.7}$$

are different for distinct subsets J. Consider $n \times M$ matrix H_C which columns are vectors $\mathbf{c}_1, \ldots, \mathbf{c}_M$. It is easy to see from (1.1.6) that the property (1.1.7) is equivalent that the matrix H_C is a search matrix for M coins and hence non-adaptive search of counterfeit coins is equivalent to constructing (M, M)-signature code for BAC.

Denote by n_M the minimal possible dimension of an (M, M)-signature code for BAC, or, the same, the minimal possible number of weightings among all non-adaptive strategies which can detect all counterfeit coins among M coins. Then

$$n_M = \frac{2M}{\log_2 M}(1 + o(1)). \tag{1.1.8}$$

for $M \to \infty$. Formula (1.1.8) as an upper bound was proved by P. Erdös and A. Reniy [12] and as a lower bound – in [23], [24] by proposing the corresponding construction of matrix $H_{\mathcal{C}}$.

Let us relax the propery (1.1.7), namely, by demanding that sums $S(J)$ are different only for subsets J of the cardinality at most t. On the language of group testing it means that we known in advance that the number of counterfeit coins is at most t. On the language of t-signature codes it means that the corresponding code $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$, by the definition, is a t-signature code for BAC. Such codes are strongly related to generalized Sidon sequences or B_t-sequences, see [25]. Recall that a Sidon sequence (or Sidon set) is a sequence $(a_1, a_2, \ldots)$ of natural numbers such that all pairwise sums $a_i + a_j$, where $i \leq j$, are different and the question is what is the largest possible size of such sequence if all its members are at most N. This problem firstly occur in the theory of Fourier series and later became a cross point of additive number theory and combinatorics, starting from the paper of [26], see also [18, 25, 27]. This notion was generalized to the case of larger sums (up to t elements) and arbitrary Abelian groups. For the Abelian group of residues by modulo $\frac{q^{t+1}-1}{q-1}$, where q is a prime power, there is the construction, given by Bose-Chowla theorem [28], of a set of $q+1$ elements such that

all its t sums are different. In particular, for $t = 2$ it gives perfect Singer sets [29].

Note that the problem of constructing t-signature code for BAC is equivalent to the problem of non-adaptive search for counterfeit coins on a spring scale when it is known a priory that the number of false coins is at most t. The following upper and lower bounds are known for t-signature codes for BAC when t is fixed but large [30]

$$\frac{\log t}{4t} + o(t^{-1}) \leq R_t^{BAC} \leq \frac{\log t}{2t} + o(t^{-1}). \tag{1.1.9}$$

It is easy to see that $R_t \leq t^{-1} \log t$. Indeed, all coordinates of MAC outputs are integers in the range $[0, t]$, hence there are $(t+1)^n$ different outputs. On the other hand, different t-subsets of the code should get different outputs, i.e., $\binom{M}{t} \leq (t+1)^n$, and hence asymptotically for $t \to \infty$ one has that $R_t \leq t^{-1} \log t$. Improvement of this simple bound follows from the "entropy method". It is worth to mention that for BAC we know that the rate of best t-signature codes has the order $t^{-1} \log t$. Explicit constructions, namely, columns of parity-check matrices of binary BCH or Goppa codes, gives $R_{t,constr}^{BAC} \geq t^{-1}$. This rate loss is rather small given the simple (with polynomial in length) decoding procedure of the corresponding codes. For other channel that we discuss below we shall "pay" more.

1.1.3 Signature codes for disjunctive channel or superimposed codes

Another well studied MAC with partial activity is OR-channel or the so-called disjunctive channel. Again, historically, firstly the corresponding problem appeared not in the context of MAC but in the context of group testing in 1943 [14]. Group testing problem consists in finding all unknown defective elements (samples) of a search space, using subsets of the search space as tests (queries). The answer for a test is "yes" if at least one defective element is in the tested subset, and "no" if there are no defective elements. It is easy to see that the group testing model is equivalent to the logical OR function, i.e., disjunction, and a non-adaptive search of defective elements, which number is at most t, among M elements is the same as a t-signature code for OR channel. These codes appeared first time in Kautz and Singleton paper [2]. They also introduces codes, called t-superimposed codes, with somewhat stronger property, namely, that for any code subset $U \subset C$, $|U| \leq t$ and any codevector $c \notin U$

$$\bigvee_{u \in U} \mathbf{u} \neq \bigvee_{u \in U} \mathbf{u} \vee \mathbf{c}. \tag{1.1.10}$$

Note that this property significantly simplify "decoding" algorithm which recovers the input codewords, since it allows the decoding by checking M codewords

instead of M^t.
If replace vectors $\mathbf{c}_1, \ldots, \mathbf{c}_M$ of t-signature code C of length n on the corresponding sets $X_1, \ldots, X_M \subset \{1, \ldots, n\}$ then the property that logical sums of t or less vectors c_i are distinct means that unions of t or less sets X_i are distinct, and the property (2.2.1) means that no set of $X_1, \ldots, X_M$ is covered by the union of t others – in context of the set systems this notion was introduced by Erdos, Frankl and Furedi [31], see also [32,33]. Therefore these codes also called *cover-free* codes.

It is well-known (and easy to check) that a t-signature code for OR-channel is $t-1$ cover-free code and t-cover-free code is a t-signature code for OR-channel. The following upper and lower bounds are known for large t, see [31], [32]

$$\Theta\left(\frac{1}{t^2}\right) \leq R_t^{OR} \leq O\left(\frac{\log t}{t^2}\right). \tag{1.1.11}$$

1.1.4 Signature codes for A-channel

There is a well-known in coding theory class of codes which is called *separating* codes and these codes play the same role for A-channel as cover-free codes for OR-channel. Separating codes has a long story started in 60s of the last century. A code C is called (s,t)-separating code if for any two code subsets $U, V \subset C$ and $U \cap V = \emptyset$, where $|U| \leq s$, $|V| \leq t$, there is a coordinate i s.t. $U_i \cap V_i = \emptyset$. It is easy to check that a signature code for A-channel is $(1, t-1)$-separating code and $(1,t)$-separating code is a signature code for A-channel. Since $MAC\ OR \preceq MAC\ A$, one has that for the binary A-channel $R_t^{OR} \leq R_t^A$. On the other hand, it is easy to see (and well known) that $R_t^A \leq 2R^{OR}$. Hence, it follows from (1.1.11) that for large t:

$$\Theta\left(\frac{1}{t^2}\right) \leq R_t^{A} \leq O\left(\frac{\log t}{t^2}\right). \tag{1.1.12}$$

1.1.5 Signature codes for B-channel

We mentioned already that B-channel provides the maximal information among the considered MACs. Therefore it looks like constructing codes for this channel should be an easy task. In contrary, not much was known about codes for B-channel in general case, only upper and lower bounds for the binary case when B-channel is the same as binary adder channel and these bounds are given by (1.1.9). One of the main goal of this Thesis is to generalize these bounds to nonbinary case.

1.1.6 Signature codes for adder by mod 2 channel

It is the most simple to analyze example of seven MACs. Indeed, a code $C = \{c_1, \dots, c_M\}$ is a t-signature code for the adder by mod 2 channel, the property that different sums of t or less vectors are different is equivalent that any $2t$ codevectors are linear independent over $GF(2)$. Then it is well known object, namely, these vectors are columns of a parity-check matrix of a binary linear code correcting t errors. Hence, at least for the case when t is fixed we know, thanks to Hamming bound and BCH-codes, that

$$R_t^{\Sigma\oplus} = \frac{1}{t} + o(1). \tag{1.1.13}$$

Note, that it is the only case of MAC when the asymptotical rate of the best signature codes is known. Moreover, even the case of adversarial noise can be solved, see [34].

1.1.7 Signature codes for weighted adder channel

It's not difficult to check that a *binary* code $C = \{c_1, \dots, c_M\}$ is a t-signature code for this channel iff any $2t$ codevectors are linear independent over the field $\mathbb{R}$ of real numbers. For binary vectors their linear independence over $GF(2)$ implies that they are linear independent over $\mathbb{R}$. Hence binary Goppa-BCH codes show that

$$R_t^{wBAC} \geq t^{-1}. \tag{1.1.14}$$

On the other hand, BAC is a particular case of wBAC, hence, the following upper bound is true

$$R_t^{wBAC} \leq O(t^{-1} \log t). \tag{1.1.15}$$

Let us also note that signature codes for wBAC is almost the same as non-adaptive search of counterfeit coins if they may have different weights and weights are unknown, see [35]

1.2 Combinatorial group testing

Group testing is a combinatorial scheme developed for the purpose of efficient identification of defective elements in a given pool of subjects. The naive solution of the search of defective elements is to test each item separately, but group testing allows to conduct tests in more efficient way. The main idea is to test the samples in groups (subsets), rather than individually, which decreases the number

of tests conducted.

The history of this problem starts with the work of Dorfman [14], where he formulated the problem in the context of the blood tests for the presence of the particular disease. In this case, blood samples of different persons were mixed and then tested. If at least one of the blood samples used in this test was "defective" then the answer was "yes". If all blood samples were "good" then the answer was "no".

This version of the problem was extensively studied in the literature and found many applications in computational molecular biology, chemical analysis and strong connections with algorithms, complexity theory, data compression, computational geometry and so on. For more detailed review of group testing applications see [36].

Mentioned applications also gave rise to many other versions of group testing problem. There are three main points of difference of such schemes. The first one is the strategy of the search. There are two possible cases: the so called adaptive and non-adaptive search. For the adaptive search questions/test are made in series in dependence of the answers for previous questions. For non-adaptive case all tests are conducted simultaneously, and based on all answer one decides about the set of defective elements. The second difference for group testing models is the answer-question model. For example, one can think of tests where each sample can participate only in a finite number of tests, and the number of samples in one test is also bounded; or one can think of threshold schemes, where one receives the answers "yes" only if the number of defective elements is bigger than some predefined amount [37], etc. The third difference is the presence of noise. Noiseless and noisy cases are considered in the literature.

There are two main models of noise for group testing. The first one is probabilistic, where the error is designed using some probability distribution, see e.g. [38]. The second one is combinatorial model, also known as adversarial noise, this is exactly the type of errors that we consider in this thesis. Probably the most famous problem of group testing with combinatorisl noise is the so-called Ulam's problem on searching with a lie. Ulam asked in his book [39] what is the minimal number of yes-no queries needed to find an unknown integer between 1 and $N = 10^6$ if one lie is allowed among answers (lie is equivalent to an error). In fact, this problem was first stated by A.Renyi in [40], so it is more correctly to call Renyi-Ulam problem (or game).

The exact answer for adaptive search algorithms and arbitrary M was given by A.Pelc in [41], see also his review paper [42]. The corresponding asymptotic result is known for general case of L false answers, namely, for fixed L and growing M the minimal number of queries behaves asymptotically as $\log_2 N + L \log_2 \log_2 N$ and it can be achieved by non-adaptive search.

In this Thesis we consider the modification of the ordinary group testing problem, namely, symmetric group testing (SGT). The use of SGT was originally motivated by applications in circuit testing and chemical component analysis [43]. As an example, consider the testing of N identically designed circuits using only serial and parallel component concatenation. In the serial testing mode, one can detect if all circuits are operational. In the parallel mode, one can detect if all circuits are non-operational. If at least one circuit is operational and one is non-operational, neither of the two concatenation schemes will be operational. Detecting efficiently which of the circuits are non-operational is exactly what symmetric group testing is aimed to.

More formally, consider the set $X = [N]$ of all samples and let $\mathcal{F} \subset X$ be a tested subset. In SGT the response on a test $\mathcal{F}$ equals 0 iff no defective elements belong to $\mathcal{F}$, equals 1 iff all elements of $\mathcal{F}$ are defective, and equals $\{0, 1\}$ otherwise. The goal is to create such family $\{\mathcal{F}_1, ..., \mathcal{F}_n\}$ of subsets (tests) of X of minimal size n that the answers for such tests allow to uniquely identify the subset of defective elements, given that their number is upper bounded by some fixed parameter d. It is convenient to consider binary characteristic vectors of the sets, i.e., we map each test $\mathcal{F} \subset X$ to the binary vector $\mathbf{f} \in \{0, 1\}^N$, where $f_i = 1$ if $i \in \mathcal{F}$ and $f_i = 0$ otherwise. Since we consider the non-adaptive version of symmetric group testing we can represent the family of tests in a form of $n \times N$ matrix H where rows represent tests and, consequently, columns $\{h^1, ..., h^N\} \subset \{0, 1\}^n$ represent "identifying" vectors for each sample from a pool. If an element $h^i_j = 1$, $i \in [N], j \in [n]$, it means that i-th sample from the pool participates in j-th test. The answer for such set of tests can be represented as a vector $a \in \{\{0\}, \{1\}, \{0, 1\}\}^n$, where a_j is the answer for the j-th test. Then, the goal of SGT is to construct such matrix that gives for different subsets of defective elements the different answer vectors $\mathbf{a}$.

1.3 Digital fingerprinting codes

With rapid development of multimedia technologies and the steady growth in the use of the Internet, a digital marketplace where a wide range of multimedia content (such as image, video, audio, speech...) is available, has become increasingly popular. However, the ease with which digital content can be accessed, retrieved and manipulated, poses the challenging task of devising methods for copyright protection and prevention of redistribution. One of the prominent techniques that is used to achieve that goal is called digital fingerprinting. The main idea consists in embedding in each copy of digital content a personalized and *undetectable* mark, called *digital watermark*. Note that the first examples of usage of the same method can be traced back to the XVIIth century. The creators-owners

of logarithm tables used to introduce tiny errors in the insignificant digits of $\log x$ for few specific values of x. Had a user of a logarithmic table sold illegal copies of it, the errors in the table would have allowed to identify who was the owner. This is an example of a standard digital watermarking technique which is widely used nowadays. Embedding of different marks into copies of different users allows to recognize the cases when a single dishonest user produces an illegal copy of distributed content. There are also the cases when the users possessing different copies of the same content collude (forming a coalition) and produce a forged version of the content based on their copies (or a single user bought few copies). That type of attacks, known as collusion attacks, is of interest and requires the development of more sophisticated techniques, based on digital watermarking, that allow to maintain the security at due level. In this thesis we omit the mark embedding procedure and concentrate on the creation of the set of fingerprints resistant to collusion attacks.

There are many different ways of modeling the broadcasting. In this Thesis we consider two models: the continuous one, usually called *multimedia model*, in which a digital content x is represented as a vector over the field of real numbers (see [**?**]), and another model where a digital content x is a vector over some finite alphabet. The later model heavily exploits secret sharing schemes.

In the multimedia model the distributor, in order to create the set of fingerprints, chooses n orthonormal noise-like signals (vectors), which length (energy) is much smaller then the length of the host signal, and then set each fingerprint as a linear combination of these vectors. The coefficients of linear combination are from $\{\pm 1\}$ or $\{0, 1\}$ and are different for different users. Thus, the vector of coefficients uniquely identify a particular user. As for embedding of the marks, the additive embedding procedure is used, i.e., the marked content is just a sum of original content vector and fingerprint vector. To create a forged copy the colluders calculate the linear combination of their copies where the sum of coefficients (of weights) equals one, which is needed to maintain the proper level of energy of the signal. The fact that users are restricted to the linear attacks and cannot manipulate the individual (basis) signal constitutes the so called *marking assumption* for multimedia version. The main problem is the same: to create a set of vectors (of coefficients) of maximal possible cardinality in such a way that the distributor can reveal at least one participant of coalition of size $\leq t$. In overwhelming majority of cases the authors considered only the case of averaging type of the collusion attack, namely, the colluders have the same impact into the resulting forged vector, i.e., all the coefficients of linear combination are equal. But from the cryptography point of view the averaging attack is much weaker than the general linear attack since in this case the strategy of colluders is known to the distributor. In the Thesis we consider the case of arbitrary weights and we

derive more efficient results (codes) for the case of stronger attack. We establish teh relationship between this problem and signature codes for the corresponding multiple access channels and propose also a family of codes with an efficient algorithm for tracing colluders.

Another model, based on secret sharing schemes, was proposed in the beginning of nineties by Chor, Fiat and Naor in [44] where they invented a combinatorial scheme for broadcast encryption known today as *traitor tracing* scheme. Consider a distributor who has some digital content to broadcast and who wants to sell the access to this content only to authorized users, i.e., users who paid for the access. To prevent illegal redistribution of the data, the distributor encrypts the data blocks with session keys and gives to each authorized user the corresponding set of keys which we will also call decoders. These decoders are needed to decipher session key and then the original content. The main challenge of broadcast encryption schemes is to make them collusion resistant. Indeed, malicious users, in order to create a pirate version of the decoder and not to reveal completely their identities, can form a coalition and create a pirate version as a mixture of their decoders. In other words, the pirate version represents users from the coalition but only partially, so for the distributor it becomes harder to identify the participants of the coalition. Assuming that the cardinality of a possible coalition is not greater than some integer t, the desired property is that once a forged decoder is found, the distributor can trace it back to at least one traitor from the corresponding malicious coalition.

There are three particular cases of tracing traitors schemes known as *codes with the identifiable parent property (IPP codes)* [45], *(binary) set systems with the identifiable parent property (IPP set systems)* [46], [47] and non-binary IPP set systems [48]. The IPP codes were extensively studied in the literature, see e.g. [49], [50], [51], also a detailed overview can be found in [52], [53]. These codes (schemes) are based on a perfect (n, n)-threshold secret sharing scheme (see [54], [55]). As for the binary IPP set systems, it started with the papers [46], [47] and the most recent results can be found in [56], [57] and [58]. These schemes are based on a perfect (w, n)-threshold secret sharing scheme. A new, the most general type of IPP systems, called non-binary IPP set systems, was introduced in [48], and further developed in [59], [60]. Non-binary IPP set systems have IPP codes and binary set systems as its partial cases. Such scheme constitute the subject matter of the fourth chapter devoted to the applications of signature codes. In particular, we shall shown how codes for such scheme are connected with the so-called "malicious MAC", when the output of a MAC is controlled by our opponent who can chooses the input symbol among accessible ones. Namely, we consider A-channel with additional property that the receiver (decoder) sees not all used elements of the input alphabet X, but only one of them, and this

element is chosen by dealer's opponent.

Chapter 2

Signature codes for A-channel

2.1 Coding for multiple access A-channels

In [15] authors introduced a MAC model called q-frequency M-user multiple access channel without intensity information, or, simply, A-channel. Formally, this channel model can be described as follows. Consider a multiple access channel with M users, where every user can occupy one of q frequencies $f_1, f_1, ..., f_q$ to transmit on at each time slot during a session. Each session consists of n time slots. For the A-channel, the output at each time slot is a symbol that represents the subset of frequencies occurred as inputs to the channel at that time slot. Now the name of the channel becomes clear — for A-channel only frequencies are known, but how many users used a particular frequency at a particular time slot is unknown, thus without intensity information. The channel for which this information is known is called B-channel and will be studied in the next chapter. The output of the entire session for A-channel consists of n such symbols, one symbol for one time slot. For simplicity, the input alphabet will be represented as $\mathcal{A} = \{1, 2, ..., q\}$ and output alphabet as a set of all binary vectors of length q. Indeed, for the A-channel the output symbol can be represented as a binary vector of length q for which the j-th coordinate ($j \in [q] := \{1, \ldots, q\}$) equals 1 if and only if frequency f_j was used by at least one user at considered time slot. In the paper [15] authors considered the case of information transmission, i.e., when each user has a code $C_i, i \in [M]$ that consists of signals corresponding to the messages. Signals are represented as q-ary vectors, each coordinate defines the frequency that is used for information transmission at the corresponding time slot. At each session user can transmit one of its signals (messages). The problem is to construct such set of codes for users that given the output of the A-channel it is possible to identify which message was sent by each user. The parameter that is studied for such systems is called the sum rate and is defined as

$$R^A_{\text{sum}} = R^A_1 + ... + R^A_M,$$

where $R^A_i = n^{-1} \log |C|_i$. In [15] both information-theoretic bounds on the achievable sum rate and constructive coding schemes were investigated.

In that chapter, we will say *signature code* to refer to signature codes for A-channel, since the whole chapter is devoted to the A-channel.

Formally, the output of the A-channel can be defined as follows. Consider a set $C \subset \{0, 1, 2, ..., q-1\}^n$ consisting of some q-ary vectors of length n, each such vector represents one user. According to the described model of A-channel, the output can be formally described as the following transfer function S_A that maps a subset of $C = \{c_1, ..., c_M\}$ to the binary matrix of size $n \times q$:

$$S_A : C \to \{0, 1\}^{n \times q} \tag{2.1.1}$$

according to the following rule : let $U \subset C$ then $S_A(U)$ is a binary matrix of size $n \times q$ such that the element

$$S_A(U)_{ij} = 1 \text{ iff } \exists u \in U \ : \ u_i = j$$

for $i \in [n]$, $j \in [q]$. In other words, the output matrix describes the presence of each frequency (element of q-ary alphabet) at each time slot. So, following the general definition of t-signature code for multiple access channel (definition 1.1.1) *a q-ary code C is called a t-signature code for adder channel if for any two different subsets $U, V \subset C$ such that $|U|, |V| \leq t$* it holds that

$$S_A(U) \neq S_A(V). \tag{2.1.2}$$

The binary case will be of a main interest in this chapter and it is convenient to consider the output $S_A(U)$ as a vector of length n over a ternary alphabet $\{0, 1, \{0, 1\}\}$. Note, that $S_A(U)_i = 1$ iff all vectors $u \in U$ have 1 at i-th position, $S_A(U)_i = 0$ iff all vectors $u \in U$ have 0 at i-th position, and $S_A(U)_i = \{0, 1\}$ iff there were both vectors with 0 and 1 at i-th position.

The main parameter that is studied in this chapter is the rate of the best t-signature code which is defined as

$$R_t^A(n) := \max n^{-1} \log_2 |C|, \tag{2.1.3}$$

where the maximum is taken over all t-signature codes C of length n. In the rest of this chapter we will consider the estimation of the rate of binary signature codes for noisy and noiseless cases and especially its asymptotic behavior as $n \to \infty$.

2.2 Upper and lower bounds via separating codes

Recovering the whole set of active users. We argue that the task of identifying all active users of a session can be solved by means of superimposed (cover-free)

codes, introduced in 1964 [2], and later investigated in [31, 32, 61], and separating codes, see [17]. Saying in other words, these two types of codes are at the same time t-signature codes for A-channel. In what follows we will give the necessary definitions and establish the connection of these notions with the specificity of the A-channel.

To get closer to the notions of superimposed and cover-free codes we start with another channel model known as OR-channel or disjunctive channel. The idea of the OR-channel is the same as the A-channel with the only difference in the outputs. As for binary A-channel, each user has its own binary vector, that is transmitted to mark the desire to be active. For OR-channel model the output of the channel for one time slot is 1 if at least one user from the set of active users has one, and 0 otherwise. Formally, we have the following definition of signature codes that can be used to recover the set of active users for OR-channel.

Definition 2.2.1. [2] A binary code C is called a t-signature code for OR-channel if for any two different subsets $U, V \subset C$ such that $|U|, |V| \leq t$

$$\vee_{v \in V} \mathbf{v} \neq \vee_{u \in U} \mathbf{u},$$

where $\mathbf{a} \vee \mathbf{b}$ is the bitwise logical OR.

Signature codes for OR-channel appeared firstly in the literature under the names uniquely decipherable of order t or *t-superimposed codes* [2]. We will use all these names to refer to signature codes for OR-channel. Recall one more notion for OR-channel that will be useful also for A-channel.

Definition 2.2.2. A binary code $C \subset \{0,1\}^n$ is called a *t-cover-free code* if for any $U \subset C$, $|U| \leq t$ and any $\mathbf{z} \in C \setminus U$ the following holds:

$$(\vee_{u \in U} \mathbf{u}) \vee \mathbf{z} \neq \vee_{u \in U} \mathbf{u}. \tag{2.2.1}$$

In other words, a code C is t-cover-free if for any $U \subset C$, $|U| \leq t$ and any $\mathbf{z} \in C \setminus U$ there exists a coordinate $k = k(U, \mathbf{z})$ such that $u_k = \{0\}$ for all $u \in U$ and $z_k = 1$. The name "cover-free" codes can be explained if we consider subsets instead of their characteristic vectors. Then, the described above property means that the family of subsets is a t-cover-free family if no set is covered by the union of t others. The notion was coined in the paper with the same name by Erdos et al. [31,61], but firstly such codes appeared in [2] under the name of zero-false-drop of order t. More over, it is well known (and easy to check) that a t-cover-free code is also a t-signature code for the OR-channel. The advantage of cover-free codes over the signature codes in general is that cover-free codes allow more simple and faster identification of the set of active users, since instead of considering all possible subsets of t or less users, one can just consider code vectors that are "covered"

by the output vector. That means that the complexity of decoding is decreased from $O(M^t)$ to $O(M)$.

Since a t-cover-free code is also a t-signature code for the OR-channel and OR-channel $\preceq$ A-channel, then Proposition 1.1.1. gives the following useful fact that any t-cover-free code is also a t-signature code for A-channel.

As it was stated above, if C is a cover-free code then one can just check if the output "covers" the user's vector or not. Formally, the set of active users can be found as

$$\widehat{U} = \{z \in C \ : \ z_i \in U_i, \ i = 1, \ldots, n\},$$

where $S(U) = (U_1, \ldots, U_n)$ is an output of the A-channel.

It is known [62], [31, 61] that the rate R_{t-cf} of the best t-cover-free code is at least $\Theta(t^{-2})$. Therefore we immediately conclude that for any fixed t there exist t-signature codes for A-channel with non-vanishing rate, i.e., *good* codes, which are capable of identifying the *entire set* of active users. But even though cover-free codes give an idea of how to construct good signature codes for the A-channel, there exists a more direct solution which employs the notion of separating systems.

Definition 2.2.3. [16] A q-ary code $C \subset \{1, 1, ..., .q\}^n$ is called (s, t)-separating code if for any two disjoint sets $V, U \subset C$ such that $|V| \leq s$, $|U| \leq t$ there exists at least one coordinate which separates them, i.e., there exists $k \in [n]$ s.t. $V_k \cap U_k = \emptyset$.

Note that $(1, t)$-separating and (t, t)-separating codes were rediscovered in [63] under the names of frameproof and secure frameproof codes respectively. For extensive describtion of results about separating codes see two excellent survey papers [17], [64]. And for recent results for the case $q \to \infty$ see [65].

Let us note that $(1, t)$-separating codes play the same role for the A-channel as superimposed codes play for the OR-channel, in particular, any $(1, t)$-separating code is also a t-signature code for A-channel. Namely, if C is $(1, t)$-separating code, then the set U of active users can be uniquely recovered as

$$\hat{U} := \{c \in C : c_i \in U_i, \ i = 1, \ldots, n\}, \tag{2.2.2}$$

where $S(U) = (U_1, \ldots, U_n)$ is the corresponding output of A-channel and $U_i := \{u_i | u \in U\}$. Indeed, $U \subseteq \hat{U}$ since $u_i \in U_i$ for any $u \in U$ and all i. Let $U \neq \hat{U}$ and $z \in \hat{U} \setminus U$, but then z cannot be separated from U what contradict to the $(1, t)$-separation property.

Denote by R_{t-sep} the largest possible rate of binary $(1,t)$-separating codes and by R_{t-cf} the largest possible rate of t- cover-free codes. The following relationship between rates of best separating and best cover-free codes is rather obvious and long known, see [2]):

$$R_{t-cf} \leq R_{t-sep} \leq 2R_{t-cf}. \tag{2.2.3}$$

The best known results for R_{t-cf} have the following form for large t, see [62]:

$$\frac{\ln 2}{t^2}(1+o(1)) \leq R_{t-cf} \leq \frac{2\log_2 t}{t^2}(1+o(1)). \tag{2.2.4}$$

Hence we now conclude that for large t:

$$\Theta(t^{-2}) \leq R_{t-sep} \leq O\left(\frac{\log_2 t}{t^2}\right). \tag{2.2.5}$$

Note that in binary separating codes we can use both types of $(1,t)$-separation, i.e., at a particular position all vectors from $U, |U| \leq t$ have zero and vector z has one or the inverse — all vectors from U have one and vector z has zero. That fact describes the advantage of separating systems over cover-free families for which only one type of separation is allowed.

At the same time, the $(1,t)$-separation property is stronger than the t-signature property for the A-channel. Indeed, signature codes demand only that for different subsets of users the corresponding outputs are also different. In particular, let us note that $(1,t)$-separating codes afford a relatively simple decoding procedure with complexity of order M instead of the brute-force complexity M^t. Therefore, it is in principle possible that t-signature codes have rate that asymptotically exceeds (2.2.5). The following statement due to [4] (Lemma 4.6) shows that it is not true.

Proposition 2.2.1. *Any t-signature code for A-channel is a $(1,t-1)$-separating code.*

Proof. Let C be a t-signature code for A-channel. This means that for any two different subsets of users $U \neq V$, $|U| \leq t$, $|V| \leq t$ their signature vectors $S(U)$ and $S(V)$ are different. Let V be any active subset of size $t-1$ and let x be any vector not from V. Set $U := V \cup x$. Then $S(U) = S(V \cup x) \neq S(V)$, which implies that there exists a coordinate k such that $x_k \notin V_k$, i.e., the k-th coordinate separates x from V. □

Denote by R_t^A the highest possible rate of binary t-signature codes for A-channel. Then finally we have

Theorem 2.2.1. *For large t*

$$\Theta(\frac{1}{t^2}) \leq R_t^A \leq O\left(\frac{\log t}{t^2}\right). \tag{2.2.6}$$

Proof. According to 2.2.1 and application of the proposition 1.1.1 to the fact that OR-channel $\preceq$ A-channel, we have

$$R_{t-sep} \leq R_t^A \leq R_{(t-1)-sep}.$$

Also, according to (2.2.3)

$$R_{t-cf} \leq R_{t-sep} \leq R_t^A \leq R_{(t-1)-sep} \leq 2R_{(t-1)-cf}.$$

So, following (2.2.5) we can conclude that

$$\Theta(t^{-2}) \leq R_t^A \leq \Theta\left(\frac{\log t}{t^2}\right).$$

□

For small t there is a substantial difference between the rate R_t^A and the rate R_{t-sep} of $(1,t)$-separating codes. Consider as example the binary case with $t = 2$. For $(1,t)$-separating codes it is known, see [17], [64] that

$$1 - \frac{\log_2 3}{2} = 0.207518 \leq R_{2-sep} \leq 0.5. \tag{2.2.7}$$

It is interesting to note that the lower bound from (2.2.7) which is due to Gilbert-Varshamov type bound random coding technique was beaten by the means of algebraic geometry (AG) codes in [66]. So, the best known lower bound for today is

$$R_{2-sep} \geq 0.207565.$$

On the other hand, it is known and easy to check that A-channel with two frequencies is the same as the binary adder channel. The adder channel also consists of users who have binary vectors, and the output is modeled as the sum (over real numbers) of the vectors of active users. So, equivalence of these channels can be seen if we replace outputs of the A channel $0, 1$ and $\{0,1\}$ on integers 0, 2 and 1 correspondingly, see [13], [15]. As for the rate R_2 of the best 2-signature codes for the adder channel the following bounds are known:

$$0.5 \leq R_2^{\Sigma} \leq 0.5753, \tag{2.2.8}$$

where the lower bound was obtained in [25,27], and the upper one in [67]. So, for small t signature codes for the adder channel provide better rate than separating codes.

Note that we consider the asymptotic behavior when fixed channel's input alphabet is fixed but the code length n tends to infinity. There are papers, see the corresponding overview [68], where another type of asymptotic behavior is considered, namely, when the code length n is fixed but the size of the alphabet q goes to infinity. Our consideration is arguably more in line with other basic results of information theory which typically assume a fixed channel alphabet and increasing code length.

Recovering at least one active user. Without restricting ourselves to the complete identification of the set of active users, we can achieve even better code rates. The following definition was coined in [69], where the idea of recovering one user instead of an entire set was inspired by the fingerprinting problem that we will discuss later.

Definition 2.2.1. A code $C \subset \{0,1\}^n$ is *$t-$single user tracing* (SUT) if from the bitwise OR of words from any subset $U \subset C$ such that $|U| \leq t$ we can find out at least one word from U.

In other words, for any family of subsets $C_1, ..., C_k \subset C$ such that $|C_j| \leq t$ the equalities

$$\bigvee_{c \in C_1} \{c\} = ... = \bigvee_{c \in C_k} \{c\} \text{ imply } \bigcap_{j=1..k} C_j \neq \emptyset.$$

Lemma 2.2.1. *If the code C is $t-$SUT then this code can be used to recover from the output $S(U)$ at least one active user participating in a particular time slot.*

Proof. This fact follows immediately from the fact that t-signature codes for OR-channel can be used for A-channel. □

According to [70], the rate $R_{t-SUT} = n^{-1} \log_2 |C|$ for $t-$SUT is the following $R \geq \frac{1}{20t} > 0$. Later, in [71] authors improved the results of [70] by considering the identification of bigger number of uses. Namely, the following definition was introduced.

Definition 2.2.2. A code $C \subset \{0,1\}^n$ is *k-out-of-t user tracing*$(k,t$-UT) if from the bitwise OR of any $l \leq t$ words from C one can identify at least $min(k;l)$ of these words.

Formally, for any family of subsets $C_1, ..., C_k \subset C$ such that $|C_j| \leq t$ the equations

$$\bigvee_{c \in C_1} \{c\} = ... = \bigvee_{c \in C_k} \{c\} \text{ imply } \bigcap_{j=1..k} C_j \geq k.$$

In [71] it was proved that if $k \leq \sqrt{t}$ then the rate $R_{k,t-UT} = n^{-1} \log_2 |C|$ of k-out-of-t user tracing code is of order

$$R_{k,t-UT} = \Theta(t^{-1}).$$

Which means that the order of the rate changes only by a constant factor while increasing the number of identifiable users. This result is more important for the multimedia digital fingerprinting codes (see section 4.2), rather than for multiple access channels.

2.3 Construction of signature codes with efficient decoding algorithm

As already remarked, $(1, t)$-separating codes afford a decoding complexity $O(nM)$ which is lower than the complexity $O(nM^t)$ of identifying active users in the general case of signature codes. At the same time, since the best t-signature codes for A-channel have non-vanishing rate, their cardinality M is an exponential function of the length n, and so the overall decoding complexity $O(n2^{Rn})$ is still too high for practical applications.

In this section we take up the problem of constructing asymptotically good t-signature codes for A-channel with decoding complexity *polynomial* in the code length. We show that this is possible, although the rate that these codes attain is smaller than the best known o rate $\Theta(t^{-2})$. More formally we have the following

Theorem 2.3.1. *There exist t-signature codes for A-channel with rate of order $\Theta(t^{-3})$ and decoding complexity polynomial in the code length.*

It the chapter 3, dedicated to applications, we will show how t-signature codes for A-channel are strongly related to digital fingerprinting codes, but for now, we will use the ideas presented in a series of works concerning digital fingerprinting. Namely, we rely on the ideas of [72] which was the first to construct digital fingerprinting codes with non-vanishing rate and polynomial-time decoding. One of the first papers where concatenation was applied for superimposed codes is [73] and the similar construction for separating codes was considered in [74].

In what follows, we will use the code construction known as concatenated codes [75]. To introduce reader to the notion we give a formal definition.

Definition 2.3.1. Let C be a $q-$ary code C of length n with the minimal code distance d and cardinality Q. And let V be a Q-ary code V of the cardinality M

and length N over an alphabet $\mathcal{A}, |\mathcal{A}| = Q$. Codes C and V are called inner and outer codes respectively. Let ρ be a bijection mapping

$$\rho : \mathcal{A} \to C,$$

which establish the 1-1 correspondence between codewords from C and elements from alphabet $\mathcal{A}$. The q-ary code $W = C \circ V$ of length nN and cardinality M with codewords constructed as follows

$$W = \{(\rho(v_1)||...||\rho(v_N)) : v = (v_1, ..., v_N) \in V\},$$

where $||$ denotes the concatenation of words, is called a *concatenated code* with inner code C and outer code V.

It is a basic fact about concatenated codes that the code distance of concatenated code satisfies $d(W) \geq dD$.

It is rather common and we will do the same, namely, to use Reed-Solomon code (RS code) as the outer code in our construction.

Definition 2.3.2. Let $\mathbb{F}_q = \{\alpha_0, ..., \alpha_{q-1}\}$ be a finite field,

$$RS = \{c = (f(\alpha_{i_1}), ..., f(\alpha_{i_n})) \mid f \in \mathbb{F}_q[x], \deg(f) < k\}.$$

It is a linear code of dimension k and code distance $d(RS) = n - k + 1$.

As in [72], we employ the idea of code concatenation with random inner codes and Reed-Solomon codes with large distance as an outer codes. Choose the binary inner code C_{inn} to be a $(1, t)$-separating code of length m and cardinality $q = 2^{\lfloor \mu m \rfloor}$, where we can take $\mu = -t^{-1}\log_2(1 - (et)^{-1})$, and hence $\mu \geq (et^2 \ln 2)^{-1}$. Such codes can be shown to exist by constructing random codes with independent coordinates and with probability of one equal to $p_1 = 1/t$. Let the outer code be a q-ary Reed-Solomon code W of length $N = q$ and rate $R_{out} = t^{-1}$ over the finite field $GF(q)$.

The codewords of the resulting concatenated code have the form

$$(\varphi(\alpha_1), \ldots, \varphi(\alpha_N)), \text{ where } (\alpha_1, \ldots, \alpha_N) \in W$$

and

$$\varphi : GF(q) \to C_{inn}$$

is a one-to-one map from the field $GF(q)$ onto the inner code C_{inn}. This construction affords the following decoding algorithm, similar to Algorithm $1'$ in [72]. Let

$$\mathbf{S} = (s_{11}, \ldots, s_{1m}, s_{21}, \ldots, s_{2m}, \ldots, s_{q1}, \ldots, s_{qm}) = (\mathbf{s}_1, \ldots, \mathbf{s}_q)$$

be an output of the A-channel. The decoding procedure is performed in two stages and heavily relies on the famous Guruswami-Sudan list decoding algorithm [76].

Algorithm 2.3.1. Decoding procedure.

1. In the first stage, we use a brute-force tracing algorithm for the inner code C_{inn} to decode vectors $\mathbf{s}_1, \ldots, \mathbf{s}_q$ as output vectors of A-channel.

   Let $U^{(1)}, \ldots, U^{(q)}$ be the corresponding subsets of C_{inn}, where $|U^{(i)}| \leq t$ and let $H_i := \varphi^{-1}(U^{(i)}) \subset GF(q)$.

2. In the second stage of the tracing algorithm, we use a soft-decoding version of the Guruswami-Sudan list decoding algorithm [76] which returns all codewords $\mathbf{w} \in W$ that satisfy the inequality

$$r(\mathbf{w}) := \sum_{i=1}^{N} r_{w_i,i} \geq \sqrt{NR_{out} \sum_{i,j} r_{ji}^2}, \tag{2.3.1}$$

   where r_{ji} are some non-negative weights assigned to the alphabet symbols. In our case we take $r_{ji} = 1$ if $j \in H_i$ and $r_{ji} = 0$ otherwise. It is easy to check that we need to find all codewords $\mathbf{w} \in W$ such that $r(\mathbf{w}) = N$.

Since the length of the inner code is logarithmic in the length of the entire concatenated codeword, the brute-force search has complexity polynomial in the code length $n = mq$. Also, note that with the chosen parameters, namely, $\sum_{i,j} r_{ji}^2 = \sum_i |H_i| \leq Nt$ and $R_{out} = 1/t$, the list decoder returns a polynomial-size list that may contain some excess codewords but they can be easily sorted out.

2.4 Noise-resistant signature codes for A-channel

In this section we consider the case of noisy outputs of the A-channel and adress the question of existance of good codes resistsnt to errors. As it was stated in the previous section, the signature codes for the A-channel are strongly connected with the cover-free codes for OR-channel. The cover-free codes that can deal with the presence of noise is already a well studied subject, the most valuable results can be found in [77], [78], [79]. As for the signature codes for A-channel, there are only few works that study the noisy case, see, for example, [80] where the probabilistic model of noise in the context of symmetric group testing was considered.

In this thesis we consider the model of adversarial errors. We assume that the output vector might be erroneous in no more than L positions, i.e., no more than L coordinates are incorrect. Nevertheless, the receiver should be able to reveal

all active users. Note, that errors might be of any type, i.e., even if the correct value is, for example, 1, the erroneous output might be both $\{0,1\}$ or 0, but the number of such errors is upper bounded by L. The goal is the same, i.e., for the given output vector to recover the set of active users even in the presence of noise. Formally it can be stated as follows:

Definition 2.4.1. A t-signature code C is said to correct up to L errors, or (t, L)-signature code for short, if for any $U, V \subset C$ such that $|U|, |V| \leq t$ and $U \neq V$ and any $S \in \{0, 1, \{0,1\}\}^n$ such that

$$d_H(S(U), S) \leq L \text{ and } d_H(S(V), S) \leq L$$

the equation $U = V$ holds.

Equivalently, a t-signature code C is said to correct up to L errors if the outputs of the channel differ in at least $D = 2L + 1$ positions. Formally, if for any $U, V \subset C$ such that $|U|, |V| \leq t$ and $U \neq V$ then $d_H(S(U), S(V)) \geq 2L + 1$, where d_H denotes the ordinary Hamming distance (here between two ternary vectors). In this section we are interested in the estimation of the rate $R_t^A(\delta)$ of (t, L)-signature codes with $\delta = \frac{2L+1}{n}$. In what follows we will prove the following lower and upper bound in the rate of (t, L)-signature codes. As for the lower bound, we show that for any $\delta < \delta_{crit} = t^{-1}(1 - t^{-1})^t$ the following holds

$$R_t^A(\delta) \geq \frac{2 \log_2 e}{t}(\delta_{crit} - \delta)^2. \tag{2.4.1}$$

Then, we prove the following upper bound was proved:

$$R_t^A(\delta) \leq \frac{1}{t-1} R(\delta),$$

where $R(\delta)$ denotes the asymptotic maximal possible rate of a code in the Hamming space with relative distance δ.

Let's firstly introduce the notion of *separating distance*. Consider the function $\Delta(a, B)$ for $a \in \{0,1\}, B \in \{0, 1, \{0,1\}\}$ defined as $\Delta(a, B) = 0$ if $a \in B$ and 1 otherwise.

Definition 2.4.2. Separating distance d_{sep} between a binary vector $c \in \{0,1\}^n$ and a ternary vector $S \in \{0, 1, \{0,1\}\}^n$ defines as

$$d_{sep}(c, S) := \sum_{i=1}^{n} \Delta(c_i, S_i) = n - |\{i \mid c_i \in S_i\}|.$$

Now we can move to the definition of t-cover-free signature code for A-channel with particular separating distance.

Definition 2.4.3. A binary code C of length n is called a *t-cover-free signature code with separating distance D* if for any subset $U \subset C$ such that $|U| \leq t$ and any code vector z out of U, i.e., $z \in C \setminus U$, there is at least D positions in which ternary vector $S(U)$ and binary vector z are separated, i.e., $d_{sep}(z, S(U)) \geq D$.

Informally, this definition means that the output of the channel differs from any vector out of the set of active users in at least D position. As it was done for the noiseless case, we can establish the connection between introduced notions.

Proposition 2.4.1. *If C is a t-cover-free signature code with separating distance D then C is also a (t, L)-signature code where $L = \lfloor \frac{D-1}{2} \rfloor$.*

Proof. Consider the inverse, then there exist two different subsets U, V s.t. $|U|, |V| \leq t$ and $d_H(S(U), S(V)) \leq D - 1$. The following relation is true for any $x \in U \setminus V$:

$$d_{sep}(x, S(V)) \leq d_H(S(U), S(V)) < D$$

which contradicts the cover-free property. □

As it was shown for the noiseless case, the cover-free property provides the efficient recovery of the set of active users. The same thing is true for the noisy case. Indeed, consider the output vector $\hat{S}$ which may differ from $S(U)$ in at most L positions. Then, the set of active users can be recovered as

$$\hat{B} := \{c \in C; d_{sep}(c, \hat{S}) \leq L\} \tag{2.4.2}$$

which is exactly the set of active users from U. Moreover, it is not difficult to check that a (t, L)-signature code is $(t-1)$-cover-free signature code with separating distance $D = 2L + 1$.

Proposition 2.4.2. *If C is a (t, L)-signature code where $L = \lfloor \frac{D-1}{2} \rfloor$ then C is also a $(t-1)$-cover-free signature code with separating distance D.*

Proof. Consider the inverse, i.e., there exist a vector $x \in C$ and a set $U \subset C, |U| \leq t$ such that $d_{sep}(x, U) < D$. Consider two subsets of active users: U and $V := U \cup \{x\}$, then

$$d_H(U, V) = d_{sep}(x, U) < D$$

which contradicts with the ability of correcting L errors. □

2.4.1 Lower bound on the rate of error-resistant signature code

The lower bound on the rate of t-signature code that is able to correct up to L errors can be established using the known results for the t-signature codes for OR-channel resistant to noise. As it was done for the noiseless case, we will use the notion of cover-free codes, or signature codes for OR-channel, but in the context where errors are present. The property of error-resistance for cover-free codes can be formulated in the following way [77].

Definition 2.4.1. Superimposed "distance" $d_{\sup}$ between two binary vectors $x, y \in \{0,1\}^n$ is defined as

$$d_{\sup}(x,y) = |\{i \ : \ y_i = 0, x_i = 1\}|.$$

Remark. Superimposed "distance" is not an ordinary distance because obviously it is not symmetric in general.

Definition 2.4.4. A binary code C is called a *t-cover-free code with superimposed distance D* if for any $U \subset C$, $|U| \leq t$ and any $\mathbf{c} \in C \setminus U$

$$d_{\sup}(c, \bigvee_{u \in U} u) \geq D$$

or, equivalently, there exists at least D coordinates in which $U_k = \{0\}$ and $c_k = 1$.

The additional condition about the superimposed distance D provides the error-tolerant property to the cover-free codes. Indeed, it can be easily checked that if a code has superimposed distance D, then $\lfloor \frac{D-1}{2} \rfloor$ errors can be corrected. Moreover, we have the following obvious relationship between t-cover-free code with superimposed distance D and t-cover-free signature codes with separating distance D.

Proposition 2.4.3. *If a code C is t-cover-free code with superimposed distance D then it is also a t-cover-free signature codes with separating distance D.*

So, this fact means that we can use the known lower bounds on the rate $R_{t,D-cf}$ of cover-free code with superimposed distance D to estimate the lower bound on the rate of t-cover-free signature codes with separating distance D. The best known result on the estimation of the asymptotic lower bound of $R_{t,D-cf}$ is due to Dyachkov, Rykov, and Rashad [77]. A bit weaker result but in a significantly simpler form was obtained in [20].

Theorem 2.4.1. *For any $0 < \varepsilon < \delta_{crit} = t^{-1}(1-t^{-1})^t$ and any natural number n there exists a binary code C of length n and rate $R \geq 2t^{-1}\varepsilon^2 \log_2 e$ such that for any t-subset $B \subset C$ and any codeword $\mathbf{c} \notin B$ there are at least $(\delta_{crit} - \varepsilon)n$ coordinates i in which $c_i = 1$ and $b_i = 0$ for all $b \in B$.*

Proof. We use random coding technique with expurgation for proof. Let us generate a binary random code C of length n and cardinality M by choosing every coordinate of a codevector equals 1 with probability t^{-1} and equals 0 with probability $1 - t^{-1}$ independently for different coordinates and different vectors. For a given vector $\mathbf{c}$ and a subset B we call a coordinate i "good" if $c_i = 1$ and $b_i = 0$ for all $b \in B$. Let χ_i be a random variable which equals 1 if i is "good" and equals 0 otherwise. Then the probability

$$p = \Pr(\chi_i = 1) = t^{-1}(1-t^{-1})^t = \delta_{crit} \tag{2.4.3}$$

and the expected number of "good" coordinates

$$E(\sum_{i=1}^{n} \chi_i = 1) = np = n\delta_{crit}. \tag{2.4.4}$$

Let us call a codevector $\mathbf{c}$ and t-subset $B \subset C$, where $\mathbf{c} \notin B$, an "ε-good couple" if the number of corresponding good coordinates is at least $n(p-\varepsilon)$, and call a couple $(\mathbf{c}, B)$ an "ε-bad couple" otherwise. The probability that a couple is bad equals to $P_{bad} = \Pr(\sum_{i=1}^{n} \chi_i = 1 < n(p-\varepsilon))$ and according to the Hoeffding inequality [81]

$$P_{bad} \leq e^{-2n\varepsilon^2}. \tag{2.4.5}$$

Hence the expected number of bad couples is $MC_{M-1}^{t} P_{bad}$ what is less than $\frac{M^{t+1}}{t!} P_{bad}$. Then we expurgate from every bad couple one element (for instance, $\mathbf{c}$) and the rest vectors of C will be a code with no bad couples. Choose maximal M in such a way that $P_{bad} M^{t+1} t! \leq M/2$, i.e., $M = \sqrt[t]{t! P_{bad}^{-1}/2}$. Therefore the resulting code without bad couples has cardinality at least $\sqrt[t]{t! P_{bad}^{-1}/2}/2$ and its rate

$$R \geq \frac{2\varepsilon^2}{t \ln 2}. \tag{2.4.6}$$

□

Let $M_t^*(n, L)$ denotes the maximum possible cardinality of the t-signature code that is able to correct up to L errors. Then let $R_t^A(n, \delta n) := n^{-1} \log_2 M_t^*(n, n\delta/2)$ denote the maximum possible rate of the (t, L)-signature code with $2L + 1 = \delta n$ and let $R_t^A(\delta) := \limsup_{n\to\infty} R_t(n, \delta n)$. Then for any $\delta < \delta_{crit} = t^{-1}(1 - t^{-1})^t$ the following holds

$$R_t^A(\delta) \geq \frac{2 \log_2 e}{t} \left(\delta_{crit} - \delta\right)^2. \tag{2.4.7}$$

Remark. It is worthy to note that $\delta_{crit} = t^{-1}(1-t^{-1})^t < (et)^{-1}$ and the Theorem remains valid if to replace δ_{crit} on $(et)^{-1}$.

2.4.2 Upper bound on the rate of error-resistant signature code

In this section we will prove the upper bound on the rate of error-resistant signature code for A-channel. The theorem 2.4.2 states the upper bound on the cardinality of a t-signature code that is able to correct up to L errors. Let $A(n', d')$ denote the maximal cardinality of a code in Hamming space of length n' and distance d', then we have the following estimation

Theorem 2.4.2.

$$M_t^*(n, L) \leq (t-1) A\left(\frac{n}{t-1}, \left\lfloor \frac{2L}{t-1} \right\rfloor + 1\right).$$

Before proving the theorem, we shall prove the following lemma.

Lemma 2.4.1. *Consider a binary code C of length n and the splitting of the coordinates of the code in $t-1$ sections $I_1, ..., I_{t-1} \subset \{1, ..., n\}$ of size $\frac{n}{t-1}$. If a code C is (t, L)-signature code then for any $c \in C$ there is an index $i \in \{1, ..., t-1\}$ such that $d_H(c|_{I_i}, a|_{I_i}) \geq \lfloor \frac{2L}{t-1} \rfloor + 1$ for any $a \in C, a \neq c$, where $a|_{I_i}$ denotes the restriction of the vector a to the positions corresponding to the indexes from I_i.*

Proof. Let's assume the opposite, i.e., there exists $c \in C$ such that for each $i \in [t-1]$ there exist a codeword $c^i \in C$ such that $d_H(c|_{I_i}, c^i|_{I_i}) \leq \lfloor \frac{2L}{t-1} \rfloor$. Then, consider two different subsets of active users $U = \{c^1, ..., c^{t-1}\}$ and $V = \{c^1, ..., c^{t-1}\} \cup \{c\}$, consider also the characteristic vectors u and v corresponding to the sets U and V. The distance $d_H(S(U), S(V)) \leq \lfloor \frac{2L}{t-1} \rfloor (t-1) = 2L$ which contradicts the fact that C is (t, L)-signature code. That concludes the proof of the lemma.

□

Now, we can move to the proof of the Theorem 2.4.2.

Proof. Let's make the same splitting of the coordinates of the code in $t-1$ sections $I_1, ..., I_{t-1} \subset \{1, ..., n\}$ of size $\frac{n}{t-1}$. Among all submatrices that correspond to the defined partitioning let's take the submatrix that contains the maximum number of (sub)words with respect to the condition of the lemma. The number of such words is at least $\frac{N}{t-1}$. We know that all such words are different from each other in at least $\lfloor \frac{2L}{t-1} \rfloor + 1$ positions, so we can estimate $\frac{N}{t-1}$ as

$$\frac{N}{t-1} \leq A\left(\frac{n}{t-1}, \left\lfloor \frac{2L}{t-1} \right\rfloor + 1\right).$$

□

The following statement reformulates the theorem in terms of the code rate. Recall that $R(\delta)$ denotes, as usual, the asymptotic maximal possible rate of a code in the Hamming space with relative distance δ. Then we have the following bound

Corollary 2.4.1.

$$R_t(\delta) \leq \frac{1}{t-1} R(\delta).$$

Proof of the corollary. From the theorem

$$\frac{N}{t-1} \leq A\left(\frac{n}{t-1}, \left\lfloor \frac{2L}{t-1} \right\rfloor + 1\right),$$

so

$$\frac{\log_2 N}{n} \leq \frac{\log_2 A\left(\frac{n}{t-1}, \lfloor \frac{2L}{t-1} \rfloor + 1\right)}{n} + \frac{\log_2(t-1)}{n}.$$

As $n \to \infty$, it can be rewritten as

$$R_t(\delta) \leq \frac{R(\delta)\frac{n}{t-1}}{n} = \frac{R(\delta)}{t-1},$$

where $\delta = \frac{2L+1}{n}$.

We mentioned in the previous section that any lower bound on the rate of the best t-cover-free code capable to correct L errors is valid for the best t-signature code correcting up to L errors. But for upper bounds the situation is converse. Namely, the same bound was proved in [77] and its proof is very simple, based just on counting arguments. We proved stronger result from which the result of [77] follows immediately.

Chapter 3

Signature codes for B-channel

3.1 Problem statement

In [15] in addition to A-channel authors also introduced a channel model called q-frequency M-user multiple access channel *with* intensity information, or, simply, B-channel. Formally, this channel model can be described as follows. Consider a multiple access channel with M users, where every user can use one of q frequencies $f_1, f_1, ..., f_q$ to transmit on at each time slot during a session. For the B-channel, the output is modeled determetistically, see Figure 2. Output at each time slot is a symbol that represents the subset of frequencies occurred as inputs to the channel at that time slot and the number of times that each frequency was used. This output model is stated in the name of the channel — for B-channel both the subset of used frequencies and how many users used a particular frequency at a particular time slot are known (intensity information). The output of the entire session for B-channel consists of n such symbols, one symbol for one time slot. For simplicity, the input alphabet will be represented as $\mathcal{A} = \{1, 2, ..., q\}$ and output alphabet as a set of all M-ary vectors of length q. Indeed, for the B-channel the output symbol can be represented as an M-ary vector of length q for which the j-th coordinate ($j \in [q]$) equals the number of times the frequency f_j was used by active users at considered time slot. In [15] the case of information transmission as well as the corresponding capacity region was considered. In this thesis we consider only the case of partial activity with activeness status, i.e., each user transmits as information its own status only: whether it is active or not. So, mathematically we have the following problem.

Consider an M-user multiple access channel with partial activity, where each of the users has one q-ary vector of length n to transmit, but at any time at most t users may transmit simultaneously. If the channel input is a set $U = \left\{\mathbf{u}^{(j)} = (u_1^{(j)}, \ldots, u_n^{(j)}) \mid j = 1, \ldots, \ell\right\} \subset \mathcal{A}^n$ of q-ary words where $|U| = \ell \leq t$, then its

output is an $n \times q$ matrix

$$S(U) = \begin{pmatrix} \mathrm{comp}(u_1^{(1)}, \ldots, u_1^{(\ell)}) \\ \vdots \\ \mathrm{comp}(u_n^{(1)}, \ldots, u_n^{(\ell)}) \end{pmatrix},$$

where $\mathrm{comp}(\cdot)$ is defined as follows.

Definition 3.1.1. The *composition of a vector* $\mathbf{y} = (y_1, \ldots, y_\ell) \in \mathcal{A}^\ell$ is a sequence $\mathrm{comp}(\mathbf{y}) = (w_1, \ldots, w_q)$ where w_a is the number of occurrences of a symbol $a \in \mathcal{A}$ in $\mathbf{y}$.

Thus, an entry $S(U)_{ij}$ of $S(U)$ is the number of elements $j \in \mathcal{A}$ occurring in the ith coordinate of vectors in U. Clearly, $w_1 + \ldots + w_q = \ell$, and the number of different compositions for fixed q and ℓ is $\binom{\ell + q - 1}{q - 1}$. In what follows, by $S(U)_i$ we denote the ith row of $S(U)$.

Example 3.1.1. Let $\mathcal{A} = \{1, 2, 3\}$, $n = 4$, and $U = \left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \\ 3 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 2 \\ 3 \\ 2 \end{pmatrix} \right\}$; then $S(U) = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 3 & 0 \\ 0 & 0 & 4 \\ 3 & 1 & 0 \end{pmatrix}$. Here, for example, the matrix entry $S(U)_{22} = 3$, since the element 2 occurs in the second coordinate of words in U exactly three times.

In what follows we refer to such a channel as a *compositional* channel or B-channel. For partial activity channels the problem is to construct such set of vectors that, given a channel output, one could correctly identify the set of active users. This requirement for B-channel can be formalized as follows.

Definition 3.1.2. A code $C \subseteq \mathcal{A}^n$ is said to be *t-signature* for B-channel if compositions of any t distinct vectors (codewords) of C or less are distinct, i.e., $S(U) \neq S(V)$ for any $U, V \subseteq C$, $U \neq V$, $|U|, |V| \leq t$.

In this thesis we study the asymptotic behavior of the rate of the best t-signature codes. Denote by $M_q(n, t)$ the maximum possible size of a q-ary t-signature code of length n, and by $R_q^B(n, t) = n^{-1} \log_q(M_q(n, t))$, the maximum possible rate of such a code. We are mostly interested in the behavior of the functions

$$\underline{R_q^B(t)} := \liminf_{n \to \infty} R_q^B(n, t) \qquad \text{and} \qquad \overline{R_q^B(t)} := \limsup_{n \to \infty} R_q^B(n, t).$$

The main result of the chapter is proving that for t large enough we have the following inequalities:

$$(q-1)\frac{\log_q t}{4t} - \frac{c_1}{4t} \le \underline{R_q^B(t)} \le \overline{R_q^B(t)} \le 2\left((q-1)\frac{\log_q t}{4t} + \frac{c_2}{4t}\right),$$

where $c_1 = c_1(q) = q(1-\log_q(2\pi))+2\log_q e$ and $c_2 = c_2(q) = q(-1+4\log_q e)$. In what follows, log is understood to be the logarithm to the base q, unless otherwise specified.

3.2 Lower Bound

The main result of this section is as follows.

Theorem 3.2.1. *There exists a number $t^* = t^*(q)$ such that for any $t \ge t^*$ we have the inequality*

$$\underline{R_q^B(t)} \ge \frac{(q-1)}{4t}\log t - \frac{q\log(\frac{q}{2\pi}) + 2\log e}{4t}. \tag{3.2.1}$$

Before proving the main result, we prove the following auxiliary lemma.

Lemma 3.2.1. *For any $k \ge 1$ we have*

$$\sum_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)^2 \le (2\pi)^{-q/2} e q^{2k+q/2} k^{\frac{1-q}{2}}.$$

Proof. Clearly,

$$\sum_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)^2 \le \left(\sum_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)\right)\left(\max_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)\right)$$
$$= q^k \left(\max_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)\right).$$

Note that the multinomial coefficient attains its maximum at values of $k_1,\dots,k_q$ close to k/q; therefore, without loss of generality, we will assume that $k_1 = \dots = k_q = k/q$ and replace the factorial with the gamma function if necessary. We obtain

$$\sum_{\substack{k_1,\dots,k_q \\ k_1+\dots+k_q=k}} \left(\frac{k!}{k_1!\dots k_q!}\right)^2 \le q^k \frac{k!}{\left(\left(\frac{k}{q}\right)!\right)^q}.$$

By Stirling's formula,

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$$

for any $n \ge 1$. We use a weak version of this inequality, namely

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \le n! \le e\sqrt{n}\left(\frac{n}{e}\right)^n.$$

The simplification of the upper bound follows from the fact that $\sqrt{2\pi}e^{\frac{1}{12n}} < e$ for $n \ge 2$, while for $n = 1$ we have $e\sqrt{n}\left(\frac{n}{e}\right)^n = 1! = 1$. It is known that

$$\Gamma(n+1) \ge n^n e^{-n}\sqrt{2\pi(n+1/6)} > n^n e^{-n}\sqrt{2\pi n}$$

for any real $n \ge 1$ (see [82]). For $n \in (0,1]$ it is known from [82] that

$$\Gamma(n+1) \ge \sqrt{2e}\left(\frac{n+0.5}{e}\right)^{n+0.5}.$$

One can easily show that

$$\sqrt{2e}\left(\frac{n+0.5}{e}\right)^{n+0.5} > n^n e^{-n}\sqrt{2\pi n}.$$

Indeed, the function $f(n) = \left(\frac{n}{n+0.5}\right)^{n+0.5}\sqrt{\pi}$ increases for $n \in (0,1]$. Hence, the maximum is attained at $n = 1$, but

$$\left(\frac{1}{1.5}\right)^{1.5}\sqrt{\pi} \approx 0.964\ldots < 1.$$

Therefore, $\Gamma(n+1) > n^n e^{-n}\sqrt{2\pi n}$ for any positive real n.

Turning back to estimating the sum of squared multinomial coefficients, we obtain

$$\sum_{\substack{k_1,\ldots,k_q \\ k_1+\ldots+k_q=k}} \left(\frac{k!}{k_1!\ldots k_q!}\right)^2 \le q^k \frac{k!}{\left(\Gamma\left(\frac{k}{q}+1\right)\right)^q} \le \frac{q^k e\sqrt{k}\left(\frac{k}{e}\right)^k}{\left(e^{-k/q}\sqrt{2\pi k/q}\left(\frac{k}{q}\right)^{k/q}\right)^q}$$
$$= \frac{eq^{2k+q/2}}{(2\pi)^{q/2}k^{q/2-1/2}} = (2\pi)^{-q/2}eq^{2k+q/2}k^{\frac{1-q}{2}}.$$

□

Now we come back to the proof of the theorem.

Proof. We use the random coding method. Consider a random code $C = \{c_1, \ldots, c_M\}$ with symbols chosen from a q-ary alphabet independently and equiprobably. Let us upper estimate the probability $\Pr(*)$ of the following event: the random code C is not t-signature.

Note that the equality $S(U) = S(V)$ implies $S(U \setminus V) = S(V \setminus U)$. Also, note that for any $U \subseteq C$ sums of entries in all rows of $S(U)$ are the same and equal the cardinality of the subset U. Thus, to check whether a code is t-signature, it suffices to consider only disjoint sets of the same cardinality no greater than t. Thus, a code is not t-signature if and only if there exist at least two disjoint subsets $U, V \subseteq C$ of cardinality $k \leq t$ with $S(U) = S(V)$. The union bound yields

$$\Pr(*) \leq \sum_{\substack{U,V \subseteq C \\ |U|=|V| \leq t \\ U \cap V = \emptyset}} \Pr(S(U) = S(V)) = \sum_{k=1}^{t} \left(\sum_{\substack{U,V \subseteq C \\ |U|=|V|=k \\ U \cap V = \emptyset}} \Pr(S(U) = S(V)) \right).$$

Note that $\Pr(S(U) = S(V)) = \prod_{i=1}^{n} \Pr(S(U)_i = S(V)_i)$. Since $\Pr(S(U)_i = S(V)_i)$ is independent of i, i.e., $\Pr(S(U)_1 = S(V)_1) = \ldots = \Pr(S(U)_n = S(V)_n)$, we have $\Pr(S(U) = S(V)) = (p(k))^n$, where $k = |U| = |V|$ and where $p(k)$ is the probability for compositions of two q-ary k-tuples with symbols chosen from a q-ary alphabet independently and equiprobably to coincide. Thus, we have the following inequality:

$$\Pr(*) \leq \sum_{k=1}^{t} \binom{M}{k} \binom{M-k}{k} (p(k))^n \leq \sum_{k=1}^{t} M^{2k} (p(k))^n. \tag{3.2.2}$$

Since the number of q-ary sequences of length k with a given composition $(k_1, \ldots, k_q)$ is $\dfrac{k!}{k_1! \ldots k_q!}$, we can represent $p(k)$ explicitly:

$$p(k) = \sum_{\substack{k_1, \ldots, k_q \\ k_1 + \ldots + k_q = k}} \left(\frac{k!}{k_1! \ldots k_q!} \right) \frac{1}{q^k} \cdot \left(\frac{k!}{k_1! \ldots k_q!} \right) \frac{1}{q^k}.$$

In what follows we will need the fact that for any $k \geq 1$ we have

$$p(k) \leq (2\pi)^{-q/2} e q^{q/2} k^{\frac{1-q}{2}}. \tag{3.2.3}$$

As it was stated in the lemma 3.2.1 for any $k \geq 1$ we have

$$\sum_{\substack{k_1, \ldots, k_q \\ k_1 + \ldots + k_q = k}} \left(\frac{k!}{k_1! \ldots k_q!} \right)^2 \leq (2\pi)^{-q/2} e q^{2k+q/2} k^{\frac{1-q}{2}}.$$

Then, clearly, we have that for any $k \geq 1$ we have

$$p(k) \leq (2\pi)^{-q/2} e q^{q/2} k^{\frac{1-q}{2}}.$$

Hence, it follows from (3.2.2) and (3.2.3) that for any k_0, $1 \leq k_0 \leq s$, we have

$$\Pr(*) \leq \sum_{k=1}^{k_0} M^{2k} p(k)^n + \sum_{k=k_0+1}^{t} M^{2k} \left(e \left(\frac{q}{2\pi} \right)^{q/2} k^{\frac{1-q}{2}} \right)^n, \tag{3.2.4}$$

and the claim of the theorem is valid if the right-hand side of (3.2.4) is strictly less than 1. To this end, it suffices to show that every term in each of the two sums is less than $1/t$. Let us begin with the terms of the second sum; for them, we have the following chain of equivalent inequalities:

$$M^{2k} \left(e \left(\frac{q}{2\pi} \right)^{q/2} k^{\frac{1-q}{2}} \right)^n < 1/t,$$

$$2k \log M + \log((2\pi)^{-nq/2} e^n q^{qn/2}) + \frac{n(1-q)}{2} \log k < \log(1/t),$$

$$2k \frac{\log M}{2nk} + \frac{\log((2\pi)^{-q/2} e q^{q/2})}{2k} + \frac{n(1-q)}{2} \frac{\log k}{2nk} < \frac{\log(1/t)}{2nk},$$

whence it follows that it suffices that the code rate satisfies the condition

$$R = \frac{\log M}{n} < F_q(k) - \frac{\log t}{2nk}, \quad \text{where} \quad F_q(k) = \frac{(q-1)}{4k} \log k - \frac{q \log\left(\frac{q}{2\pi} \right) + 2 \log e}{4k}.$$

Let us check the function $F_q(k)$ for monotonicity. Since

$$\frac{d}{dk} F_q(k) = (4k^2 \ln q)^{-1} \left(q \ln \frac{q}{2\pi} + q + 1 - (q-1) \ln k \right),$$

the function $F_q(k)$ strictly decreases for $k > \hat{k}$, where

$$\hat{k} = \left(\frac{q}{2\pi} \right)^{\frac{q}{q-1}} e^{\frac{q+1}{q-1}}.$$

Let $k_0 = \lceil q^{3/2} \rceil$. Since $\hat{k} < k_0$ for $q \geq 2$, the function $F_q(k)$ thereby strictly decreases for $k \geq k_0$. Hence, if $R < F_q(t)$, then

$$R < F_q(k) - \frac{\log t}{2nk}$$

for all $k \in [k_0, t]$ and n large enough, and therefore, as was noted above, each term of the second sum is strictly less than $1/t$.

Now we consider the first term and require that $M^{2k}p(k)^n < 1/t$ for any $k \in [1, k_0]$, i.e., $2k \log M + n \log p(k) < -\log t$, or equivalently,

$$R < \hat{F}(k) - \frac{\log t}{2kn}, \quad \text{where} \quad \hat{F}(k) = \frac{\log(1/p(k))}{2k}.$$

Similarly to the above, we obtain that if $R < R_1$, where $R_1 = \min\{\hat{F}(1), \ldots, \hat{F}(k_0)\}$, then for n large enough each term of the first sum in (3.2.4) is less than $1/t$. Hence, $R_q^B(t) \geq \min\{F_q(t), R_1\}$.

Since $F_q(t)$ tends to zero (monotonically for $t \geq k_0$), there exists t^* such that $F_q(t) \leq R_1$ for $t \geq t^*$. □

3.3 Upper Bound

Theorem 3.3.1. *Let t and q be constants with $t \geq q \geq 2$; then*

$$\overline{R_q^B(t)} \leq (q-1)\frac{\log t}{2t} + q\frac{-1 + 4\log e}{2t}.$$

Proof. Consider an arbitrary q-ary t-signature code C of length n and size M. Denote by U a discrete random variable uniformly distributed on the set of t-element subsets of C:

$$\Pr(U = A) = \begin{cases} \binom{M}{t}^{-1} & \text{if } A \subseteq C,\ |A| = t, \\ 0 & \text{otherwise.} \end{cases}$$

The main idea of the proof, inspired by [83, 84], is estimating from above and below the entropy of the random variable $S(U)$. These estimates will give an upper bound on the rate of an s-compositional code.

1. Let us estimate $\boldsymbol{H}(S(U)) = \boldsymbol{H}_q(S(U))$ from below. Since the code is t-signature, values of $S(U)$ for different values of U are also different, which implies that $\boldsymbol{H}(S(U)) = \boldsymbol{H}(U)$. Since U is uniformly distributed on t-element subsets A of C, we have

$$\boldsymbol{H}(U) = -\sum_{A:\,|A|=t} \Pr(U = A)\log(\Pr(U = A)) = \log\binom{M}{t} \geq s\log\frac{M}{t}(1 + o(1)).$$

2. Now we estimate $\boldsymbol{H}(S(U))$ from above. In this case the main idea consists in estimating the entropy of entries of the matrix $S(U)$ and applying the entropy property

$$\boldsymbol{H}(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} \boldsymbol{H}(X_i).$$

To simplify the notation, introduce vectors $\boldsymbol{w}(i) = (w_i^1, \dots, w_i^q)$, where w_i^j is the number of codewords with an element $j \in Q$ in the ith coordinate, and a vector $\boldsymbol{k} = (k_1, \dots, k_q)$; i.e., $k_j \geq 0$ for all j, and $k_1 + \ldots + k_q = t$. Then

$$\Pr\left(S(U)_i = \boldsymbol{k}\right) = p(\boldsymbol{k}, \boldsymbol{w}(i)) = \frac{\binom{w_i^1}{k_1}\binom{w_i^2}{k_2}\dots\binom{w_i^q}{k_q}}{\binom{M}{s}}.$$

Hence, for all $i = 1, \dots, n$, the random vector $S(U)_i$ has a multivariate hypergeometric distribution, and

$$\boldsymbol{H}(S(U)_i) = \boldsymbol{H}(\mathrm{Hyp}(M, t, \boldsymbol{w}(i))) = -\sum_{\boldsymbol{k}:\ k_1+\ldots+k_q=t} p(\boldsymbol{k}, \boldsymbol{w}(i)) \log(p(\boldsymbol{k}, \boldsymbol{w}(i))).$$

Put $p_j = \dfrac{w_i^j}{M}$.

Lemma 3.3.1. *For a fixed t and $M \to \infty$, for any $\boldsymbol{k}$ and $\boldsymbol{w}(i)$ we have the inequality*

$$-p(\boldsymbol{k}, \boldsymbol{w}(i)) \log(p(\boldsymbol{k}, \boldsymbol{w}(i))) \leq \frac{t!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} \log \frac{k_1! \dots k_q!}{t! (p_1)^{k_1} \dots (p_q)^{k_q}} \left(1 + O\left(\frac{1}{M}\right)\right)$$

Proof. Note that

$$p(\boldsymbol{k}, \boldsymbol{w}(i)) = \frac{\binom{w_i^1}{k_1}\binom{w_i^2}{k_2}\dots\binom{w_i^q}{k_q}}{\binom{M}{t}} \leq \frac{(w_i^1)^{k_1} \dots (w_i^q)^{k_q}}{k_1! \dots k_q!} \frac{t!}{M^t} \left(\frac{M}{M-t}\right)^t,$$

i.e.,

$$\begin{aligned} p(\boldsymbol{k}, \boldsymbol{w}(i)) &\leq \frac{s!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} \left(1 + \frac{t}{M-t}\right)^t \\ &= \frac{t!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} \left(1 + O\left(\frac{1}{M}\right)\right). \end{aligned} \tag{3.3.1}$$

Since the function $-x \log x$ is strictly increasing in the interval $[0, e^{-1}]$, the claim of the lemma is valid if

$$\frac{s!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} (1 + O(1/M)) < e^{-1}.$$

If $t!(k_1! \dots k_q!)^{-1} p_1^{k_1} \dots p_q^{k_q}(1 + O(1/M)) \geq e^{-1}$, then for all j with $k_j > 0$ we have

$$p_j \geq p_j^{k_j} \geq p_1^{k_1} \dots p_q^{k_q} \geq (et!(1 + O(1/M)))^{-1} \geq c > 0,$$

where c is a positive constant. Thus, for any j, either k_j is zero or $w_i^j \geq cM$. Then

$$p(\boldsymbol{k}, \boldsymbol{w}(i)) \geq \frac{\prod\limits_{j:\, k_j>0} (w_i^j - k_j + 1)^{k_j}}{k_1! \dots k_q!} \frac{s!}{M^t} \geq \frac{t!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} \left(1 - \frac{t}{cM}\right)^t.$$

The obtained inequality and inequality (3.3.1) imply that in the considered case we have $p(\boldsymbol{k}, \boldsymbol{w}(i)) = \dfrac{t!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q}(1 + O(1/M))$. Since the derivative of $-x \log x$ is bounded in the interval $[e^{-1}, 1]$, the claim of the lemma holds in this case with equality, which completes the proof of the lemma. □

The lemma implies

$$\begin{aligned}\lim_{M\to\infty} \boldsymbol{H}(\mathrm{Hyp}(M, t, \boldsymbol{w}(i))) &\leq \sum_{k_1+\dots+k_q=t} \frac{s!}{k_1! \dots k_q!} p_1^{k_1} \dots p_q^{k_q} \log \frac{k_1! \dots k_q!}{t! p_1^{k_1} \dots p_q^{k_q}} \\ &= \boldsymbol{H}(\mathrm{Mult}(t, p_1, \dots, p_q)).\end{aligned}$$

It was proved in [85] that the entropy of the polynomial distribution attains its maximum at $p_1 = \dots = p_q = \dfrac{1}{q}$, which implies

$$\lim_{M\to\infty} \boldsymbol{H}(S(U)_i) \leq \boldsymbol{H}(\mathrm{Mult}(t, 1/q, \dots, 1/q)).$$

Now the problem reduces to estimating the entropy of the polynomial distribution at the point $(1/q, \dots, 1/q)$. It [86] there was proved an upper bound on the entropy of the polynomial distribution in the general case. Namely, it was proved that for any admissible values of the parameters q and t one has

$$\begin{aligned}\boldsymbol{H}(\mathrm{Mult}(t, p_1, \dots, p_q)) &\leq \frac{1}{2} \log \left((2\pi e t)^{q-1} p_1 \dots p_q\right) - \frac{q}{2} \log(2\pi) \\ &\quad + \left(\frac{3}{2} \sum_{i=1}^q \frac{p_i}{q_i} + (t-1) \sum_{i=1}^q \frac{p_i^2}{q_i} - t + \frac{1}{2}\right) \log e \\ &\quad - \sum_{i=1}^q \left(t p_i + \frac{1}{2}\right) \log \frac{p_i}{q_i} - \frac{1}{2} \sum_{i=1}^q (1 - p_i)^t \log t e q_i,\end{aligned}$$

where $q_i = \max\left\{\dfrac{1}{t}, p_i\right\}$.

In our case, $t \geq q$; therefore, $p_1 = \dots = p_q = q_1 = \dots = q_q = 1/q$ and

$$\begin{aligned}\boldsymbol{H}(\mathrm{Mult}(t, 1/q, \dots, 1/q)) &\leq \frac{1}{2} \log \left((2\pi e t)^{q-1} q^{-q}\right) - \frac{q}{2} \log(2\pi) + \left(\frac{3}{2} q - \frac{1}{2}\right) \log e \\ &\quad - \frac{1}{2} \sum_{i=1}^q \left(1 - \frac{1}{q}\right)^t \log t e \frac{1}{q} \leq \frac{q-1}{2} \log t + \frac{q}{2} \log(e^4 q^{-1}).\end{aligned}$$

Combining the upper and lower bounds and using the sub-additivity property of the entropy, $\boldsymbol{H}(X_1,\ldots,X_n) \le \sum_{i=1}^{n} \boldsymbol{H}(X_i)$, we conclude that, as $M \to \infty$, we have

$$t \log \frac{M}{t} \le \boldsymbol{H}(S(U)) \le n \left(\frac{q-1}{2} \log t + \frac{c_2}{2} \right) (1 + o(1)),$$

where $c_2 = q \log\left(\frac{e^4}{q}\right)$. Since the code rate is $R = \frac{\log M}{n}$, these inequalities imply

$$R \le \frac{(q-1)\log t + c_2}{2t}(1 + o(1)) + \frac{t \log t}{n};$$

therefore, as $n \to \infty$ and $M \to \infty$ with t fixed, we obtain the following upper bound on the asymptotic rate of q-ary t-signature codes:

$$\overline{R_q^B(t)} \le (q-1)\frac{\log t}{2t} + \frac{c_2}{2t}.$$

The theorem is proved. □

3.4 Binary adder and B-channels, and their generalizations

It is clear that in the binary case B-channel and adder channel coincide (almost, see remark below), and one of natural questions is how to generalize them to the non-binary case. The most interesting generalization is introduced by us the weighted adder channel which is of independent interest, and will be applied in Chapter 4 to construction of multimedia digital fingerprinting codes.

Remark. Let us remind that for binary adder channel its i-th output symbol is just the number of ones at the i-th position of vectors of active users. And for binary B-channel its i-th output symbol is the number of zeros and ones. It looks like that B-channel provides more information. The difference is actually not so important, because it is possible to receive the same amount of information from the adder channel by increasing the code length by one. Indeed, let us add extra coordinate to user's vector and set its value equals to one. Then, the value at this position of the output vector will be equal to the number of active users. Asymptotically, this additional coordinate will not make any changes for the rate of corresponding signature codes, which means that these models of channels are equivalent.

Let us continue with an unexpected observation that non-binary B-channel investigated in previous sections is not, in fact, so much nonbinary. Namely, consider q-ary B-channel in the corresponding binary form where a symbol $i \in$

$\{1,\dots,q\}$ corresponds to a binary vector of length q with 1 in the i-th position, and all other symbols equal to zero. Thus, a q-ary signature code of length n and the cardinality M becomes binary, its size is still M, and the length of binary codewords is $n_2 = qn$, i.e., becomes q times as large. Coming back to the original problem setting of communication channels with multiple frequency modulation let us asume now that users are allowed to employ not a single but any number of frequencies, i.e. they can transmit *arbitrary* binary vectors. Then transmission over a q-ary B-channel turns into transmission over a binary adder or B-channel. Therefore, according to [30, 83], we have the following upper and lower bounds on the best possible rate R_2 of an *arbitrary* binary code:

$$\frac{\log_q t}{4t\log_q 2} = \frac{\log_2 t}{4t} \lesssim R_2 = \frac{\log_2 M}{n_2} \lesssim \frac{\log_2 t}{2t} = \frac{\log_q t}{2t\log_q 2}, \tag{3.4.1}$$

whence for the best possible rate R_q (recalculated to the base q) follows

$$\frac{q\log_q t}{4t} \lesssim R_q = \frac{\log_q M}{n} \lesssim \frac{q\log_q t}{2t}. \tag{3.4.2}$$

Thus, first, we obtain a more general upper bound, which is only in $\frac{q}{q-1}$ times bigger than the upper bound of Theorem 3.3.1. Second, the more intriguing is our lower bound of Theorem 3.2.1, since we consider only q different "signals" instead of 2^q ones allowed in (3.4.1), whereas the obtained rate is worse by a factor of $\frac{q}{q-1}$ only.

Now let us consider possible generalizations of the binary adder channel. First generalization is rather obvious, namely, to consider nonbinary input alphabet. Note, that a choice of the corresponding set of integers is not unique even for $q = 3$. Indeed, one can chose $X = \{0, 1, 2\}$, but other can chose $X = \{0, 1, 3\}$ and these channels as well as signature codes to them clearly not equivalent. Some recent results on signature codes for nonbinary adder channel, with or without noise, can be found in [87].

Here we pay attention to more interesting from our point of view generalization of binary adder channel, which was introduced in the Thesis under the name weighted binary adder channel, see section 1.1.7. and [88]. The generalization consists in adding "weights" to the active users. Indeed, in all previous formulation we considered only the fact of being active which was modeled as (specifically defined) multiplication of the corresponding vector by 1 if user is active or by 0 if user is inactive. The natural step towards the generalization of such model would be to consider the arbitrary weights for vectors of active users. As a motivation of such definition one can consider a wireless (mobile) network where the energy of received signals is dependent on how far these users are from the receiver, aka

base station. More formally, the output y for such channel, called weighted binary adder channel, is defined in the following way is modeled as the most general case of adder channel, namely,

$$y = \sum_{j \in J} \lambda_j x_j, \tag{3.4.3}$$

where J denotes the set of active users with $|J| \leq t$. The coefficients λ_j, called weights or gains, are any positive real numbers. The problem is to construct a set of M code words (vectors) such that the sum of code words of any set of t or less active users is uniquely decodable, i.e., subsets of active users and corresponding outputs are in 1-1 correspondence. Note, that the case when all λ-s equal 1 is the standard formulation of adder channel with partial activity.

Formally a signature code for the weighted binary adder channel (wBAC) can be formulated as follows.

Definition 3.4.1. The set of n-dimensional binary vectors $h_j = (h_{1j}, ..., h_{nj}) \in \{0,1\}^n$ is called a t-signature code of length n for wBAC iff for any subsets $J, J' \subset \{1, ..., M\}$ such that $|J|, |J'| \leq t$ the following equality

$$\sum_{j \in J} \lambda_j h_j = \sum_{j \in J'} \lambda'_j h_j \tag{3.4.4}$$

implies that $J = J'$ (and $\lambda_j = \lambda'_j$ for all j).

Denote by $M(n,t)$ the maximal possible cardinality of a binary t-signature code of length n for wBAC. , i.e., the maximal cardinality of a set of binary vectors in n-dimensional Euclidean space for which the condition (3.4.4) holds.

Theorem 3.4.1.

$$M(n,t) \geq 2^{\lfloor n/t \rfloor}.$$

Proof. It is easy to see that (3.4.4) is equivalent to the linear independence (over real numbers) of any $2t$ vectors h_j . In order to construct such a set let us consider an irreducible binary Goppa code of length 2^m with $r \leq tm$ redundancy symbols which corrects t errors. Then, all 2^m columns of a parity-check matrix of this code forms the desired binary t-signature code of length r for wBAC. Indeed, any $2t$ columns are linear independent over the field of residues by module 2, and hence they are linear independent over the field of rational numbers and also over the real numbers since in all cases dependency means that the determinant of the corresponding $t \times t$ minor equals to 0.

□

So, in terms of the rate $R_t^{wBAC} := n^{-1} \log_2 M(n,t)$, the theorem 3.4.1 give the following result:

$$R_t^{wBAC} \geq t^{-1}(1 + o(1)). \tag{3.4.5}$$

Let us also denote by $n(M,t)$ the minimal possible dimension n such that $M(n,t) \geq M$, i.e., the minimal dimension in which there exists a code of the cardinality M with any $2t$ code vectors are linear independent, or equivalently, there is a t-signature code for wBAC with M code vectors. Now Theorem 3.4.1 can be rewritten in the following way:

$$n(M,t) \leq t\lceil \log_2 M \rceil. \tag{3.4.6}$$

Now let us show how signature codes for wBAC are related to the non-adaptive search of counterfeit coins on a precision scale. Recall the statement of this problem. There are M coins. Let us enumerate them and let $x_1, \ldots, x_M$ be their weights with at least $M - t$ of them being of equal weight, say a. Denote $I = \{i : \ x_i = a\}$ and $J = [n] \setminus I$, with $|J| \leq t$. There is a precision scale that allows to know the exact weight of any subset of coins. Any non-adaptive search with n weightings is uniquely defined by its $n \times M$ binary *search matrix* H which i-th row is the characteristic vector of the i-th weighted subset of coins. The property that a given non-adaptive search defined by H can find all weights is equivalent to the property that if $H\mathbf{x}^T = H\mathbf{y}^T$ then $\mathbf{x} = \mathbf{y}$.

Denote by $Q(M,t)$ the minimal number n of non-adaptive weightings which allows to find weights for all coins.

The next result can be found implicitly in [35]

Proposition 3.4.1. $n(M,t) \leq Q(M,t) \leq 2t + 1 + n(M,t)$

Let us prove that any $2t$ columns of H are linear independent over $\mathbb{R}$ and hence $Q(n,t) \geq n(M,t)$. Indeed, let assume the inverse. Consider $2t$ columns which are linear dependent, i.e.,

$$\sum_{k=1}^{2t} \lambda_k \mathbf{h}_{i_k} = 0,$$

where $\mathbf{h}_j$ is the j-th column of H. Then $H\mathbf{x}^T = H\mathbf{y}^T$, where $x_{i_k} = \lambda_k$ for $k = 1, \ldots, t$ and the rest $x_i = 0$, versus $y_{i_k} = \lambda_k$ for $k = t+1, \ldots, 2t$ and the rest $y_i = 0$.

Now let us show that $Q(M,t) \leq 2t + 1 + n(M,t)$. Let H_0 be $n(M,t) \times M$ matrix, in which any $2t$ columns are linear independent, and let I_m be $m \times m$ unit matrix. Construct a matrix

$$H = \left(\begin{array}{c|c} I_{2t+1} & \mathbf{0} \\ \hline H_0 & \end{array} \right)$$

Let $\mathbf{s} = (s_1, \ldots, s_n) = H\mathbf{x}^T$. The following algorithm finds $\mathbf{x}$.
First of all, $a := maj\{s_1, \ldots, s_{2t+1}\} = maj\{x_1, \ldots, x_{2t+1}\}$. This the point which differs signature codes for wBAC and the considered problem, because wBAC corresponds to the case when a is known and equals to zero!
Then choose a subset $L \subset [M]$ s.t. $|L| = t$ and solve (if possible) the following system of linear equations $H_0\mathbf{x}^T = \mathbf{s}_0$, where $\mathbf{s}_0 = (s_{2t+2}, \ldots, s_n)$, $x_j : j \in L$ are unknown variables and $x_i = a$ for all $i \notin L$. For $L = J$ this system has the solution and any two solutions will give a linear dependence between at most $2t$ columns of H_0 what contradicts to the choice of H_0.

It follows from Proposition 3.4.1 and Theorem 3.4.1 that

$$Q(n,t) \leq 2t + 1 + n(M,t) \leq t \log_2 n(1 + o(1)). \tag{3.4.7}$$

It was previously known that $Q(n,t) = O(t \ln n)$ [35].
The best known lower bound

$$Q(n,t) \geq 2\frac{t}{\log_2 t} \log_2 n(1 + o(1))$$

follows from the known upper bound on the cardinality of t-signature codes for the binary adder channel, see [30, 83].

There are at least three open questions:

1. what is $Q(n,t)$ for $t = fixed$ and $n \to \infty$?
2. what is $Q(n,t)$ for $t = \lambda n$ and $n \to \infty$?
3. to develop "decoding" algorithm which finds weights for all coins with low polynomial complexity for $t = fixed$.

Chapter 4

Applications

In this chapter we present three possible practical applications of signature codes. The first two applications come from the digital right management technologies, and are known as multimedia digital fingerprinting codes and digital fingerprinting codes with traceability property. Such codes represent the base of the technology that is aimed to ease the search of the illegal redistribution in the case of collusion attacks, i.e., when a group of users collude to produce a forged copy of the content. The last application is the so-called symmetric group testing problem — modification of general group testing problem characterized by more information as the answer. In what follows we give all necessary definitions and explanations that are needed to show the relationship of these problems to the construction of signature codes for A& B-channels.

4.1 Multimedia digital fingerprinting codes

We start with the multimedia digital fingerprinting (MDF) codes. In this section we will show how different reformulations of such problem can be reduced to the problem of constructing signature codes for binary A-channel and for weighted adder channel.

It is not a secret, that with the rapid development of multimedia technologies and the steady growth in the use of the Internet, a digital marketplace where a wide range of multimedia content (such as image, video, audio, speech...) is available, has become increasingly popular. However, the ease with which digital content can be accessed, retrieved and manipulated, poses the challenging task of devising methods for copyright protection and prevention of redistribution. One of the prominent techniques that is used to achieve that goal is called digital fingerprinting. The main idea consists in embedding in each copy of digital content a personalized mark, also called *fingerprint* or *watermark*. In what follows we will use both names. Although, today this type of techniques is in popular demand, the first examples of usage of same method can be found even in XVII century. A

creator-owner of logarithm tables used to introduce tiny errors in the insignificant digits of $\log x$ for few specific values of x. Had a malicious user of a logarithmic table sold illegal copies of it, the errors in the table would have allowed to identify who was the owner. This is an example of a standard digital watermarking technique which is widely used nowadays. Embedding of different marks into copies of different users allows to recognize the cases when a single dishonest user produces an illegal copy of distributed content. There are also the cases when the users possessing different copies of the same content collude (forming a coalition) and produce a forged version of the content based on their copies. Such type of attacks, also known as collusion attacks, is of interest and requires the development of more sophisticated techniques that allow to maintain the security at due level. In this chapter we omit the mark embedding procedure and concentrate on the creation of the set of fingerprints resistant to collusion attacks.

There are many different ways of modeling the broadcasting, in this thesis we consider the continuous model, usually called *multimedia model*, when the digital content x is represented by a vector over the field of real numbers (often called as host signal). The continuous model was firstly considered in the papers [89, 90] in the beginning of this century. In order to create the set of fingerprints, the distributor chooses n orthonormal noise-like signals (vectors) with small energy compare to the host signal, and forms fingerprints as linear combinations of these vectors. The coefficients of linear combination are from $\{-1, +1\}$ or $\{0, 1\}$ and are different for different users. Thus, the vector of coefficients uniquely identify a particular user. As for embedding of the marks, the additive embedding procedure is used, i.e., the marked content is just a sum of original content vector and fingerprint vector. To create a forged copy the colluders calculate the linear combination of their copies where the sum of coefficients (of weights) equals one, which is needed to be sure that the host signal was not changed during fingerprint embedding. The fact that users are restricted to the linear attacks and cannot manipulate the individual (basis) signal constitutes the so-called *marking assumption for multimedia model*. Assuming that the size of the coalitions is at most t the main problem is to create a set of vectors (of coefficients) of maximal possible cardinality in such a way that the distributor can reveal with zero-error of identification at least one participant of the malicious coalition or maybe the whole coalition.

Formally, suppose that the multimedia content is represented as a real-valued vector $\mathbf{x} \in R^m$, see [4, 89–91]. To prevent unauthorized redistribution of $\mathbf{x}$, the dealer constructs a set of watermarks using a linear modulation scheme that employs orthonormal vectors $\{\mathbf{f}_i \in R^m \mid i = 1, \ldots, n,\ n \leq m\}$ of *noise-like* signals. The fingerprint $\mathbf{w}_j$ of the $j-$th user ($j \in \{1, \ldots, M\}$) is represented as

follows:

$$\mathbf{w}_j = \sum_{i=1}^{n} h_{ij}\mathbf{f}_i, \tag{4.1.1}$$

where $h_{ij} \in \{+1, -1\}$ is used for antipodal signals and $h_{ij} \in \{0, 1\}$ for on-off keying type of modulation. The j-th user receives the vector

$$\mathbf{y}_j = \mathbf{x} + \mathbf{w}_j,$$

where it is assumed that $\|\mathbf{x}\|_2 \gg \|\mathbf{w}_j\|_2$ in order that the watermarking scheme do not introduce significant changes in the host signal. A group of users, called *colluders* or traitors, aims to create an unauthorized copy of the content such that the problem of tracing back to the source of leakage becomes a difficult task for a distributor. An important concept in digital fingerprinting, the *Marking Assumption* of [63], in the context of multimedia fingerprinting can be expressed in the following way: we assume that the members of the pirate coalition $J \subset \{1, \ldots, M\}$ cannot manipulate individual signals $\mathbf{f}_j$, and are limited to *linear attacks*. By a linear attack we mean that in order to generate a forged copy $\hat{\mathbf{y}}$ of the host content the pirates compute a linear combination of their copies $\mathbf{y}_j$ with some coefficients λ_j

$$\hat{\mathbf{y}} = \sum_{j \in J} \lambda_j \mathbf{y}_j, \tag{4.1.2}$$

where $\lambda_j > 0$ for all j and $\sum_{j \in J} \lambda_j = 1$. Using these assumptions, we see that the copy of the signal that they form

$$\hat{\mathbf{y}} = \mathbf{x} + \sum_{j \in J} \lambda_j \mathbf{w}_j \tag{4.1.3}$$

is still a usable copy of the host signal $\mathbf{x}$. The task of the dealer is to find the entire coalition J or at least one of its members based on the knowledge of the vector $\mathbf{T} = \mathbf{T}(J, \{\lambda_j\}) = (\tau_1, \ldots, \tau_n)$, where

$$\tau_k = (\widehat{\mathbf{y}} - \mathbf{x}, \mathbf{f}_k) = \left\langle \sum_{i=1}^{n} \sum_{j \in J} \lambda_j h_{ij} \mathbf{f}_i, \mathbf{f}_k \right\rangle = \sum_{j \in J} \lambda_j h_{kj} \tag{4.1.4}$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product. Equivalently,

$$\mathbf{T} = \sum_{j \in J} \lambda_j \mathbf{h}_j, \tag{4.1.5}$$

where $\mathbf{h}_j = (h_{1j}, \ldots, h_{nj})$ is a vector of coefficients used to create fingerprint for j-th user via linear combination.

Definition 4.1.1. A binary code $C = \{\mathbf{h}_1, ..., \mathbf{h_M}\} \subset \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n}}$ is called t-MDF code if for any coalition $J \subset [M], |J| \leq t$ it is possible to identify at least one user from J given $\mathbf{T} = \sum_{j \in J} \lambda_j \mathbf{h}_j$ for any choice of λ-s s.t. $\lambda_j > 0, j \in J$ and $\sum_{j \in J} \lambda_j = 1$.

Very often we need stronger property, namely, that a code can reveal the whole coalition.

Definition 4.1.2. A binary code $C = \{\mathbf{h}_1, ..., \mathbf{h_M}\} \subset \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n}}$ is called *strong* t-MDF code if for any coalition $J \subset [M], |J| \leq t$ it is possible to identify the whole coalition J given $\mathbf{T} = \sum_{j \in J} \lambda_j \mathbf{h}_j$ for any choice of λ-s s.t. $\lambda_j > 0, j \in J$ and $\sum_{j \in J} \lambda_j = 1$.

In overwhelming majority of works devoted to multimedia fingerprinting, starting from [89] were limited to the *averaging attack* of traitors, i.e., all the coefficients of linear combination are equal, or, equivalently, an attack with $\lambda_j = |J|^{-1}$ under the premise that this choice is the most powerful available attack. From cryptography point of view the averaging attack assumption is incorrect since in this case the coalition's strategy is known to the distributor which contradicts to the basic principles of cryptography. In other words, identifying the pirates under attacks with unknown λ-s is a priory much stronger task comparing with the case when all λ-s are known. Hence, in the Thesis we consider the case of positive arbitrary weights and propose more efficient results comparing to the existing results for stronger case (with averaging attack assumption). Moreover, even under this (oversimplified) assumption, all the families of multimedia fingerprinting codes known in the literature have code rate that tends to zero as the code length approaches infinity [4], [90].

In what follows we will show that the idea of using separating codes that was explained in the first chapter allows to prove the existence of good, i.e., positive-rate multimedia fingerprinting codes. Namely, it turned out that the quantization of the MDF problem allows to prove that the rate of the best codes is at least $\Theta(t^{-2})$, where t is the size of the traitor coalition. Moreover, we will show that these codes have a much stronger property than usual MDF codes, namely, they reveal the whole coalition rather than just a single pirate. For this reason we call them *strong* multimedia codes, see definition 4.1.2. It can be shown that the general case, i.e., without quantization, is equivalent to the signature codes for weighted binary adder channel. So, the lower bounds on the size of such families allow to receive even better results for the rate of strong MDF codes.

Quantized version. As earlier, the dealer attempts to find the coalition J via the vector $\mathbf{T} = (\tau_1, \ldots, \tau_n)$, see (4.1.4) and (4.1.5). Let $h_{ij} \in \{0, 1\}$, i.e.,

assume an on-off keying code modulation. Then $\tau_k \in [0,1]$, where

$$\begin{aligned}
&\tau_k = 0 \text{ iff } h_{kj} = 0 \text{ for all } j \in J,\\
&\tau_k = 1 \text{ iff } h_{kj} = 1 \text{ for all } j \in J,\\
&0 < \tau_k < 1 \text{ otherwise.}
\end{aligned}$$

It was suggested in [4, 89] to consider the following reduced *discrete/ quaintized* model when the dealer knows only whether $\tau_k = 0$, or $\tau_k = 1$, or that $0 < \tau_k < 1$, but does not know the exact value of τ_k in the last case. Denote by $F(.)$ the following mapping of the segment $[0,1]$ onto the ternary alphabet consisting of the elements $0, 1$ and $\{0,1\}$:

$$F(\tau) = \begin{cases} \tau & \text{if } \tau \in \{0,1\},\\ \{0,1\} & \text{if } 0 < \tau < 1. \end{cases}$$

The task of the dealer is to identify the coalition J or at least a part of it based on the knowledge of the following vector

$$S = F(\mathbf{T}) = (F(\tau_1), \ldots, F(\tau_n)). \tag{4.1.6}$$

To accomplish this, the dealer should construct a binary code $C = \{\mathbf{h}_1, ..., \mathbf{h}_M\}$ of cardinality M and length n. Given a coalition J, we refer to the corresponding set of codevectors $U = \{\mathbf{h}_j : j \in J\} \subset C$ also as a coalition. Let denote by

$$U_k = \{u_k : \mathbf{u} \in U\}$$

the set of values at the k-th coordinate of the vectors $\mathbf{u} \in U$. Then, the coalition U creates the following *unique* forged multimedia fingerprint whose coordinates are from ternary alphabet $\{0, 1, \{0,1\}\}$ which is exactly the vector

$$S(U) = F(\mathbf{T}) = (U_1, \ldots, U_n). \tag{4.1.7}$$

We call the vector $S(U)$ the *signature* of the coalition U. The goal of the dealer is to construct a code C such that any coalition $U, |U| \leq t$ can be uniquely recovered from its signature $S(U)$.

Now it is easy to see that constructing such code is equivalent to constructing a *t-signature code* for A-channel of [15] with the number of "frequencies" $q = 2$.

Let us recall some concepts from the theory of fingerprinting codes, see [45, 49, 63, 72]. Suppose that the set of possible forgeries that can be created by a coalition U is given by

$$D(U) = \{\mathbf{x} = (x_1, \ldots, x_n) : x_k \in U_k,\ k = 1, \ldots, n\}. \tag{4.1.8}$$

The set $D(U)$ is called the descendant set of the coalition.

Definition 4.1.3. [45] The code C is a code with the t-identifiable parent property, or a t-IPP code for short, if for any $\mathbf{x}$ either

$$\bigcap_{U: x \in D(U),\ |U| \le t} U \neq \emptyset, \tag{4.1.9}$$

or there is no coalition $U \subset C, |U| \le t$ such that $\mathbf{x}$ is one of its descendants, i.e., $\mathbf{x} \notin D(U)$ for any $U \subset C,\ |U| \le t$.

A significant difference between these (traditional) IPP codes and multimedia fingerprinting codes is that in multimedia fingerprinting a coalition can create only a single forged fingerprint, contrary to the large set of descendants in the case of IPP-codes. In other words, the model of multimedia fingerprinting assigns the label $\{0, 1\}$ to the positions of the forgery in which the IPP model allows inserting any binary element as long as it follows the Marking Assumption. Therefore, in multimedia the dealer's goal can be much more ambitious, namely, he attempts to identify all the members of the coalition instead of finding at least one of its members which is the task for the IPP-code problem (note that binary t-IPP codes of cardinality more than 2 do not exist).

Definition 4.1.4. [4, 92]. A binary code C is a multimedia code with a strong t-IPP property, or a strong t-MDF code for short, if for any two distinct coalitions $U, V \subset C, |U| \le t, |V| \le t$ their signatures are different:

$$S(U) = (U_1, \ldots, U_n) \neq S(V) = (V_1, \ldots, V_n). \tag{4.1.10}$$

Strong t-MDF codes were defined in [4], Definition 4.1, where they are called $\bar{t}$-*separable codes*. But, actually, this definition coincides with the general definition 2 for signature codes for A-channel. Another definition of strongly separable codes was given in [93].

Now we can use the results received for the signature codes for A-channel to establish new results for the MDF codes. According to the theorem 2.2.1 and (2.2.6) we have the following best known bound on the rate of t-signature codes for A-channel or, equivalently, strong t-MDF code:

$$\Theta(t^{-2}) \le R^*_{t-MDF} \le O\left(\frac{\log t}{t^2}\right). \tag{4.1.11}$$

Note, that the equivalence to MAC problem allowed not only to improve the estimation of the rate of corresponding codes but also to consider more general strategies of the coalitions, where we did not make any restrictions on the used weights. Also, the construction of signature codes and the algorithm of finding the set of active users proposed in the first chapter can be used for the MDF model.

So, the result achieved in the first chapter can be reformulated for MDF codes in the following way: there is a construction of strong t-MDF codes with the rate of order at lest t^{-3} and with decoding complexity polynomial in the length of the code.

If we don't restrict ourselves to the complete identification of the participants of malicious coalitions than we also can use the results about the recovery of at least one user which goes perfectly along with the initial tracing traitor problem from [44] and [45] where IPP codes were introduced. Indeed, as it was introduced in the definition 4.1.9, usually the recovery of at least one user is demanded. So, if we go back to *single user tracing* codes for OR channel, we can see that they provide exactly the solution for quantized MDF problem with partial recovery. That means that in such case we can attain much bigger rate, namely of order t^{-1}.

Another important improvement that we can get from the results for MAC is the noisy case. Indeed, in the first chapter we considered the case of erroneous output vector which in MDF model correspond to the signature of the coalition. The presence of noise makes MDF problem more realistic, since there are at least three sources of errors: due to transmission, to calculation of T, or adversarial noise.

Without quantization

Let us come back to the general formulation of MDF codes in definition 4.1.1 when the dealer tries to reveal a malicious coalition or at least one of its members from the knowledge of vector $\mathbf{T}$. In fact this problem is equivalent to the problem of constructing signature codes for weighted binary adder channel. Indeed, the problem of signature codes for WbAC is to construct a subset of binary vectors such that all possible linear combinations of t or less vectors are different. The MDF problem is the same, since we consider the most general strategy of the coalition without any restrictions on the value of coefficients. It was shown in Theorem 3.4.1 that columns of a parity-check matrix of Goppa or BCH codes, correcting t errors, form a t-signature code for weighted binary adder channel, hence, they form a t-strong MDF code what results in the following lower bound on the rate of t-strong MDF codes

$$R_{t-MDF} \geq \frac{1}{t}(1 + o(1)).$$

This solution improves the best known bound for quantized/ hard decoding reformulation stated in (4.1.11).

4.2 Constant weight IPP codes

In this section we consider another model of data redistribution with protection from illegal redistribution. Namely, we present a new approach to broadcast encryption and the corresponding "fingerprinting" codes and show how they generalize (in some sense) the idea of signature codes for multiple access channels.

Consider a distribution model where a dealer uses a broadcast channel to transmit some digital content to a wide audience. In order to restrict the access to the distributed content only for the authorized users (who paid for the access) the distributor should use broadcast encryption schemes. For the first time such schemes were considered in [94]. In what follows we will be interested in broadcast encryption schemes resistant to the so-called *collusion attacks* [44]. Such type of attacks can be described as follows.

To prevent unauthorized users from accessing the data, the distributor encrypts the data blocks with session keys and gives each authorized user the corresponding personal decoder, consisting of the personal set of keys needed to decrypt the data. Note that different users receive different decoders. Malicious users, who want to resell the access to the distributed content without revealing their identities, can form a group (coalition of traitors) and, based on their common knowledge (present keys and decoders), create a forged decoder. This type of forgery constitutes the main idea of a collusion attack. So, assuming that the cardinality of a possible coalition is not grater than some integer t, the main problem is to construct such set of decoders (for authorized users) that for a given unauthorized decoder (pirate version), the distributor will be able to identify at least one of the sources of the leakage even if this unauthorized copy was produced by a coalition.

The problem of data protection against such collusion attacks has given rise to the well known concept of *tracing traitors* (TT) [44]. As a base of TT-schemes, in [44] it was proposed to use different types of perfect secret sharing schemes (SSS, for short), which were discovered in [54, 55]. For the moment three main tracing traitor schemes are known. Historically the first scheme is known as *codes with the identifiable parent property (IPP codes).* Such scheme is based on the simplest (n, n)-threshold SSS and was proposed in [44], then, it was further developed in [45, 49–51]; interested reader may address to the detailed overviews [52, 95, 96]. Another known scheme, based on arbitrary (w, n)-threshold SSS, was proposed in [46, 47] and is known under the name of *set systems with the identifiable parent property (IPP set systems).* The most recent results can be found in [56–58, 97, 98]. The generalization of these two schemes was proposed in [60]. It is also based on (w, n)-threshold SSS as for IPP set systems but uses an encryption process similar to one used for IPP codes. In this section we shall call this generalized scheme as *non-binary IPP set systems.*

In this section we investigate the particular case of IPP-type schemes, known as *tracing traitors schemes with traceability property* or *traceability schemes*, for short. The main idea of traceability schemes is to create such set of decoders that a malicious user (participant of the coalition) can be found as the "nearest" decoder to the forged one. In fact, the first tracing traitors schemes constructed in [44] have the traceability property, namely, the malicious users can be recovered as the nearest in Hamming metric codevector to the forged vector (decoder). They were further studied in [99, 100]. The systematic study of traceability set systems has been started in [47, 101]. An original approach to construction of traceability set systems via constant-weight codes was proposed in [102]. Unfortunately there were some mistakes in evaluation of error-correcting codes parameters, which led to wrong results as it was remarked in [103]. The correct version of constructing traceability set systems via *binary* constant-weight codes was given in [56].

The non-binary IPP set systems with traceability constitutes the subject matter of this section. Our main result is the existence of such schemes with non-vanishing rate. In thi section we provide a short reminder of the basics of non-binary IPP set systems, namely, we show how (w, n)-SSS is incorporated in it and explain the traceability paradigm for such scheme. Then, we prove GV-bound for non-binary IPP set systems with traceability and, finally, we define the effective rate of IPP-schemes what allows to compare different schemes with traceability property. As a concluding remark for the section we formulate the related open problem and explain the connection with signature codes.

Non-binary IPP set systems

Consider the following broadcasting scenario where the distributor delivers some digital content x to M users. In order to prevent illegal redistribution, the distributor transmits the content x in an encrypted form $z = \varphi(x, \sigma)$ obtained by using some secret key $\sigma \in K$, which serves as a session key and should be changed for distributing another portion of digital content. Firstly the key σ is matched with the set of shares $s_1, \ldots, s_n$ according to perfect (w, n)-threshold Secret Sharing Scheme [54, 55]. Let us recall that a secret sharing scheme is called a *perfect* (w, n)*-threshold* secret sharing scheme if any w shares out of n are enough to recover the secret σ and any less number of shares provides no a posteriori information about the secret. Initially, in [44] authors proposed to use perfect (n, n)-SSS and then encrypt each share on q different keys. Different shares are encrypted on different sets consisting of q keys, i.e., overall there are nq encrypted shares and q encrypted versions of each share. This idea gave rise to a notion of IPP codes [45]. Then, general case of w-out-of-n threshold perfect SSS was used in [46, 47] for constructing IPP set systems. For such model each share is encrypted using only a single key. Different shares are encrypted on different keys, so, overall, there are n encrypted shares.

In [60] it was proposed to combine the main ideas of these two schemes. More

precisely, it was proposed to use w-out-of-n threshold perfect SSS and encrypt each share on q different keys as it was done in [44]. Formally, the share s_i is encrypted q times on the keys from the set $\mathcal{A}_i = \{\alpha_i^1, ..., \alpha_i^q\}$. Encrypted version of shares are transmitted along with the encrypted content z. During the initial stage (before the transmission) the j-th user receives the set consisting of w decryption keys that are then used to decrypt w shares and, so, to decrypt the secret key σ (according to the chosen SSS). Formally, j-th user receives the subset $\mathcal{D}_j \subset \bigcup_{i\in[n]} \mathcal{A}_i$ consisting of w different keys needed to decrypt w different shares, i.e., $|\mathcal{D}_j| = w$ and $|\mathcal{D}_j \cap \mathcal{A}_i| \leq 1$ for all $i \in [n]$.

In what follows we will move from subset representation of users' decoders to vector representation. Indeed, consider some ordering of keys for each set $\mathcal{A}_i$ and map each key to a symbol of q-ary alphabet $\mathbf{A}_q^* = \{1, 2, ..., q\}$, for example by mapping α_i^k to $k \in \mathbf{A}_q^*$ for all $i \in [n]$. Define also the $(q+1)$-ary alphabet $\mathbf{A}_q = \{0, 1, ..., q\}$. Then, instead of considering the subset $\mathcal{D}_j$ we will consider the corresponding characteristic vector $\mathbf{c}^{(j)} \in \mathbf{A}_q^n$ such that its i-th coordinate $c_i^{(j)} = k$ if $\alpha_i^k \in \mathcal{D}_j$ and $c_i^{(j)} = 0$ if $\mathcal{D}_j \cap \mathcal{A}_i = \emptyset$ (absence of the key for i-th share). Note that the resulting vector $\mathbf{c}^{(j)}$ has exactly w non-zero coordinates, i.e., it has weight $wt(\mathbf{c}^{(j)}) = w$ over $q+1$-ary alphabet $\mathbf{A}_q$.

For this model the collusion attack proceeds in the following way. A malicious coalition $U \subset \mathbf{A}_q^n$ in order to create a working forged "decoder" has to collect at least w different keys that can decrypt w different shares. The participant of the coalition can do so by taking at least w different keys among those keys that belong to them. Thus, the set of all forged decoders that the coalition U can create equals to

$$\langle U \rangle_w = \{\mathbf{y} \in P_1^*(U) \times ... \times P_n^*(U) \ : \ wt(\mathbf{y}) \geq w\}, \tag{4.2.1}$$

where

$$P_i^*(U) = \{u_i \ : \ \mathbf{u} \in U\} \cup \{0\} \tag{4.2.2}$$

is the i-th "projection" of the coalition U. Informally, it means that the participants of the coalition can take one of the keys among those that they have for any given share. If no one has a key for a particular share, then we assume that they cannot guess the possible key.

Set systems with identifiable parent property

Now we are ready to formulate the identifiable parent property for such scheme.

Definition 4.2.1. [60] A $(q+1)$-ary constant-weight code $C \subset \mathbf{A}_q^n$ of weight w is (t, w, q)-IPP code if for any vector $\mathbf{y} \in \mathbf{A}_q^n$ s.t. $wt(\mathbf{y}) \geq w$ either

$$\bigcap_{U\subset C:\, \mathbf{y}\in\langle U\rangle_w,\, |U|\leq t} U \neq \emptyset, \tag{4.2.3}$$

or there is no $U \subset C$ such that $|U| \leq t$ and $\mathbf{y} \in \langle U \rangle_w$.

Such property guarantees that at least one malicious user will be identified correctly. Note that if $w = n$ then the definition 4.2.1 transforms to a definition of t-IPP codes [45], and for the case $q = 1$ it transforms to (t, w)-IPP set systems [46].

Set systems with traceability property

In order to formulate the traceability concept for the new type of tracing traitors schemes, i.e., q-ary IPP set systems, we need the following "proximity measure" $S(\mathbf{x}, \mathbf{y})$ between two vectors $\mathbf{x}, \mathbf{y} \in \mathbf{A}_q^n$ defined as

$$S(\mathbf{x}, \mathbf{y}) = |\{i \mid \mathbf{x}(i) = \mathbf{y}(i) \neq 0\}|, \tag{4.2.4}$$

i.e., $S(\mathbf{x}, \mathbf{y})$ is the number of coinciding *non-zero* coordinates. The function $S(\mathbf{x}, \mathbf{y})$ is obviously related to the Hamming distance $d_H(\mathbf{x}, \mathbf{y})$, namely,

$$d_H(\mathbf{x}, \mathbf{y}) = wt(\mathbf{x}) + wt(\mathbf{y}) - 2S(\mathbf{x}, \mathbf{y}) - J(\mathbf{x}, \mathbf{y}), \tag{4.2.5}$$

where $J(\mathbf{x}, \mathbf{y}) = \{l \ : \ x_l \neq 0, y_l \neq 0, x_l \neq y_l\}$.

The traceability property can be formulated as follows.

Definition 4.2.2. A $(q+1)$-ary constant weight code $C \subset \mathbf{A}_q^n$ of weight w is called a (t, w, q)-traceability set system ((t, w, q)-TSS code, for short) if for any coalition $U \subset C$, $|U| \leq t$ and any $\mathbf{y} \in \langle U \rangle_w$, it holds

$$S(\mathbf{y}, \mathbf{v}) < \max_{\mathbf{u} \in U} S(\mathbf{y}, \mathbf{u}) \tag{4.2.6}$$

for any $\mathbf{v} \in C \setminus U$.

Remark 4.2.1. Note that for the case $w = n$ this definition is equivalent to the definition of t-IPP codes with the traceability property. For the case $q = 1$ this definition is equivalent to the definition of t-IPP set systems with traceability property. In the general case the given definition is more convenient than a similar one based on the Hamming distance as we can see from the next lemma.

The following lemma establishes a sufficient condition on a (t, w, q)-set system to have t-traceability property, which is similar to the original approach of [44]:

Lemma 4.2.1. *A $(q+1)$-ary constant-weight code $C \subset \mathbf{A}_q^n$ of weight w is a (t, w, q)-TSS code if for any $\mathbf{u}, \mathbf{v} \in C$ it holds*

$$S(\mathbf{u}, \mathbf{v}) < w/t^2. \tag{4.2.7}$$

Proof. Consider any coalition $U \subset C$, $|U| \leq t$ and any $\mathbf{y} \in \langle U \rangle_w$. Then, $\max_{\mathbf{u} \in U} S(\mathbf{u}, \mathbf{y}) \geq w/t$ since $wt(\mathbf{y}) \geq w$. On the other hand, for any $\mathbf{v} \in C \setminus U$,

$$S(\mathbf{v}, \mathbf{y}) < \sum_{u \in U} S(\mathbf{v}, \mathbf{u}) < t \cdot \frac{w}{t^2} = \frac{w}{t},$$

which concludes the proof. □

According to Remark 1, Lemma 1 gives for IPP codes the same results as in [44], namely, a q-ary code C with the minimal code distance $d_H(C) > (1-t^{-2})n$ has the t-traceability property. As for t-IPP set systems, Lemma 1 coincides with Lemma 61 from [101].

Let $M_q(n, t, w)$ denote the maximal possible cardinality of (t, w, q)-TSS code of length n. Define the lower asymptotic bound on the rate of best (t, w, q)-TSS code as

$$R_t(\omega, q) = \liminf_{n \to \infty} n^{-1} M_q(n, t, \lfloor n\omega \rfloor). \tag{4.2.8}$$

We will be interested in the maximal possible rate of q-ary t-IPP set systems with traceability as

$$R_t(q) = \max_{\omega} R_t(\omega, q). \tag{4.2.9}$$

In the next section we will establish the Gilbert-Varshamov type bound on the size of (t, w, q)-TSS codes.

Gilbert-Varshamov bound for non-binary IPP set systems

Let $L_q(n, w, T)$ denote the maximum possible number of codewords in a $(q+1)$-ary code C of length n and constant weight w with $S(\mathbf{u}, \mathbf{v}) < T$ for any $\mathbf{u}, \mathbf{v} \in C$. To establish the lower bound for $L_q(n, w, T)$ we employ Gilbert-Varshamov type bound similar to GV-bound for constant weight codes.

Define the "ball" $B_z(n, w, T)$ of radius T with the center at $\mathbf{z}$ as the set of all vectors $\mathbf{x}$ of weight w such that $S(\mathbf{x}, \mathbf{z}) \geq T$. Let us denote the "size" of the ball as $B(n, w, T)$ since it is the same for all $\mathbf{z}$ s.t. $wt(\mathbf{z}) = w$. It is easy to see that

$$B(n, w, T) = \sum_{s,u: s \geq T, s+u \leq w} \binom{w}{s} \binom{w-s}{u} \binom{n-w}{w-(s+u)} (q-1)^u q^{w-(s+u)}, \tag{4.2.10}$$

where $s = S(\mathbf{x}, \mathbf{z})$ and $u = |\{l : x_l \neq 0, z_l \neq 0, x_l \neq z_l\}|$. The standard Gilbert-type arguments show that

$$L_q(n, w, T) \geq \frac{\binom{n}{w} q^w}{B(n, w, T)}, \tag{4.2.11}$$

From Lemma 1 and the equation (4.2.11) we have the following theorem

Theorem 4.2.1.

$$M_q(n,t,w) \geq \frac{\binom{n}{w} q^w}{B(n,w,wt^{-2})}. \tag{4.2.12}$$

We shall use the following simple upper bound on the size of the ball $B_z(n,w,T)$

$$B(n,w,T) \leq n^2 \max_{s,u:\ s\geq T, s+u\leq w} \left[\binom{w}{s} \binom{w-s}{u} \binom{n-w}{w-(s+u)} (q-1)^u q^{w-s-u} \right] \tag{4.2.13}$$

and the well known approximation of binomial coefficient

$$\binom{n}{k} = 2^{n(H(k/n)+o(1))} \text{ for } k \leq n/2,$$

where $H(x) = -(x\log_2 x + (1-x)\log_2(1-x))$ is the binary entropy function. Then from (4.2.12), by substituting $w = \omega n, s = yw, u = zw$, next corollary follows:

Corollary 4.2.1.

$$R_t(\omega, q) \geq H(\omega) - \max_{y,z:\ y\geq t^{-2}, y+z\leq 1, z\geq 0} F_q(\omega, y, z), \tag{4.2.14}$$

where

$$F_q(\omega,y,z) = \omega H(y) + \omega(1-y) H\left(\frac{z}{1-y}\right) + (1-\omega) H\left(\frac{\omega(1-y-z)}{1-\omega}\right) +$$

$$+\omega z \log_2(q-1) - \omega(y+z)\log_2 q. \tag{4.2.15}$$

Remark 4.2.2. It is easy to see from (4.2.5), (4.2.10) and (4.2.11) that in the case of t-IPP codes, which corresponds to $w = n$, $s + u = w$, the GV-type bound (4.2.12) coincides with the result of [44]. In the case $q = 1$ we have t-IPP set systems and the bound (4.2.14) was obtained in [56].

For the next simple case $q = 2$ the optimization problem (4.2.14) transforms to

$$R_t(2) = \max_{\omega} \min_{y,z} H(\omega) + \omega(y+z) -$$

$$- \left(\omega H(y) + \omega(1-y) H\left(\frac{z}{1-y}\right) + (1-\omega) H\left(\frac{\omega(1-y-z)}{1-\omega}\right) \right) \tag{4.2.16}$$

subject to $\omega, z \geq 0$, $y \geq t^{-2}, y + z \leq 1$, and t is integer greater than 1. The corresponding numerical optimization gives that for $t = 2$

$$R_2(2) \geq 0.03602,$$

which is achieved for $\omega = 0.1156$, i.e. for $w/n = 0.1156$, and for $t = 3$

$$R_3(2) \geq 0.006314,$$

which is achieved for $\omega = 0.048$.

Consider also the case $q = 3$. The corresponding numerical optimization gives that for $t = 2$

$$R_2(3) \geq 0.05369,$$

which is achieved for $\omega = 0.172$, and for $t = 3$

$$R_3(3) \geq 0.00946,$$

which is achieved for $\omega = 0.073$. Note, that numerical results for the case $q = 1$, i.e., the case of t-IPP set systems, can be found in [56, 98].

How to compare tracing traitors schemes?

In order to compare different tracing traitors schemes we need to return to the origin of this subject, namely to [44], where it was suggested to consider the total number $N = nq$ of transmitted "blocks" containing encrypted shares, i.e., consider N as the "block length" and correspondingly calculate the *effective* rate of (t, w, q)-TSS code C as

$$R^{\text{eff}} = N^{-1} \log_2 |C|.$$

In the case of IPP set systems ($q = 1$) the effective rate equals to the ordinary code rate, since $q = 1$ and $N = n$.

Define the *maximal possible effective rate* of (t, w, q)-TSS codes as

$$R_t^{\text{eff}} = \max_q R_t^{\text{eff}}(q),$$

where $R_t^{\text{eff}}(q) = q^{-1} R_t(q)$.

Let us compare numerically the new traceability scheme with the known ones in the particular case of coalitions of size two and three. For $t = 2$ and $q = 1$ in [56] it was proved that $R_2^{\text{eff}}(q = 1) = 0.0181$, this bound was later improved in [98] using combinatorial methods, and the best known bound for today is $R_2^{\text{eff}}(q = 1) = 0.0219$. For the case $t = 3$ from [98] we have $R_3^{\text{eff}}(q = 1) = 0.00365$. As for the new scheme from (4.2.16) we have $R_2^{\text{eff}}(q = 2) = 0.018, R_2^{\text{eff}}(q = 3) = 0.0179$, it can be shown that $R_2^{\text{eff}}(q)$ decreases with the growth of q. If we consider 2-IPP traceability codes ($w = n$) the corresponding effective rate achieves its maximum at $q = 18$ and is equal to 0.0162, and for the case $t = 3$ the maximum is at $q = 43$ and is equal to 0.00301. So, we can conclude that for now the best effective rate R_t^{eff} for $t = 2$ is achieved at $q = 1$ and is equal to 0.0219 which is due to binary 2-IPP set systems with traceability

property [98]. The same can be said for the case $t = 3$, the best effective rate is also due to binary 3-IPP set systems with traceability and is equal to 0.00365.

Connection to MACs and signature codes In this section we introduced generalized IPP-schemes with the traceability property that allow to investigate uniformly t-IPP codes and t-IPP set systems with the traceability property as two marginal cases of non-binary IPP set systems.

How the effective rate of the best general t-IPP schemes with traceability behaves for $t \to \infty$ is still an open question. It is known that the effective rate of t-IPP set systems with traceability $R_t^{\text{eff}}(q = 1) = t^{-4+o(1)}$. Indeed, it was proved in [57] that t-traceability set systems is a t^2-cover-free family [61], therefore, it follows from the known upper bound on the cardinality of t-cover-free families, see [61], [32], that $R_t(1) = O(t^{-4+o(1)})$. On the other hand, the GV-bound shows that $R_t(1) \geq c_1 t^{-4}$, where $c_1 > 0$ is some constant.

We conjecture that for large t

$$R_t^{\text{eff}} = t^{-4+o(1)}.$$

As for the connection to MAC, the described problem also can be considered as information transmission problem. Users forming a coalition and producing a forged vector can be considered as active users of A-channel which output is controlled by a malicious opponent. In this model the receiver (decoder) sees only one element (from the input alphabet) among all elements used by a coalition, and this element is chosen by the opponent in a way preventing the dealer from finding the coalition or even a single element of it.

4.3 Symmetric group testing

Let us start from a short history of the research about group testing (GT) problem. Then we investigate a particular case of GT, namely, symmetric group testing and finally, establish the connection with signature codes for A-channel.

Group testing is a combinatorial scheme developed for the purpose of efficient identification of defective elements in a given pool of subjects. The naive solution of the search of defective elements is to test each item separately, but group testing allows to conduct tests in more efficient way. The main idea is to test the samples in groups (subsets), rather than individually, which decreases the number of conducted tests.

The history of this problem starts with the work of Dorfman [14], where he formulated the problem in the context of the blood tests for the presence of the particular disease. In this case, blood samples of different persons were mixed and

then tested. If at least one of the blood samples used in this test was "defective" then the answer was "yes". If all blood samples were "good" then the answer was "no". There are many applications of group testing in different areas of science, in particular, in computational molecular biology. For more detailed review of group testing applications see [36].

There are three main points of difference of group testing schemes.

- The first one is the strategy of the search. There are two possible cases, namely, adaptive and non-adaptive search. For the adaptive search questions/test are made in series in dependence of the answers for previous questions. For non-adaptive case all tests are conducted simultaneously, and based on all answers one decides about the set of defective elements.

- The second difference for group testing models is the answer-question model. For example, one can think of tests where each sample can participate only in a limited number of tests, and the number of samples in one test is also upper bounded; or one can think of threshold schemes, where one receives the answers "yes" only if the number of defective elements is bigger than some predefined amount, see [37].

- The third difference is presence or absence of noise. Noiseless case are mainly considered in the literature. And a noisy case could be with random nature of errors or an adversarial errors.

Group testing with noise There are two main models of noise for group testing. The first one is probabilistic, where errors are generated according to some probability distribution, see e.g. [38]. The second model, known as adversarial noise, has a combinatorial nature and it is exactly the type of errors that we consider in this section.

Probably the most famous problem of group testing with adversarial noise is the so-called Ulam's problem on searching with a lie. Ulam asked in his book [39] what is the minimal number of yes-no queries needed to find an unknown integer between 1 and $N = 10^6$ if one lie is allowed among answers (lie is equivalent to an error). In fact, this problem was first stated by A.Renyi in [40], so it is more correctly to call Renyi-Ulam problem.
The exact answer for adaptive search algorithms and arbitrary N was given by A.Pelc in [41], see also his review paper [42]. The corresponding asymptotic result is known for general case of L false answers, namely, for fixed L and growing N the minimal number of queries behaves asymptotically as $\log_2 N + L \log_2 \log_2 N$ and it can be achieved by non-adaptive search.
The Renyi-Ulam problem is about finding a single defective element in presence

of L erroneous tests results. A generalization of this problem to the case of many defective elements and its relation to error-correcting codes was considered in [104], [105].

Symmetric group testing In this section we consider the modification of the ordinary group testing problem, namely, symmetric group testing (SGT) and consider the case of the presence of adversarial noise. Informally, for a given tested subset $\mathcal{F} \subset X$ of the ground set X the response of SGT scheme equals 0 iff no defective elements belong to $\mathcal{F}$, equals 1 iff all elements of $\mathcal{F}$ are defective, and equals $\{0, 1\}$ otherwise. Note that answers of SGT scheme provide more information than the ordinary group testing. Namely, SGT allows to distinguish the cases when a tested group consists of only defective elements and when it consists of both defective and good elements.

The use of SGT was originally motivated by applications in circuit testing and chemical component analysis [43], see also [106]. As an example, consider the testing of N identically designed circuits using only serial and parallel component concatenation. In the serial testing mode, one can detect if all circuits are operational. In the parallel mode, one can detect if all circuits are non-operational. If at least one circuit is operational and one is non-operational, neither of the two concatenation schemes will be operational. Detecting efficiently which of the circuits are non-operational is exactly what symmetric group testing is aimed to.

More formally, consider the set $X = [N]$ of all samples and let $\mathcal{F} \subset X$ be a tested subset. In SGT the response on a test $\mathcal{F}$ equals 0 iff no defective elements belong to $\mathcal{F}$, equals 1 iff all elements of $\mathcal{F}$ are defective, and equals $\{0, 1\}$ otherwise. The goal is to create such family $\{\mathcal{F}_1, ..., \mathcal{F}_n\}$ of subsets (tests) of X of minimal size n that the answers for such tests allow to uniquely identify the subset of defective elements, given that their number is upper bounded by some fixed parameter d. It is convenient to consider binary characteristic vectors of the sets, i.e., we map each test $\mathcal{F} \subset X$ to the binary vector $\mathbf{f} \in \{0, 1\}^N$, where $f_i = 1$ if $i \in \mathcal{F}$ and $f_i = 0$ otherwise. Since we consider the non-adaptive version of symmetric group testing we can represent the family of tests in a form of $n \times N$ matrix H where rows represent tests and, consequently, columns $\{h^1, ..., h^N\} \subset \{0, 1\}^n$ represent "identifying" vectors for each sample from a pool. If an element $h^i_j = 1$, $i \in [N], j \in [n]$, it means that i-th sample from the pool participates in j-th test. The answer for such set of tests can be represented as a vector $a \in \{\{0\}, \{1\}, \{0, 1\}\}^n$, where a_j is the answer for the j-th test. Then, the goal of SGT is to construct such matrix that gives for different subsets of defective elements the different answer vectors $\mathbf{a}$ which is formally stated in the definition 4.3.1.

Let $\text{supp}(x)$ denote the set of positions where a vector x has non-zero values,

i.e., $\mathrm{supp}(x) = \{i \mid x_i \neq 0\}$. And let $< x, y >_A$ be an A-scalar product, i.e.,

$$< x, y >_A = \begin{cases} \{0\}, \text{ if } \mathrm{supp}(x) \cap \mathrm{supp}(y) = \emptyset, \\ \{1\}, \text{ if } |\mathrm{supp}(x) \cap \mathrm{supp}(y)| = |\mathrm{supp}(x)|, \\ \{0,1\}, \text{ otherwise.} \end{cases}$$

Consider a binary matrix H of size $n \times N$ with rows $\{f_1, ..., f_n\} \subset \{0,1\}^N$ and some binary vector y of length N, then the product $H \cdot_A y$ is defined as

$$H \cdot_A y := (< f_1, y_1 >_A, < f_2, y_1 >_A, ..., < f_n, y_1 >_A).$$

(characteristic vectors of the tests). So, if H is a search matrix, then such A-scalar products represent the answer for the tests (rows of H) if y is a characteristic vector of a set of defective elements.

Definition 4.3.1. A matrix H of size $n \times N$ with columns $\{h^1, ..., h^N\} \subset \{0,1\}^n$ is called t-SGT matrix, if for any two vectors vectors $y_1, y_2 \in \{0,1\}^N, y_1 \neq y_2$ (vectors of defective elements) such that $|\mathrm{supp}(y_i)| \leq t, i = 1, 2$, the products $H \cdot_A y_1, H \cdot_A y_2$ are also different.

Correspondingly, the set of columns of t-SGT matrix is called a *t-SGT code*. It is not difficult to see that the definition of the t-SGT codes coincides with the definition of t-signature codes for A-channel. So, all the results received in the Chapter 1 are also valid for t-SGT codes, including the lower and upper bounds on the rate, construction of corresponding codes with an efficient decoding algorithm. Also, the noisy case of SGT can be considered in the same way as it was done for t-signature codes for A-channel.

We assume that the output vector (vector of answers to the tests) might be erroneous in no more than L positions, i.e., no more than L answers are incorrect. The goal of SGT in the presence of noise is for the given answer vector to recover the set of defective elements even if some of the answers are incorrect. Formally it can be stated as follows:

Definition 4.3.2. A t-SGT is said to correct up to L errors, or (t, L)-SGT code for short, if for any $y_1, y_2 \in \{0,1\}^N$ such that $|\mathrm{supp}(y_i)| \leq t, i = 1, 2$ and $y_1 \neq y_2$ the equation

$$H \cdot_A y_1 + e_1 = H \cdot_A y_2 + e_2, |\mathrm{supp}(e_i)| \leq L, i = 1, 2$$

implies $y_1 = y_2$.

So, the estimation of upper and lower bound for the t-signature codes for A-channel that can correct L errors can be applied to (t, L)-SGT code.

Conclusion

In the thesis we derived some new results on signature codes for three classes of multiple access channels, namely, A-channel, B-channel and weighted binary adder channel, and we developed the uniform approach which allows us to apply these results to digital fingerprinting codes and for non-adaptive symmetric group testing. As for the open problems that still remain unsolved for the considered types of MACs we can name the following questions.

The first one concerns the constructions and decoding procedures for t-signature codes of the considered MAC. In chapter 2 it was explained how to construct such codes for the A-channel, however, the rate of proposed codes is lower than the optimal one. So, the question of attaining the higher code rate is open. As for the B-channel, no constructions are known for q-ary case, although the B-channel provides the most information about the inputs. As for the weighted binary adder channel, the lower bound proved in this thesis was obtained in a constructive way, i.e., by exploiting the existent class of codes, namely BCH codes. However, the decoding procedure remains unclear since for the weighted channel we have to recover the input vectors and the corresponding weights that were used for the transmission. This problem is different from the decoding of BCH codes which demands algebra over finite fields and not real numbers.

The second open questions concern the noisy case. It would be interesting to investigate the effect of the adversarial noise on the rate of t-signature codes for B-channel, in a way similar to what was done for A-channel. Also, explicit constructions of such codes that can be resistant to noise stay as the open question.

As the last direction of the open questions, it might be interesting to consider the recovery of only a subset of active users of the channel. The problem is inspired by the DRM and multimedia digital fingerprinting application where the distributor wants to find at least one user from the malicious coalition, i.e., the recovery of the whole coalition is not necessary. It was shown in the second chapter that for the case of A-channel it can be done by employing the so-called single user tracing and k-out-of-t-user tracing codes. Still, the analogous questions for B-channel and weighted binary adder channel remains open.

Bibliography

[1] Claude Shannon. Two-way communication channels. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[2] W Kautz and Roy Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

[3] EE Egorova. On multimedia digital fingerprinting codes. In *Proceedings of ITaS 2015*, pages 1199–1204, 2015.

[4] Minquan Cheng and Ying Miao. On anti-collusion codes and detection algorithms for multimedia fingerprinting. *IEEE transactions on information theory*, 57(7):4843–4851, 2011.

[5] Elena Egorova, Marcel Fernandez, Grigory Kabatiansky, and Moon Ho Lee. Signature codes for the a-channel and collusion-secure multimedia fingerprinting codes. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 3043–3047. IEEE, 2016.

[6] Elena Egorova and Valeriya Potapova. Signature codes for a special class of multiple access channel. In *2016 XV International Symposium Problems of Redundancy in Information and Control Systems (REDUNDANCY)*, pages 38–42. IEEE, 2016.

[7] EE Egorova and VS Potapova. Compositional restricted multiple access channel. *Problems of Information Transmission*, 54(2):116–123, 2018.

[8] Rudolf Ahlswede. Multi-way communication channels. In *Second International Symposium on Information Theory: Tsahkadsor, Armenia, USSR, Sept. 2-8, 1971*, 1973.

[9] H Liao. A coding theorem for multiple access communications. In *Proc. Int. Symp. Information Theory, Asilomar, CA, 1972*, 1972.

[10] Or Ordentlich and Ofer Shayevitz. A vc-dimension-based outer bound on the zero-error capacity of the binary adder channel. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2366–2370. IEEE, 2015.

[11] Staffan Söderberg and Harold S Shapiro. A combinatory detection problem. *The American Mathematical Monthly*, 70(10):1066–1070, 1963.

[12] Paul Erdos and Alfréd Rényi. On two problems of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 8:229–243, 1963.

[13] Shih-Chun Chang and E Weldon. Coding for t-user multiple-access channels. *IEEE Transactions on Information Theory*, 25(6):684–691, 1979.

[14] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[15] Shih-Chun Chang and J Wolf. On the t-user m-frequency noiseless multiple-access channel with and without intensity information. *IEEE Transactions on Information Theory*, 27(1):41–48, 1981.

[16] Arthur D Friedman, Ronald L Graham, and Jeffrey D Ullman. Universal single transition time asynchronous state assignments. *IEEE Transactions on Computers*, 100(6):541–547, 1969.

[17] Yurii L'vovich Sagalovich. Separating systems. *Problemy Peredachi Informatsii*, 30(2):14–35, 1994.

[18] Gérard Cohen, Simon Litsyn, and Gilles Zémor. Binary b2-sequences: a new upper bound. *Journal of Combinatorial Theory, Series A*, 94(1):152–155, 2001.

[19] János Körner and Gábor Simonyi. Separating partition systems and locally different sequences. *SIAM journal on discrete mathematics*, 1(3):355–359, 1988.

[20] Grigory Kabatiansky, Marcel Fernandez, and Elena Egorova. Multimedia fingerprinting codes resistant against colluders and noise. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–5. IEEE, 2016.

[21] Elena Egorova. Symmetric group testing with noise. In *2019 XVI International Symposium" Problems of Redundancy in Information and Control Systems"(REDUNDANCY)*, pages 99–103. IEEE, 2019.

[22] Elena Egorova, Marcel Fernandez, Grigory Kabatiansky, and Moon Ho Lee. Signature codes for weighted noisy adder channel, multimedia fingerprinting and compressed sensing. *Designs, Codes and Cryptography*, 87(2-3):455–462, 2019.

[23] Bernt Lindström. On a combinatory detection problem i. *I. Magyar Tud. Akad. Mat. Kutató Int. Közl*, 9:195–207, 1964.

[24] David G Cantor and WH Mills. Determination of a subset from certain combinatorial properties. *Canadian Journal of Mathematics*, 18:42–48, 1966.

[25] Bernt Lindström. Determination of two vectors from the sum. *Journal of Combinatorial Theory*, 6(4):402–407, 1969.

[26] Paul Erdös and Pál Turán. On a problem of sidon in additive number theory, and on some related problems. *Journal of the London Mathematical Society*, 1(4):212–215, 1941.

[27] Bernt Lindström. On b2-sequences of vectors. *Journal of number Theory*, 4(3):261–265, 1972.

[28] Raj Chandra Bose and Sarvadaman Chowla. Theorems in the additive theory of numbers. Technical report, North Carolina State University. Dept. of Statistics, 1960.

[29] James Singer. A theorem in finite projective geometry and some applications to number theory. *Transactions of the American Mathematical Society*, 43(3):377–385, 1938.

[30] Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. On a coding model for a multiple-access adder channel. *Problemy Peredachi Informatsii*, 17(2):26–38, 1981.

[31] Paul Erdös, Peter Frankl, and Zoltán Füredi. Families of finite sets in which no set is covered by the union of two others. *Journal of Combinatorial Theory, Series A*, 33(2):158–166, 1982.

[32] Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.

[33] Miklós Ruszinkó. On the upper bound of the size of the r-cover-free families. *Journal of Combinatorial Theory, Series A*, 66(2):302–310, 1994.

[34] Thomas Ericson and Vladimir I. Levenshtein. Superimposed codes in the hamming space. *IEEE transactions on information theory*, 40(6):1882–1893, 1994.

[35] Nader H Bshouty and Hanna Mazzawi. On parity check (0, 1)-matrix over zp. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, pages 1383–1394. SIAM, 2011.

[36] Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 2000.

[37] Thach V Bui, Minoru Kuribayashi, Mahdi Cheraghchi, and Isao Echizen. Efficiently decodable non-adaptive threshold group testing. *IEEE Transactions on Information Theory*, 2019.

[38] George K Atia and Venkatesh Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Transactions on Information Theory*, 58(3):1880–1901, 2012.

[39] Stanislaw M Ulam. *Adventures of a Mathematician*. Univ of California Press, 1991.

[40] Alfréd Rényi. On a problem of information theory. *MTA Mat. Kut. Int. Kozl. B*, 6:505–516, 1961.

[41] Andrzej Pelc. Solution of ulam's problem on searching with a lie. *Journal of Combinatorial Theory, Series A*, 44(1):129–140, 1987.

[42] Andrzej Pelc. Searching games with errors—fifty years of coping with liars. *Theoretical Computer Science*, 270(1-2):71–109, 2002.

[43] Milton Sobel, Satindar Kumar, and Saul Blumenthal. Symmetric binomial group-testing with three outcomes. In *Statistical Decision Theory and Related Topics*, pages 119–160. Elsevier, 1971.

[44] Benny Chor, Amos Fiat, and Moni Naor. Tracing traitors. In *Annual International Cryptology Conference*, pages 257–270. Springer, 1994.

[45] Henk DL Hollmann, Jack H Van Lint, Jean-Paul Linnartz, and Ludo MGM Tolhuizen. On codes with the identifiable parent property. *Journal of Combinatorial Theory, Series A*, 82(2):121–133, 1998.

[46] Michael J Collins. Upper bounds for parent-identifying set systems. *Designs, Codes and Cryptography*, 51(2):167–173, 2009.

[47] Douglas R Stinson and Ruizhong Wei. Combinatorial properties and constructions of traceability schemes and frameproof codes. *SIAM Journal on Discrete Mathematics*, 11(1):41–53, 1998.

[48] Elena Egorova. How to combine ipp codes and ipp sets systems - proof of the concept. *16th International Workshop on Algebraic and Combinatorial Coding Theory, Svetlogorsk,Russia, September 2017.*

[49] Alexander Barg, Gérard Cohen, Sylvia Encheva, Gregory Kabatiansky, and Gilles Zémor. A hypergraph approach to the identifying parent property: the case of multiple parents. *SIAM Journal on Discrete Mathematics*, 14(3):423–431, 2001.

[50] Noga Alon, Gérard Cohen, Michael Krivelevich, and Simon Litsyn. Generalized hashing and parent-identifying codes. *Journal of Combinatorial Theory, Series A*, 104(1):207–215, 2003.

[51] Jessica N Staddon, Douglas R Stinson, and Ruizhong Wei. Combinatorial properties of frameproof and traceability codes. *IEEE transactions on information theory*, 47(3):1042–1049, 2001.

[52] Simon R Blackburn. Combinatorial schemes for protecting digital content. *Surveys in combinatorics*, 307:43–78, 2003.

[53] Grigory Kabatiansky. On the tracing traitors math. In *International Conference on Codes, Cryptology, and Information Security*, pages 371–380. Springer, 2019.

[54] George Robert Blakley et al. Safeguarding cryptographic keys. In *Proceedings of the national computer conference*, volume 48, 1979.

[55] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[56] Elena Egorova and Grigory Kabatiansky. Analysis of two tracing traitor schemes via coding theory. In *International Castle Meeting on Coding Theory and Applications*, pages 84–92. Springer, 2017.

[57] Yujie Gu and Ying Miao. Bounds on traceability schemes. *IEEE Transactions on Information Theory*, 64(5):3450–3460, 2017.

[58] Elena Egorova, Marcel Fernandez, and Grigory Kabatiansky. A construction of traceability set systems with polynomial tracing algorithm. In *International Symposium on Information Theory, 2019. ISIT 2019. Proceedings.* IEEE, 2019.

[59] Elena Egorova, Marcel Fernandez, and Grigory Kabatiansky. A new class of traceability schemes. *WCC 2019: The Eleventh International Workshop on Coding and Cryptography, Saint-Jacut-de-la-Mer, France, April 2019.*

[60] Egorova E. On generalization of ipp codes and ipp set systems. *Problems of Information Transmission*, 55(3), 2019.

[61] Paul Erdös, Peter Frankl, and Zoltán Füredi. Families of finite sets in which no set is covered by the union of r others. *Israel Journal of Mathematics*, 51(1):79–89, 1985.

[62] Arkadii G D'yachkov, Il'ya Viktorovich Vorob'ev, NA Polyansky, and V Yu Shchukin. Bounds on the rate of disjunctive codes. *Problems of Information Transmission*, 50(1):27–56, 2014.

[63] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[64] Gérard D Cohen and Hans Georg Schaathun. *Asymptotic overview on separating codes.* Number 248. Department of Informatics, University of Bergen, 2003.

[65] Chong Shangguan, Xin Wang, Gennian Ge, and Ying Miao. New bounds for frameproof codes. *IEEE Transactions on Information Theory*, 63(11):7247–7252, 2017.

[66] Hugues Randriambololona. (2, 1)-separating systems beyond the probabilistic bound. *Israel Journal of Mathematics*, 195(1):171–186, 2013.

[67] Gérard Cohen, Simon Litsyn, and Gilles Zémor. Binary b2-sequences: a new upper bound. *Journal of Combinatorial Theory, Series A*, 94(1):152–155, 2001.

[68] Simon R Blackburn. Probabilistic existence results for separable codes. *IEEE Transactions on Information Theory*, 61(11):5822–5827, 2015.

[69] Miklós Csűrös and Miklós Ruszinkó. Single-user tracing and disjointly superimposed codes. *IEEE transactions on information theory*, 51(4):1606–1611, 2005.

[70] Noga Alon and Vera Asodi. Tracing a single user. *European Journal of Combinatorics*, 27(8):1227–1234, 2006.

[71] Noga Alon and Vera Asodi. Tracing many users with almost no rate penalty. *IEEE transactions on information theory*, 53(1):437–439, 2006.

[72] Alexander Barg, G Robert Blakley, and Gregory A Kabatiansky. Digital fingerprinting codes: Problem statements, constructions, identification of traitors. *IEEE Transactions on Information Theory*, 49(4):852–865, 2003.

[73] A Quang. Bounds on constant weight binary superimposed codes. *PROBLEMS CONTROL INF. THEORY.*, 17(4):223–230, 1988.

[74] Yurii L'vovich Sagalovich. Concatenated codes of automaton states. *Problemy Peredachi Informatsii*, 14(2):77–85, 1978.

[75] G David Forney. Concatenated codes. 1965.

[76] Venkatesan Guruswami and Madhu Sudan. Improved decoding of reed-solomon and algebraic-geometric codes. In *Proceedings 39th Annual Symposium on Foundations of Computer Science (Cat. No. 98CB36280)*, pages 28–37. IEEE, 1998.

[77] AG Dyachkov, VV Rykov, and AM Rashad. Superimposed distance codes. *Problems of Control and Information Theory-problemy Upravleniya i Teorii Informatsii*, 18(4):237–250, 1989.

[78] Anthony J Macula. Error-correcting nonadaptive group testing with de-disjunct matrices. *Discrete Applied Mathematics*, 80(2-3):217–222, 1997.

[79] Yongxi Cheng. Advances in group testing. *Handbook of Combinatorial Optimization*, pages 93–144, 2013.

[80] Amin Emad, Jun Shen, and Olgica Milenkovic. Symmetric group testing and superimposed codes. In *2011 IEEE Information Theory Workshop*, pages 20–24. IEEE, 2011.

[81] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[82] Necdet Batir. Inequalities for the gamma function. *Archiv der Mathematik*, 91(6):554–563, 2008.

[83] AG Djackov. On a search model of false coins. In *Topics in Information Theory (Colloquia Mathematica Societatis Janos Bolyai 16, Keszthely, Hungary). Budapest, Hungary: Hungarian Acad. Sci*, page 163170, 1975.

[84] L Gyorfi, Sándor Gyori, Bálint Laczay, and M Ruszinko. Lectures on multiple access channels. *Web: http://www.szit.bme.hu/~gyori/AFOSR_05/book.pdf*, 5, 2013.

[85] P Mateev. On the entropy of the multinomial distribution. *Theory of Probability & Its Applications*, 23(1):188–190, 1978.

[86] Yuichi Kaji. Bounds on the entropy of multinomial distribution. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1362–1366. IEEE, 2015.

[87] Vladimir Gritsenko, Grigory Kabatiansky, Vladimir Lebedev, and Alexey Maevskiy. Signature codes for noisy multiple access adder channel. *Designs, Codes and Cryptography*, 82(1-2):293–299, 2017.

[88] Peter Mathys. A class of codes for a t active users out of n multiple-access communication system. *IEEE Transactions on Information Theory*, 36(6):1206–1219, 1990.

[89] Wade Trappe, Min Wu, Z Jane Wang, and KJ Ray Liu. Anti-collusion fingerprinting for multimedia. *IEEE Transactions on Signal Processing*, 51(4):1069–1087, 2003.

[90] Minquan Cheng, Lijun Ji, and Ying Miao. Separable codes. *IEEE Transactions on Information Theory*, 58(3):1791–1803, 2011.

[91] KJR Liu, W Trappe, ZJ Wang, M Wu, and H Zhao. Multimedia fingerprinting forensics for traitor tracing, 2005.

[92] Minquan Cheng, Hung-Lin Fu, Jing Jiang, Yuan-Hsun Lo, and Ying Miao. Codes with the identifiable parent property for multimedia fingerprinting. *Designs, Codes and Cryptography*, 83(1):71–82, 2017.

[93] Jing Jiang, Minquan Cheng, and Ying Miao. Strongly separable codes. *Designs, Codes and Cryptography*, 79(2):303–318, 2016.

[94] Amos Fiat and Moni Naor. Broadcast encryption. In *Annual International Cryptology Conference*, pages 480–491. Springer, 1993.

[95] G. Kabatiansky. Traceability codes and their generalizations. *Problems of Information Transmission*, 55(3):283–294, 2019.

[96] G. Kabatiansky. On the tracing traitors math. In *International Conference on Codes, Cryptology, and Information Security*, pages 371–380. Springer, 2019.

[97] Y. Gu, M. Cheng, G. Kabatiansky, and Y. Miao. Probabilistic existence results for parent-identifying schemes. *IEEE Transactions on Information Theory*, 2019.

[98] E. Egorova and I. Vorobyev. New lower bound on the rate of traceability set systems. *XVI International Symposium Problems of Redundancy in Information and Control Systems*, pages 93–98, 2019.

[99] Tina Lindkvist, J Lofvenberg, and M Svanstrom. A class of traceability codes. *IEEE Transactions on Information Theory*, 48(7):2094–2096, 2002.

[100] Simon R Blackburn, Tuvi Etzion, and Siaw-Lynn Ng. Traceability codes. *Journal of Combinatorial Theory, Series A*, 117(8):1049–1057, 2010.

[101] Douglas R Stinson and Ruizhong Wei. Key preassigned traceability schemes for broadcast encryption. In *International Workshop on Selected Areas in Cryptography*, pages 144–156. Springer, 1998.

[102] Reihaneh Safavi-Naini and Yejing Wang. New results on frame-proof codes and traceability schemes. *IEEE Transactions on Information Theory*, 47(7):3029–3033, 2001.

[103] Jacob Lofvenberg and Jan-Åke Larsson. Comments on" new results on frame-proof codes and traceability schemes. *IEEE Transactions on Information Theory*, 56(11):5888–5889, 2010.

[104] G Kabatiansky, V Lomakov, and S Vladuts. On codes correcting errors in channel and syndrom. *Probl. Inf. Transm*, 51(2):50–57, 2015.

[105] Grigory Kabatiansky, Serge Vaduts, and Cedric Tavernier. On the doubly sparse compressed sensing problem. In *IMA International Conference on Cryptography and Coding*, pages 184–189. Springer, 2015.

[106] Frank K Hwang. Three versions of a group testing game. *SIAM Journal on Algebraic Discrete Methods*, 5(2):145–153, 1984.