2018-07-12

The expected security performance of random linear binary codes in syndrome coding

Zhang, K

10.1049/iet-com.2017.1243
IET Communications
Institution of Engineering and Technology

IET Communications

Research Article

IET Journals
The Institution of Engineering and Technology

Expected security performance of random linear binary codes in syndrome coding

ISSN 1751-8628
Received on 15th November 2017
Revised 10th March 2018
Accepted on 8th April 2018
E-First on 12th July 2018
doi: 10.1049/iet-com.2017.1243
www.ietdl.org

Ke Zhang[1] ✉, *Martin Tomlinson*[2], *Mohammed Z. Ahmed*[2], *Xiaolin Ma*[1]

[1]*Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Wuhan University of Technology, Wuhan, People's Republic of China*

[2]*University of Plymouth, UK*

✉ *E-mail: kzhang@whut.edu.cn*

Abstract: In this study, random codes are applied to the classical syndrome coding scheme to achieve secrecy of communications. By analysing the effect of the values of the columns of the parity check matrix on the resulting security level of communications, a code design method is presented which constructs a class of random codes, termed random permutation codes, which achieve high security levels and are easily generated. A theoretical analysis method is presented which determines the security level achieved by randomly chosen, linear binary codes, and compared with simulation results obtained by Monte Carlo analysis. The results verify the theoretical approach. In particular, the theoretical method is also suitable for analysis of long codes having a large number of parity check bits which are beyond evaluation by computer simulation. The results show that the security performance of any randomly chosen permutation code is close to that of the best equivocation code having the same code parameters. This has the practical advantage in syndrome coding of being able to use an ephemeral code for each communication session, thereby providing forward secrecy, a desired feature of modern, secure communication systems.

1 Introduction

The study of randomly chosen codes is a common topic in information theory starting from Shannon's random chosen code analysis in 1948 [1]. Most random code research focuses on using random codes for error correction so as to increase the reliability of digital communications [2–5]. Besides the reliability of communications, another significant research topic in present-day communications is physical layer security, modelled by the wiretap model of Wyner [6] in which the information gained by an eavesdropper, intercepting a communication link, is analysed. The secrecy of communications in this model is measured by the equivocation of the eavesdropper caused by imperfect interception. In the wiretap model, secrecy capacity is defined as the maximum rate at which a message can be reliably received by the legitimate receiver whilst communicating zero knowledge to the eavesdropper and is well studied for a large class of channels [7, 8].

Syndrome coding is an important secure coding scheme in physical layer communications. There has been considerable research aimed at the design of codes which achieve secrecy capacity for several specific wiretap channels based on the structure of Wyner's syndrome coding scheme [9, 10]. In [11], Csiszár and Körner studied the secrecy of randomly chosen codes in syndrome coding for the broadcast channel. In [12], Cohen demonstrated that a randomly chosen code in syndrome coding can ensure the security of communication by showing that the most likely syndrome does not have too high a probability of occurrence.

Recently, Chen and Vinck [13] proved as the code length tends to infinity, both reliability and security can be guaranteed, for random codes applied to syndrome coding in the model where both the main channel and the eavesdropper channel is a binary symmetric channel (BSC). Zhang *et al.* [14] presented a code design method to construct codes having the best performance, the best equivocation codes (BECs) for syndrome coding. These codes achieve the highest security level for any given set of code parameters. In this study, we assume that the legitimate receiver has an error-free communication channel whilst the eavesdropper, with an imperfect interception, experiences a BSC, with a non-zero transition probability α. This can be arranged by the nature of the communication system physical design or by providing the legitimate users with private error correction by means of hidden Goppa codes [15].

Modern, secure communication systems commonly feature a protocol known as perfect forward secrecy in which an ephemeral, encryption key is used for each communication session. The idea is to limit the effects of any security breach in the event of an adversary breaking an encryption key. In this context, a new code, unknown to the eavesdropper, needs to be used for each communication session [16]. Typically, for each session, a new code is generated by a key derivation function (KDF) starting from a secret seed known by both legitimate users. In such a system we want these codes to provide secrecy levels comparable to the BECs. The BECs themselves cannot be used because there are insufficient inequivalent codes for any given code parameters.

To achieve perfect forward secrecy, choosing codes randomly is a potential solution. However, as is shown below, some randomly chosen codes have poor performance. By analysing the code properties of these poor codes, a class of random codes known as random permutation codes (RPCs) is defined. It is shown that these codes may easily be generated from a KDF so as to provide forward secrecy and good performance.

In the following, this paper is organised as follows: Section 2 briefly reviews syndrome coding and the analysis of the equivocation for the BSC. In Section 3, we study random codes, analyse the effect of choices for the columns of the parity check matrix on the equivocation, and propose the RPC class of codes for syndrome coding. Section 4 presents both a theoretical analysis technique and a simulation methodology to determine the security level of a given (n,k) RPC employed in syndrome coding. Several examples of (n,k) RPCs are given in calculating the equivocation using both methods as confirmation of the validity of the theoretical approach. Section 5 gives simulation results for different code parameters.

2 Outline of syndrome coding

In the wiretap channel model considered here, it is arranged by system design that the main channel is error free and the eavesdropper channel is a BSC with an error probability α.

A binary (n,k) linear block code is defined by a $k \times n$ generator matrix, $\boldsymbol{G}$ or equivalently by a parity check matrix, $\boldsymbol{H}$, and is the basis of syndrome coding. The relationship between error patterns and parity check syndromes is usually provided by a syndrome look up table [12], which is assumed to be known by the sender, the legitimate receiver, and the eavesdropper. We denote the bit length of the syndrome, $n-k$ by m for brevity.

The sender, Alice, encodes a m-bit message M into a n-bit vector $\boldsymbol{X}$ as follows:

- The m-bit syndrome, S_T, is set equal to the message to be sent, M so that $S_T = M$.
- An independent, random n-bit codeword, C_T, is generated from a random k-bit vector $\boldsymbol{D}_R$ by $C_T = \boldsymbol{D}_R \times \boldsymbol{G}$.
- Based on the error-syndrome look up table, or calculated algorithmically, the n-bit error patten e corresponding to S_T is added to the codeword C_T to form the transmitted n-bit vector, $\boldsymbol{X}$ so that $\boldsymbol{X} = C_T \oplus e$. The information rate is given by $R = m/n$.

The legitimate receiver, Bob, receives the error-free output of the main channel, Y, and uses the parity check matrix of the code to determine the message as follows: $M = Y \times \boldsymbol{H}^{\mathrm{T}} = X \times \boldsymbol{H}^{\mathrm{T}} = S_{\mathrm{T}}$.

The eavesdropper, Eve, receives the bit stream of the eavesdropper BSC channel, usually containing bit errors, Z, and estimates the message, as follows $\hat{M} = S_E = Z \times \boldsymbol{H}^{\mathrm{T}} = (X+E) \times \boldsymbol{H}^{\mathrm{T}} = S_T + S_e$.

The secrecy of communications, as a function of a specific code and BSC transition probability, is traditionally measured by the equivocation of the eavesdropper decoder output [14]. This is given by

$$H(M|\hat{M}) = H(S_e) = -\sum_{S_e} p(S_e)\log_2 p(S_e), \tag{1}$$

in which, for the (n,k) code, there are 2^m distinct syndromes, S_e. The probability mass function (pmf) of the syndromes is denoted by $p(S_e)$. The equivocation can be calculated, based on the pmf of the syndromes caused by the errors from the BSC, $p(S_e)$, as a function of the parity check matrix, $\boldsymbol{H}$ of the (n,k) code.

3 Random permutation codes (RPCs)

3.1 Random codes

A random binary code may be generated by repeatedly constructing a $n \times m$ matrix with random 1s and 0s until a full rank matrix is obtained. The probability of this matrix being full rank with m independent rows is asymptotically 0.2887 [17]. The full rank matrix can be put into systematic format $\boldsymbol{H}$ by using Gauss–Jordan elimination

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 & \cdots & 0 & h_{0m} & \cdots & h_{0(n-1)} \\ 0 & 1 & \cdots & 0 & h_{1m} & \cdots & h_{1(n-1)} \\ \vdots & \vdots & \cdots & \vdots & \vdots & h_{ji} & \vdots \\ 0 & 0 & \cdots & 1 & h_{(m-1)m} & \cdots & h_{(m-1)(n-1)} \end{pmatrix}, \tag{2}$$

in which $0 \le j \le m-1$, $m \le i \le n-1$ and $h_{j,i}$ has a value of 0 or 1.

Alternatively, $\boldsymbol{H}$ may be constructed directly by appending the identity sub-matrix with random 1s and 0s for the $h_{j,i}$ values and the matrix is always full rank.

We can represent $\boldsymbol{H}$ by n packed integers as follows: $b_i = \sum_{j=0}^{m-1} h_{ji} \cdot 2^j$, in which $0 \le b_i \le 2^{n-k} - 1$. Then the parity check matrix in (2) can be represented by the following integer sequence (*Form 1*):

$$[1, 2, 4, \ldots, 2^{m-1}, b_m, \ldots, b_{n-1}]. \tag{3}$$

We can construct a random code, by choosing the values of $b_m, \ldots, b_{n-1}$ randomly between 1 and $2^m - 1$.

Random codes of this type can be put into a standard form. Since a permutation of the columns of the parity check matrix produces an equivalent code, the $b_m, \ldots, b_{n-1}$ integers may be placed in any order. We will order the parity check matrix as an identity matrix, followed by the integers that are distinct, followed by those integers that are repeats of earlier integers (*Form 2*):

$$[1, 2, 4, \ldots, 2^{m-1}, p_m, \ldots, p_{n_1}, q_{n_1+1}, \ldots, q_{n-1}], \tag{4}$$

in which $1, 2, \ldots, 2^{m-1}, p_m, \ldots, p_{n_1}$ are the distinct integers, $p_m, \ldots, p_{n_1}$ are selected from $b_m, \ldots, b_{n-1}$, and $q_{n_1+1}, \ldots, q_{n-1}$ are the repeated integers from $b_m, \ldots, b_{n-1}$, which have already appeared in $1, 2, \ldots, 2^{m-1}, p_m, \ldots, p_{n_1}$. Of course it is possible for a code to have no repeated integers in which case $n_1 = n-1$ and the q set of integers is empty. The parity check matrix in *Form 1* of a random code will be reconstructed in *Form 2*.

We next explore the effect of the repeated integers, the q set of integers, on the equivocation.

3.2 Effect of repeated columns of the parity check matrix

We can represent any error pattern as follows:

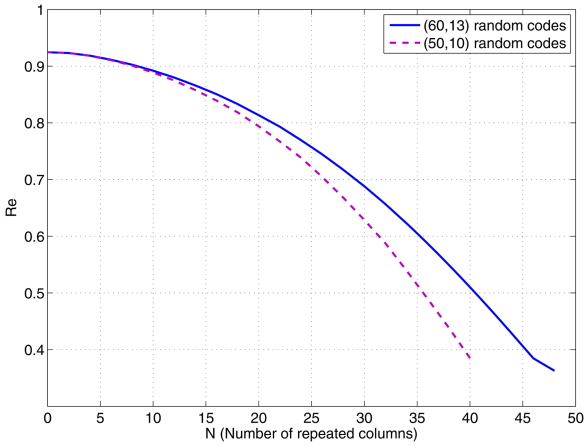

Fig. 1 *Equivocation rate versus the number of repeated columns of the parity check matrix*

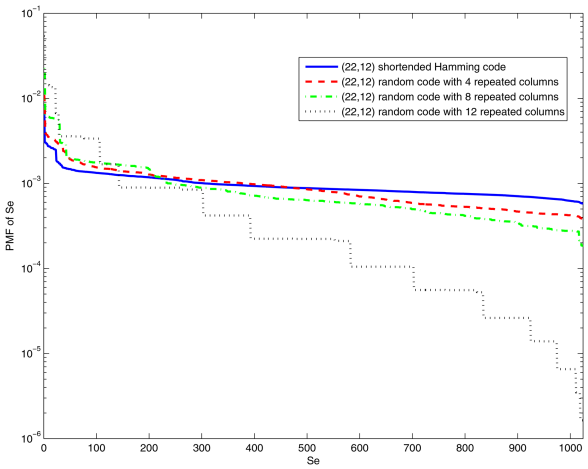

Fig. 2 *Pmf of the syndromes in rank order for different numbers of repeated columns*

$$e = [e_1\, e_2 \cdots e_n] = [e_1\, 0 \cdots 0] + [0\, e_2 \cdots 0] + \cdots + [0\, 0 \cdots e_n], \quad (5)$$

where each term denotes a 1-bit error event, $e_i = 1$ with a probability of occurring of α and $e_i = 0$ with a probability of occurring of $1 - \alpha$. As the code is a linear code, the syndrome derived from any error pattern is the modulo 2 sum of the syndromes arising from each single bit error pattern

$$S_e = e \times \boldsymbol{H}^{\mathrm{T}} = [e_1\, e_2 \cdots e_n] \times \boldsymbol{H}^{\mathrm{T}} \quad (6)$$

$$= b_1\delta(e_1 - 1) \oplus b_2\delta(e_2 - 1) \cdots \oplus b_n\delta(e_n - 1)\,. \quad (7)$$

Since the probabilities of $e_1, e_2, \ldots, e_n$ are independent, the probability of S_e is the product of the probabilities of n separate error events. From [14], there is the following theorem, replicated here for convenience

Theorem 1: The pmf of S_j for $j = 0$ to $2^m - 1$ may be defined as $p(S_j) = \beta(j)$ where $\beta(j)$ are coefficients of the probability generating function using the Z transform, denoted as $p_z(\mathcal{S})$ and this depends only on the columns of the parity check matrix and α.

$$p_z(\mathcal{S}) = \sum_{j=0}^{2^m - 1} \beta(j) Z^j = \prod_{i=0}^{n-1} ((1 - \alpha) + \alpha Z^{b_i}), \quad (8)$$

where b_i are the packed integer representations of the columns of the parity check matrix and exponent sums of powers of Z are added modulo 2.

Equation (8) shows that the choice of b_i is the main influencing factor on the pmf of S_j. From the security viewpoint, we want the pmf of S_j to be as uniform as possible since then the equivocation will be as high as possible. If $b_i = b_j$, where $0 \le i, j \le n - 1$, then $b_i \oplus b_j = 0$ and a double error event on bit positions i and j counts as no error event. Single errors in bit positions i and j will double the probability of syndrome value b_i at the expense of other syndrome probabilities. The net effect will be to cause the pmf of the syndromes to be less uniform than the case where there are no repeated columns of the parity check matrix.

The effect of repeated columns of randomly chosen codes is shown by Monte Carlo analysis of the equivocation rate where the average equivocation for a large number of random codes is determined. Fig. 1 plots the average equivocation versus the number of repeated columns of the parity check matrix for code parameters $(n = 60, m = 13)$ and $(n = 50, m = 10)$. The results clearly demonstrate that as the number of repeated columns is increased, the equivocation rate is reduced.

Based on Theorem 1, we are able to calculate the pmf of the syndromes of any code by evaluating

$$p_z(\mathcal{S}) = \prod_{i=0}^{n-1} ((1 - \alpha) + \alpha Z^{b_i})\,. \quad (9)$$

Also, by ordering the syndrome probabilities in decreasing order for each evaluated code it is easy to demonstrate the effect of the repeated columns of the parity check matrix on the syndrome pmf, on average. Fig. 2 plots the pmf of S_e for randomly, permuted, shortened Hamming codes with parameters $(22, 12)$ and also for the same parameter, random codes with different numbers of repeated columns in the parity check matrix. It is clearly evident from Fig. 2 that as the number of repeated columns increases, the pmf of the syndromes deviates further from a uniform distribution explaining why repeated column codes have poorer equivocation.

3.3 Random permutation codes

The parity check matrix of a Hamming code may be represented in a packed integer form by

$$[1, 2, 4, \ldots, 2^{m-1}, b_m, \ldots, b_{n-1}], \quad (10)$$

where $n = 2^m - 1$ and each integer is distinct.

We can construct a random code of length $2^m - 1$, by choosing the values of $b_m, \ldots, b_{n-1}$ randomly between 3 and $2^m - 1$ such that all of the integers including those that define the identity part of the parity check matrix are distinct. We call such a random code a RPC because the resulting parity check matrix is defined by a permutation of the sequence of integers from 1 to $2^m - 1$.

It is apparent that for $n = 2^m - 1$ the resulting RPC is always a permuted Hamming code and there exists some syndrome re-ordering that will produce an identical syndrome pmf. Therefore for $n = 2^m - 1$, all RPCs have the same equivocation as a Hamming code with the same parameters.

When $n < 2^m - 1$ the RPC is a shortened, permuted Hamming code but not all of these codes will be the same.

Since the repeated columns of the parity check matrix degrade the equivocation of the code, we have the following conclusions:

- RPCs are a subset of random codes and on average will have better equivocation than random codes.
- All RPCs of length $2^m - 1$ have identical equivocation and the same equivocation as a Hamming code of the same length.
- The parity check matrix of the worst performing random codes will have many repeating columns.
- In relation to cryptography, RPCs may be generated by a random permutation oracle which represents the ideal cipher in cryptography. Columns of the parity check matrix are defined by the ideal cipher.

It is apparent that any RPC may be constructed easily by generating a random code whilst taking measures to prevent repeated columns of the parity check matrix. This will produce a well performing code for syndrome coding and has a much lower computation complexity than deriving an optimum BEC [14] code. The parity check matrix of the RPC is defined as follows:

$$[p_0 = 1, p_1 = 2, \ldots, p_{m-1} = 2^{m-1}, p_m, \ldots, p_{n-1}], \quad (11)$$

in which $p_m, \ldots, p_{n-1}$ is a truncated, random permutation of the integers $3 \le i \le n - 1$ excluding $p_0, \ldots, p_{m-1}$. When $n = 2^m - 1$, the RPC will be a permuted Hamming code which is known to be optimum [14].

4 Security performance of RPCs when used in syndrome coding

There are two ways to calculate the average equivocation of the random permutation (n,k) codes

- Evaluate the average equivocation per code by simulation of a large number of error patterns generated by a BSC, then average over a large number of (n,k) RPCs using Monte Carlo analysis to calculate the average equivocation over the ensemble of RPCs.
- Analyse the effects of discrete error events on the BEC on (n,k) RPCs in terms of syndrome statistics and combine the probabilities to determine the average equivocation by theoretical analysis.

4.1 Monte Carlo simulation

We employ Monte Carlo simulation to analyse the mean of the equivocation for RPCs by generating $|C|$ ($|C|$ denotes the number of codes simulated) binary linear (n,k) RPCs, and calculating the equivocation of each code by using (1) averaged over a large number of messages. The mean of the equivocation for a large number, $|C|$, of randomly chosen (n,k) permutation codes is given by the following equation:

$$\mathbb{E}[H] = \frac{1}{|C|}\sum_{j} H_j(M|\hat{M}) \tag{12}$$

$$= \frac{1}{|C|}\sum_{j}\Big(-\sum_{S_e} p_j(S_e)\cdot \log_2 p_j(S_e)\Big), \tag{13}$$

where $p_j(S_e)$ denotes the probability of each syndrome S_e for the jth code. There are $|S_e| = 2^m$ distinct syndromes, S_e, in total for each (n,k) code.

Expanding (13), we have

$$\mathbb{E}[H] = -|S_e| \times \frac{1}{|S_e| \times |C|}\sum_{S_e}\sum_{j} p_j(S_e)\cdot \log_2 p_j(S_e). \tag{14}$$

In (14), for each code there are $|S_e| = 2^m$ values of $p_j(S_e)$ for all 2^m syndromes, and there are $|C|$ RPCs. Therefore, $|S_e| \times |C|$ values of $p(S_e)$ define a sample space S, and can be accumulated to obtain the pmf, $f_S(x)$, of $p(S_e)$, which is defined as

$$f_S(x) = Pr(\{x \in S : S = x\}) = p(p(S_e)). \tag{15}$$

Then (14) can be represented as

$$\mathbb{E}[H] = -|S_e| \times \sum_{p(S_e)} p(p(S_e)) \times p(S_e)\cdot \log_2 p(S_e), \tag{16}$$

$$= -|S_e| \times \sum_{x \in S} f_S(x) \times x \cdot \log_2 x. \tag{17}$$

Monte Carlo simulation works well for RPCs with small values of m, but for codes with $m \geq 30$ Monte Carlo simulation becomes impractical, with long computer runs. For these codes, the theoretical error event analysis method described below may be used to determine the average equivocation.

4.2 Error event-based equivocation analysis

We show below that the pmf $f_S(x)$ in (17), defined in connection with Monte Carlo analysis above, can be described in terms of a number of Poisson distributions following different weight error events occurring on the communication channel.

For a BSC with an error probability, α, there are a total of 2^m syndromes, S_e, for each evaluated RPC and whose probability is derived from a total of 2^n different error events, where each event is denoted by e. Considering the expansion of a polynomial $p(x)$ representing the independent probabilities of error events

$$\begin{aligned} p(x) &= \big((1-\alpha) + \alpha x\big)^n \\ &= (1-\alpha)^n + n(1-\alpha)^{n-1}\alpha x \end{aligned} \tag{18}$$

$$+\binom{n}{2}(1-\alpha)^{n-2}\alpha^2 x^2 + \cdots + \alpha^n x^n. \tag{19}$$

Each error event produces a resulting syndrome S_e from 2^m possible syndromes and the overall probability from all 2^n error events is $p(S_e)$. It is evident that

$$p(x = 1) = \sum p(S_e) = 1.$$

The probability of a given error event e is a function of the number of bit errors that occurred

$$p(e) = \alpha^{w(e)} \times (1-\alpha)^{n-w(e)}, \tag{20}$$

where $w(e)$ is the Hamming weight of e. The resulting syndrome is given by

$$S_e = e \times \boldsymbol{H}^{\mathrm{T}}. \tag{21}$$

Based on the weight of each error pattern, the 2^n error patterns may be divided into $n+1$ types of distinct weight error events: the 0-error event, …, the i-error events, …, and the n-error event, $0 \leq i = w(e) \leq n$. All the error patterns of the same weight have the same probability as given by (20). The code through its parity check matrix $\boldsymbol{H}$ determines the relationship between each error pattern and the resulting syndrome, as indicated by (21). As the code is linear, since there are 2^n error patterns and 2^m syndromes in total, there are exactly 2^{n-m} distinct error patterns that produce the one same syndrome in each case. These different error patterns are the result of adding, modulo 2, all of the 2^{n-m} codewords to the one same syndrome. This is because the syndrome of a codeword is zero and all codewords are distinct. Accordingly, the probability of each syndrome is determined by the summation of the probabilities of 2^{n-m} distinct error patterns.

Typical syndromes (TSs) are termed those syndromes that are not weight 0-events or weight 1-events or weight 2-events. Following from (18)

$$p(S_e) = p(S_e^0) + p(S_e^1)(l) + p(S_e^2) + p(S_e^{\mathrm{TS}}) \tag{22}$$

and the pmf of S_e is given by the convolution of the pmfs of the syndromes resulting from the separate weight error events. Since the codes are RPCs there is much that can be said about increasing weight error events and the pmf of S_e.

1. TSs are the syndromes produced by all error patterns of weight higher than 2 since these are no 'a priori' deterministic if the error event has a weight higher than 2. Weight 2-events are also not a priori deterministic except that it is known that $p(S_e = 0) = 0$, given a weight 2 error event. Since the equivocation is averaged over all randomly generated codes the probability of a given $p(S_e)$ value due to all weight i-error events will follow a Poisson distribution with parameters related to i. All of the syndromes that are not due to the 0-error event or 1-error events or 2-error events may be treated in the same way. Namely all $2^m - \binom{n}{2} - \binom{n}{1} - 1$ syndromes will have the same pmf of $p(S_e)$, $f_S(x)$.

2. The syndromes, which are generated by 1-error events, are denoted by S_e^1. Each separate single bit error event produces a distinct syndrome because each column of $\boldsymbol{H}$ is distinct. Since the syndromes in the pmf may be placed in any order each RPC produces exactly the same pmf due to single error events. There are $\binom{n}{1}$ 1-error events which map into distinct syndromes, S_e^1, and have a probability of $\alpha \times (1-\alpha)^{n-1}$ for each syndrome. We have

$$p(S_e^1) = p(S_e^{\mathrm{TS}}) + \alpha \times (1-\alpha)^{n-1}. \tag{23}$$

3. The zero syndrome, S_e^0, which is contributed by 0-error event and 3 to n error events. 0-error event contributes an additional $(1-\alpha)^n$ to $p(S_e^0)$, so

$$p(S_e^0) = p(S_e^{\mathrm{TS}-}) + (1-\alpha)^n, \tag{24}$$

where $p(S_e^{TS-})$ indicates that two-error events are excluded because the columns of the parity check matrix for RPCs are distinct and two columns can never sum to zero.

Evaluated over all of the randomly chosen codes, there will be a pmf for the distribution of $p(S_e)$, $f_S(x)$. Since each of the error events is independent, the pmf of $p(S_e)$, given by the sum of independent events, may be derived from the convolution of their pmfs as follows:

$$f_S(x) = f_S(x)_0 * f_S(x)_1 * \cdots * f_S(x)_i * \cdots * f_S(x)_n, \tag{25}$$

in which $f_S(x)_i * f_S(x)_j$ denotes the convolution between $f_S(x)_i$ and $f_S(x)_j$.

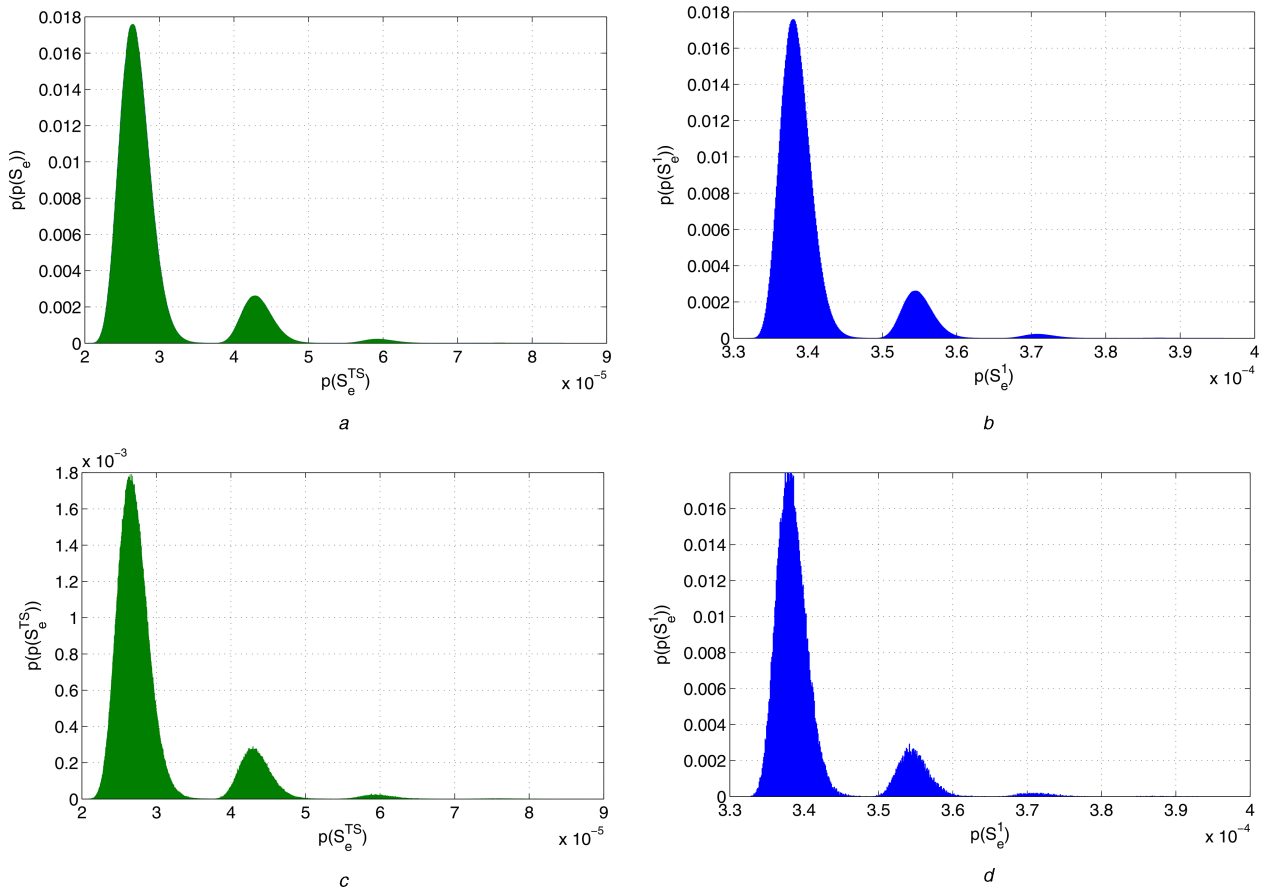

Fig. 3 *Pmfs*

(a) Pmf $f_S(x)$ for $S \in S_e^{TS}$)) obtained by event-based convolution, ***(b)*** Pmf $f_S(x)$ for $S \in S_e^1$)) obtained by event-based convolution, ***(c)*** Pmf $f_S(x)$ for $S \in S_e^{TS}$)) obtained by Monte Carlo simulation, ***(d)*** Pmf $f_S(x)$ for $S \in S_e^1$)) obtained by Monte Carlo simulation

Considering each code, there are $\binom{n}{i}$ error patterns in total for i-error events. Each i-error event maps into one syndrome out of 2^m syndromes. The average number of occurrences of an i-error event mapping into the same syndrome over $|C|$ codes is $\lambda = (|C|\binom{n}{i})/2^m$.

With the code being linear, there are 2^k error patterns producing the same syndrome and all error events are binomially distributed. The quantity of the error events over all codes is large enough that the Poisson distribution is applicable to describe the number of occurrences l that a particular syndrome is produced averaged over all i-error events

$$p(l) = \frac{\lambda^l \cdot \mathrm{e}^{-\lambda}}{l!}, \tag{26}$$

in which $p(l)$ is the pmf of l occurrences and $l!$ is the factorial of l.

For each code, one i-error event contributes the additive probability of $\alpha^i \times (1-\alpha)^{n-i}$ to the corresponding syndrome. Therefore, the pmf $f_S(x)_l$, for $x = p(S_e)$, of the syndrome arising from l occurrences, is given by

$$f_S(x)_l = p(l) = \frac{\lambda^l \cdot \mathrm{e}^{-\lambda}}{l!}, \tag{27}$$

where

$$x = p(S_e) = l \cdot \alpha^i \times (1-\alpha)^{n-i}. \tag{28}$$

It should be noted that the syndrome probabilities fall into four distinct classes and the cases of $S_e = S_e^0$, $S_e = S_e^1$, $S_e = S_e^2$ and $S_e S_e^{TS}$ are all distinct. Applying to (17) produces

$$\begin{aligned}
\mathbb{E}[H] &= -|S_e| \times \sum_{x \in S} f_S(x) \times x \cdot \log_2 x \\
&= -|S_e| \times \sum_{x \in S_e^0} f_S(x) \times x \cdot \log_2 x \\
&\quad -\binom{n}{1} \cdot |S_e| \times \sum_{x \in S_e^1} f_S(x) \times x \cdot \log_2 x \\
&\quad -\binom{n}{2} \cdot |S_e| \times \sum_{x \in S_e^2} f_S(x) \times x \cdot \log_2 x \\
&\quad -(2^m - \binom{n}{1} - \binom{n}{2} - 1 \cdot |S_e| \sum_{x \in S_e^{TS}} f_S(x) \times x \cdot \log_2 x .
\end{aligned} \tag{29}$$

For an average (n,k) RPC, the mean of the equivocation can be evaluated according to the analysis method above and compared with the results derived from the Monte Carlo method. An example of RPCs is considered next for code parameters $(n = 100, k = 85)$.

4.3 Code example

The $(10, 085)$ RPCs with $(\alpha = 0.05)$ are an example of mid-range code parameters that are amenable to both methods of average equivocation analysis. The Monte Carlo evaluation method is compared with the $f_S(x)$ pmf analysis method described above for the derivation of the average equivocation rate.

For (100,85) RPCs, where $m = 15$, there are 2^{15} syndromes in total and there are a total of $2^n = 2^{100}$ error events making exhaustive evaluation impossible. However there are only $n = 100$ different weight error events and only 100 convolutions need to be carried out to determine $f_S(x)$ given by (25). There are only four different syndrome types needed to construct $f_S(x)$

1. The zero syndrome $S_e = 0$, with $x = p(S_e = 0)$).

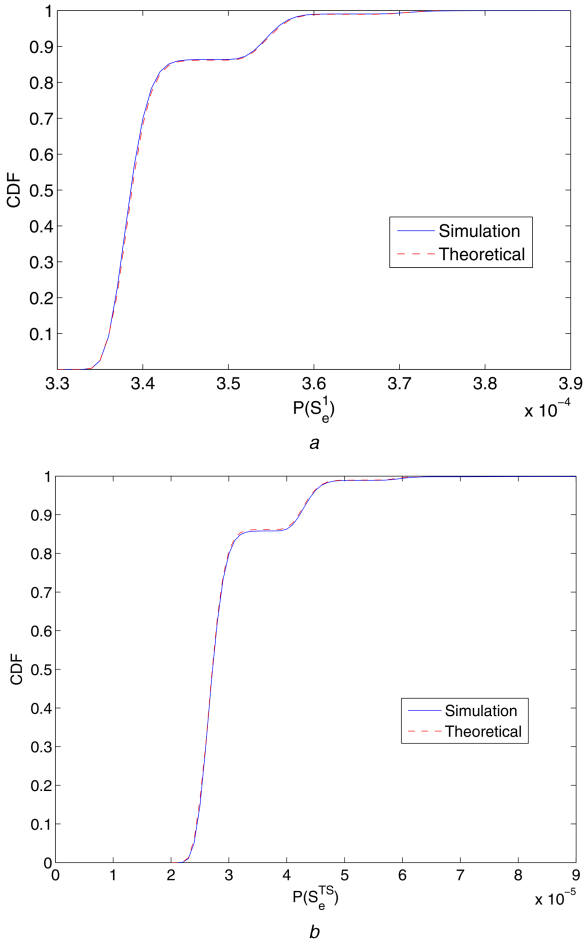

Fig. 4 *Cumulative distribution functions (CDFs) of $p(S_e^1)$ and $p(S_e^{TS})$*
(a) CDF of $p(S_e^1)$, ***(b)*** CDF of $p(S_e^{TS})$

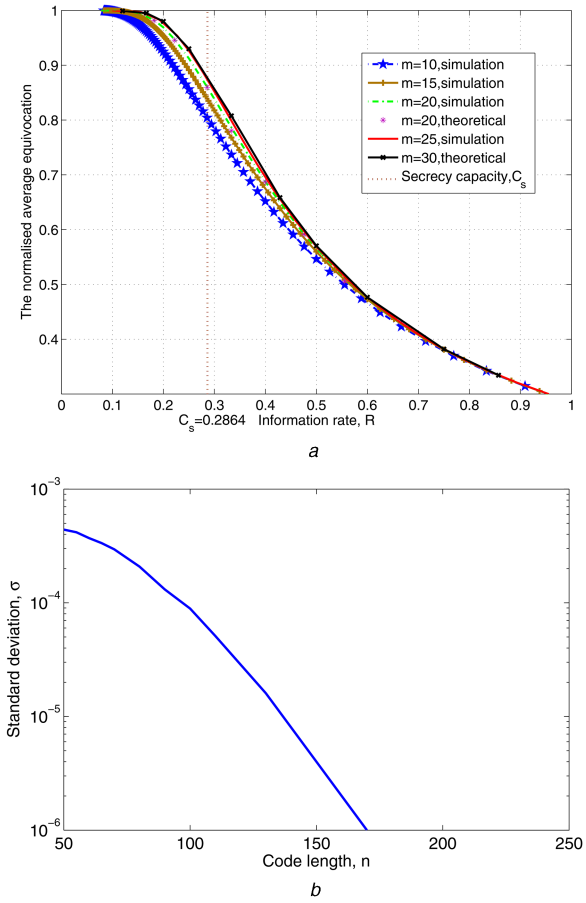

Fig. 5 *Results of Monte Carlo simulation and event-based analysis*
(a) Mean equivocation rate versus information rate for RPCs with different values of m, ***(b)*** Standard deviation of the average equivocation rate versus n for RPCs with $m = 20$

2. There are 100 S_e^1 syndromes, with $x = p(S_e = S_e^1)$).
3. There are 4950 S_e^2 syndromes, with $x = p(S_e = S_e^2)$).
4. There are $2^{15} - 5051$ syndromes with $x = p(S_e = S_e^{TS})$.

Using this method, the pmf $f_S(x)$ has been evaluated and it also has been evaluated by the method of Monte Carlo based simulation of 10.000 RPCs, for comparison purposes.

Fig. 3*a* shows a plot for $S \in S_e^{TS}$)) and Fig. 3*b*) shows a plot for $S \in S_e^1$)). Figs. 3*c* and *d* show plots for the two similar pmfs obtained by Monte Carlo simulation. It can be seen that the convolution method and Monte Carlo simulation produce almost identical results.

To obtain a quantitative analysis of how closely matched are the results the Kolmogorov–Smirnov (K-S) test may be employed [18]. Fig. 4 shows the CDFs of $p(S_e^1)$ and $p(S_e^{TS})$. It can be seen in each case that the plots are virtually on top of each other. The closeness is borne out by the K–S supremum values of $D_n = 0.01689$ and $D_n = 0.02304$, respectively.

For both methods, the calculation of the average equivocation rate may be carried out for typical values of the BSC transition probability by means of evaluation of equations (17) and (29), respectively. We have considered $\alpha = 0.05$. Monte Carlo analysis produces $\mathbb{E}[R_e] = 0.9905$ whilst event-based analysis produces $\mathbb{E}[R_e] = 0.9886$ indicating a close agreement between the two methods.

5 Results for a range of RPC parameters

In this section, some results are given for the average equivocation rate for different RPC parameters. Results are also compared between the Monte Carlo method and the event-based analysis method. The effectiveness of using RPCs in syndrome coding to achieve secrecy is also demonstrated.

5.1 Monte Carlo simulation compared with the event-based analysis method

Fig. 5*a* shows the mean of the equivocation rate versus the information rate, R, for RPCs with various values for the code parameter, m, noting that the code length is $n = m/R$. Curves are plotted based on the event-based analysis method for $m = 20$ and 30, and for Monte Carlo simulation using 10,000 different RPCs for $m = 10, 15, 20, 25, 30$. The differences in results between the two equivocation evaluation methods, which are shown for $m = 20$ and 30, are all $< 0.3\%$, indicating a close agreement. Moreover, the theoretical analysis method is also much faster for those codes with large values of n and m, which can take a long time using the alternative Monte Carlo simulation. For some code parameters, the theoretical analysis method is the only practical way to evaluate the average equivocation rate.

One advantage of Monte Carlo analysis is that it does show the variation in equivocation rate between different RPCs with the same code parameters. Fig. 5*b* shows the standard deviation of R_e for 10,000 different RPCs. It can be seen that the longer the code the lower the standard deviation. Additionally, it is evident that at $m = 20$ the standard deviation is insignificant for codes longer than 50 bits in length implying that any RPC is as good as any other RPC with the same length and value of $m = 20$.

Fig. 5*a* also shows that for the BSC with error probability $\alpha = 0.05$, RPCs with information rate < 0.2 can achieve an equivocation rate approaching 1, implying perfect secrecy (when the equivocation rate tends to 1, the communication is perfect secrecy) for syndrome coding using these codes.

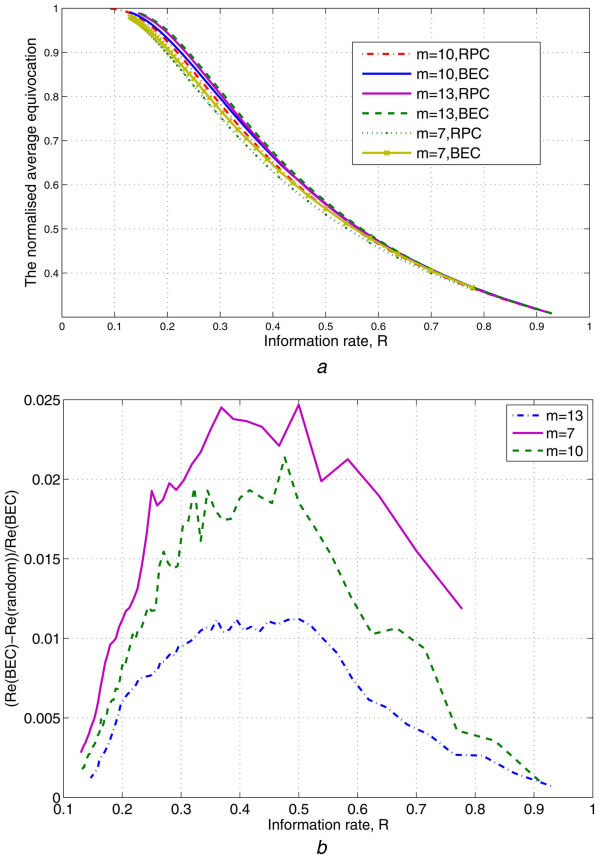

Fig. 6 *RPCs versus the best codes, BECs*
(a) Mean of the equivocation rate versus information rate for RPCs compared with the best codes, BEC, ***(b)*** Normalised difference of R_e between RPC and BEC

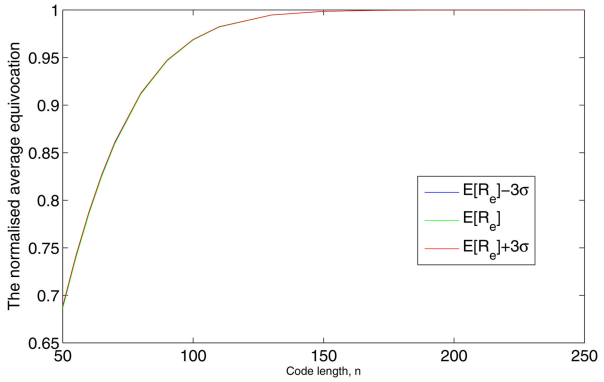

Fig. 7 *Mean of the equivocation rate versus n for RPCs with* $m = 20$ *evaluated over 10,000 codes*

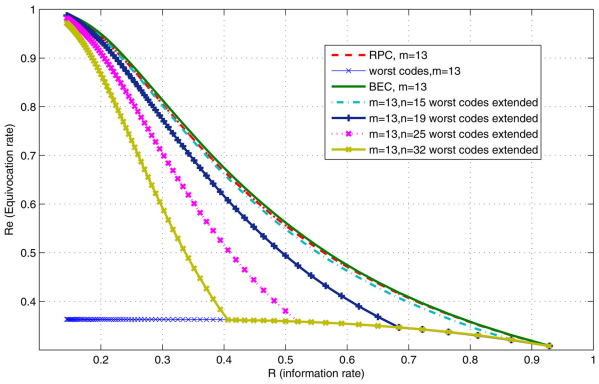

Fig. 8 *Mean of the equivocation rate of RPC and random codes*

5.2 RPCs in comparison with optimised codes, the BECs

Design methods are given in [14] that show how to construct codes with the highest achievable level of equivocation for syndrome coding. The resulting codes are known as the BECs, which are the codes achieving the highest security level in syndrome coding. In this section, we compare the security of RPCs with that of the corresponding BEC having the same code parameters.

Fig. 6*a* plots the average equivocation rate of RPCs, R_e(RPC), and the equivocation rate of corresponding BEC, R_e(BEC) for $m = 7, 10, 13$. Fig. 6*a* shows that R_e(RPC) is very close to R_e(BEC) for these values of m. More clearly, Fig. 6*b* shows the normalised difference of R_e between these two classes of codes, $(R_e(\text{BEC}) - R_e(\text{RPC}))/R_e(\text{BEC})$. The normalised difference is $< 2.5\%$ and decreases as the value of m gets larger.

The design procedures for BECs mean that these are best constructed off-line. If a communication application requires several different codes then BECs will need to be stored by the application. On the other hand, RPCs can easily be generated on-the-fly as needed. Since RPCs have performance comparable to BECs, RPCs are the preferred choice for these applications.

In some applications, it is desirable to have forward secrecy where a different code is used for each communication session. A fixed code such as a BEC is undesirable, even though the BEC has the ultimate performance. Since RPCs are generated randomly and a different code is used for each communication session, perfect forward secrecy may be guaranteed.

5.3 Security level versus code length

In this section, we explore the relationship between the code length, n, and the mean of the equivocation rate, R_e for RPCs with a given value of m for the BSC operating point, $\alpha = 0.05$. Fig. 7 shows the average equivocation rate R_e as a function of codelength $n$and indicated 3σ limits in performance. $Re_{\text{ave}} \pm 3\sigma$ versus code length. n for $m = 20$. Also shown in Fig. 7 are the 3σ limits so that there will be 99% of codes with $R_e - 3\sigma \leq R_e \leq R_e + 3\sigma$, where σ is the standard deviation of R_e. Fig. 7 shows that there is a monotonic relationship between codelength and the average equivocation rate.

From the standard deviation of R_e shown in Fig. 5b, it is evident that the longer the code the lower the standard deviation. For a code longer than 150 bits, any randomly chosen permutation code achieves almost perfect secrecy.

5.4 RPCs compared with the worst unconstrained random codes

The main use of RPCs is in modern communication applications using syndrome coding that requires forward secrecy where an ephemeral code is needed for each communication session. Rather than using RPCs an unconstrained random code could be used instead. As discussed above, the equivocation achieved by unconstrained random codes can be bad. The very worst unconstrained random codes will have many repeated columns of the parity check matrix and will have corresponding poor levels of secrecy, leaking information to the eavesdropper.

Fig. 8 shows the equivocation rate of the worst RPCs ($R_e - 3\sigma$ performance) in comparison with the worst unconstrained randomly chosen codes. These are random codes (generated in the following ways: extended shortened Hamming codes with repeated columns and extending the worst code with random columns for the code. Fig. 8 shows extreme cases of repeated columns and their deleterious effect in decreasing the value of the equivocation rate. For all these codes $m = 13$. Also shown in Fig. 8 for comparison purposes is the equivocation rate for the best codes, the BECs.

6 Conclusions

In this study, we described a class of (n,k) randomly chosen codes, RPCs, for the syndrome coding scheme which is designed to prevent leakage of information to an eavesdropper. We studied the effect of the columns of the parity check matrix on the security levels achieved as measured by the equivocation rate. A theoretical

method based on the weight of error events in the BSC was described for calculating the average equivocation rate that is achieved by RPCs for a given bit error rate, and for given code parameters. Both this method and the Monte Carlo simulation method have been used to evaluate the equivocation rate of codes with given code parameters. It has been shown that for randomly chosen codes it is the repeated columns of the parity check matrix that degrades the achieved security. Repeated columns can never occur for RPCs.

The security performance of a wide range of (n,k) RPCs has been studied. The error event weight analysis method models the pmf of syndrome probabilities and makes it possible to calculate the mean of the equivocation rate for general (n,k) RPCs, including those code parameters where exhaustive evaluation is impossible. Results have been compared with those from Monte Carlo simulation showing a close agreement. The error event weight analysis method is useful because it enables the average equivocation rate to be calculated for those codes whose parameters are such that Monte Carlo analysis cannot be carried out in a reasonable computation time.

The security performance of various RPCs has been presented. The results show that the longer the code the higher the mean of the equivocation rate and the lower the standard deviation. The results also show that the lower the information rate, the higher is the mean of the equivocation rate and the secrecy achieved. When 20 or more parity bits are used almost complete secrecy is achieved by any RPC whose code length is longer than around 120 bits, for a BSC bit error rate of $\alpha = 0.05$.

The Monte Carlo simulation results also show that the security level of any RPC is virtually the same as that achieved by using an optimum code, a BEC, for the same code parameters. This means that RPCs are ideal for those communication applications which require perfect forward secrecy where a different, ephemeral code is used for each communication session.

7 Acknowledgments

This work was partially supported by the China Postdoctoral Science Foundation (grant no. 2015M582287), the Fundamental Research Funds for the Central Universities (WUT:2017 IVA 052) and the National Natural Science Foundation of China (grant no. 61502361).

8 References

[1] Shannon, C.E.: 'A mathematical theory of communication', *Bell Syst. Tech. J.*, 1948, **27**, (3), pp. 379–423, 623–656

[2] Gallager, R.G.: '*Information theory and reliable communication*' (John Wiley and Sons, New York, 1968)

[3] Gallager, R.G.: 'The random coding bound is tight for the average code', *IEEE Trans. Inf. Theory*, 1973, **19**, (2), pp. 244–246

[4] MacMullan, S.J., Collins, O.M.: 'A comparison of known codes, random codes, and the best codes', *IEEE Trans. Inf. Theory*, 1998, **44**, (7), pp. 3009–3022

[5] Barg, A., Forney, G.D.: 'Random codes: minimum distances and error exponents', *IEEE Trans. Inf. Theory*, 2002, **48**, (9), pp. 2568–2573

[6] Wyner, A.D.: 'The wire-tap channel', *Bell Syst. Tech. J.*, 1975, **54**, (8), pp. 1355–1367

[7] Leung-Yan-Cheong, S.K., Hellman, M.E.: 'The Gaussian wire-tap channel', *IEEE Trans. Inf. Theory*, 1978, **24**, (4), pp. 451–456

[8] Liang, Y., Poor, H.V., Shamai, S.: 'Information theoretic security', *Found. Trends Commun. Inf. Theory*, 2008, **5**, (4–5), pp. 355–580

[9] Thangaraj, A., Dihidar, S., Calderbank, A.*, et al.*: 'Application of LDPC codes to that wiretap channel', *IEEE Trans. Inf. Theory*, 2007, **53**, (8), pp. 2933–2945

[10] Liu, D., Lin, S., Poor, H.V.*, et al.*: 'Secure nested codes for type II wire-tap channels'. IEEE Information Theory Workshop, Lake Tahoe, CA, USA, September 2007, pp. 337–342

[11] Csiszár, I., Körner, J.: 'Broadcast channels with confidential messages', *IEEE Trans. Inf. Theory*, 1978, **24**, (3), pp. 339–348

[12] Cohen, G., Zémor, G.: 'Syndrome-coding for the wiretap channel revisited'. IEEE Information Theory Workshop (ITW2006), Chengdu, China, October 2006, pp. 33–36

[13] Chen, Y., Vinck, A.J.H.: 'On the binary symmetric wiretap channel'. Proc. Int. Zurich Seminar on Communications, Zurich, Switzerland, March 2010, pp. 17–20

[14] Zhang, K., Tomlinson, M., Ahmed, M.Z.*, et al.*: 'Best binary equivocation code construction for syndrome coding', *IET Commun.*, 2014, **8**, (10), pp. 1696–1704

[15] Albrecht, M., Cid, C., Paterson, K.G.*, et al.*: 'NTS-KEM'. Available at https://csrc.nist.gov/Projects/Post-Quantum-Cryptography/Round-1-Submissions, accessed January 2018'

[16] Menzies, A., van Oorscot, P.C., Vanstone, S.: '*Handbook of applied cryptography*' (CRC Press, Boca Raton, 1996)

[17] Dumer, I.I., Farrell, P.G.: 'Erasure correction performance of linear block codes'. Algebraic Coding 1993, 1994 (LNCS, 781)

[18] Justel, A., Peña, D., Zamar, R.: 'A multivariate Kolmogorov-Smirnov test of goodness of fit', *Stat. Probab. Lett.*, 1997, **35**, (3), pp. 251–259