INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

A Graph Theoretic Approach to Randomness Test Based on the Overlapping Blocks

Muhiddin Uğuz

Department of Mathematics, Middle East Technical University, Ankara, Turkey
muhid@metu.edu.tr

Research Paper Received: 27.04.2023 Revised: 08.06.2023 Accepted: 19.06.2023

Abstract—Cryptographic parameters such as secret keys, should be chosen randomly and at the same time it should not be so difficult to reproduce them when necessary. Because of this, pseudorandom bit (or number) generators take the role of true random generators. Outputs of pseudorandom generators, although they are produced through some deterministic process, should be random looking, that is not distinguishable from true random sequences. In other words they should not follow any pattern. In this paper we propose a new approach using graph theory, to determine the expected value of the index at which a fixed pattern start to appear in a random sequence for the first time. Using the method proposed, a recursion for the number of paths of length n starting from a pattern and never coming back to that pattern can be computed. By means of these recursions, we obtain the probabilities for the indexes at which a fixed pattern appears in the sequence for the first time. Using these expected values and comparing them with the observed values a randomness test can be defined. In this work patters are traced through the sequence in an overlapping manner.

Keywords—Cryptography, Statistical Randomness Testing, NIST Test Suite, Recursion

1. Introduction

The concept of random sequences is vital in cryptography and also in many other fields varying from statistic to computer simulations. In cryptography, random sequences are needed not only in symmetric key encryption or key generation but also for generation of primes for RSA encryption, initialization vectors, salts in hash functions and the like.

Sources of True Random Number Generators (TRNG) are usually some complicated physical events such as lightnings or atmospheric or thermal noises and hence reproduction of them are usually very difficult if not impossible and hence they are not practical in cryptographic applications. The solution for this problem is Pseudo Random Number Generators (PRNG). PRNG produces random looking sequences from short random seeds, by making use of deterministic algorithms. Sequences produced by PRNGs must behave like those obtained from TRNGs, that is, they should not contain any recognizable pattern or an order.

In order to be used safely in cryptography, PRNG's and their outputs must be tested in terms of randomness from many different aspects. A set of statistical randomness tests, called a test suite, can be used to make sure that there is no weakness in the randomness of the sequence that will be used. There

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

are many documents outlining how these statistical randomness test can be designed, [1] gives a detailed information on this.
There are many tests defined in the literature [2], [3], [4], [5], [6], [7], [8], [9] are some of them. Also many test suites are available in the literature [10], [11], [12], [13], [14], [15].

The number of rounds at which a block cipher achieves randomness is one of the most important design criteria. Soto et. al.[16] used this idea and analyzed AES competition finalist algorithms from this point of view, using the NIST test suite [15]. In the NIST test suite, there are two randomness tests considering the number of occurrences of a predefined template in a sequence, namely the overlapping template matching test and the non overlapping template matching test. Their computations is valid only for the template $B = 111111111$. In fact, the probabilities changes depending on the period of the template. In [17], the classification of all possible templates according to their period is given and for each template the exact values of the probabilities are evaluated using generating functions. Finally a new statistical randomness test is proposed.

In this work, we propose a new approach to calculate the probabilities for overlapping templates using a graph theoretical method. Using the obtained values a randomness test can be defined following the steps described in [1].

The organization of the paper is as follows. In section 2, we propose the problem, and then define reversed graph with its transition matrix, and we state and prove two theorems. In section 3, we give probability values for the pattern 010, and in section 4, we give recursions with initial values for the other patterns of length 3. In section 5, we listed characteristic polynomials for all patterns of length 4, from which recursions and then probability values can be derived. In section 6, we derived an explicit formula for the generating function of the probability sequence of the pattern 010, and computing its derivative at 1, we obtain the expected value of the first occurrence of the pattern in a random sequence. We finish the paper with a conclusion.

Table 1.
A binary sequence example

index j	1	2	3	4	5	6	7	8	9	10	$\cdots$
r_j	1	0	0	1	1	0	**0**	**1**	**0**	1	$\cdots$

2. Overlapping Blocks

Let $\{r_i\} = r_1, r_2, r_3, \ldots$ be a binary sequence and $P = b_1 b_2 \cdots b_l$ be a fixed pattern. In this paper the formulas for the followings are given:

- For each k, the probability P_k, corresponding to the first occurrence of the pattern P to be at the position k, is calculated.
- Let j be the first observed position of P in the sequence. The expected value of j is calculated.

For example, considering the binary patterns of length three, for the case $P = 010$, corresponding to the integer 2, consider the sequence $\{r_i\}$ given in the binary sequence example given in Table 1.

The first occurrence of the pattern $P = 010$ is at the seventh position. Using the integers a_j corresponding to $(r_j, r_{j+1}, r_{j+2})_2 \in \{0, 1, 2, 3, 4, 5, 6, 7\}$, one can express the same binary sequence $\{r_j\}_{j=1}^{n}$ in the form $\{a_j\}_{j=1}^{n-2}$ as in the Corresponding Integer Sequence example given in Table 2. From this equivalent point of view, the problem we are interested is to determine the index at which a pattern, $2 = (010)_2$ as an example, is expected to be observed for the first time.

This way, any binary sequence $\{r_i\}_{i=1}^{l}$ can be identified with the corresponding integer sequence $\{a_i\}_{i=1}^{l-2}$.

Table 2.
Corresponding Integer Sequence

index j	1	2	3	4	5	6	7	8	$\cdots$
a_j	4	1	3	6	4	1	**2**	5	$\cdots$

Table 3.
Examples of paths and the corresponding sequence

Path 1	Sequence
2	**010**
3, 7, 6, 5, 2	0111**010**
3, 7, 7, 6, 4, 1, 2	011110**010**

Notice that even if the sequence $\{r_i\}$ is a random binary sequence on the set $\mathbf{Z}_2$, the corresponding sequence $\{a_i\}$ will not be a random sequence on the set $\mathbf{Z}_8$. If $a_i = 0$, as an example, a_{i+1} can have only two values, namely 0 or 1. In fact there are only two possible values for a_{i+1} depending on a_i and r_{i+2}. More clearly, modulo 8, a_i is either $2a_{i-1}$ or $2a_{i-1} + 1$ depending on whether r_{i+2} is 0 or 1.

Consider the directed graph, called **adjacent graph** with eighth vertices corresponding to binary patterns $a_j = (r_j, r_{j+1}, r_{j+2})$, given below. Each vertex of this directed graph has two successor vertices and two predecessor vertices.

Consider a path on this graph starting from a vertex and terminating as soon as it reaches to the vertex **2**. Any such path corresponds to a binary sequence whose first three terms are determined by the initial vertex. Table 3 lists three of such binary sequences and their corresponding paths.

Using this graph theoretic terminology, one can list all of the paths starting from the vertex **0**, from the vertex **1**, ... , and finally from the vertex **7** and terminating as soon as they reach to vertex **2** and listing their length, one can compute the probability of this length to be k. An easier method is to make use of the *reverse graph* obtained by reversing the orientations of the paths. Considering all paths of length k, starting from the vertex **2** and not coming back to it, the desired probability can be calculated. For this purpose, both of the two edges that direct to the vertex **2** are deleted in the graph **Reverse Graph with Edges to Vertex** 2 **Deleted** below.

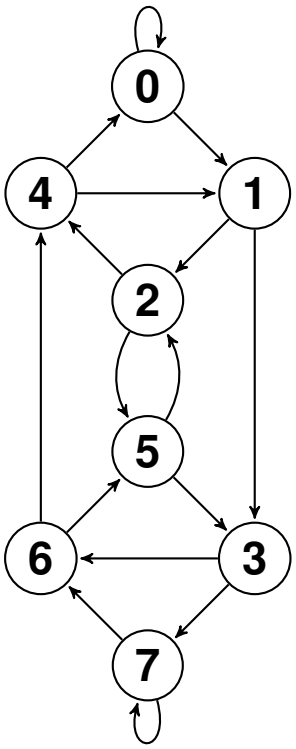

Figure 1. Adjacent Graph

Table 4.
Examples of paths starting from 2 and the corresponding sequences

Path 1	Path 2 (reversed)	Sequence
2	**2**	**010**
3, 7, 6, 5, **2**	**2**, 5, 6, 7, 3	0111**010**
3, 7, 7, 6, 4, 1, **2**	**2**, 1, 4, 6, 7, 7, 3	011110**010**

Table 4 lists three of such binary sequences and their corresponding paths.

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

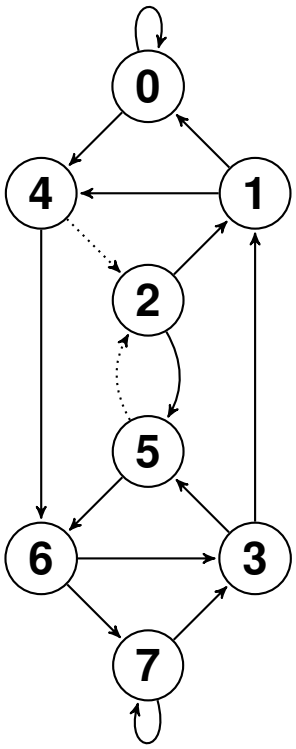

Figure 2. Reverse Graph with Edges to Vertex 2 Deleted

Let l_k denote the number of all paths, starting from the vertex **2** of the reversed graph with edges to **2** deleted, of length (after the initial 2) equal to k. Note that the pattern **2** is not considered when the length is determined. For $k = 1$ and $k = 2$ complete list of all paths are as follows:

- There are two paths of length $k = 1$, namely $2, 1$ and $2, 5$ and hence

$$l_1 = 2.$$

- There are three paths of length $k = 2$, namely $2, 1, 0$; $2, 1, 4$ and $2, 5, 6$ and hence

$$l_2 = 3.$$

Definition 1 *The transition matrix A of a directed graph is defined as the square matrix with entries A_{ij} equal to the number of edges in the graph from the vertex i to the vertex j*

Let T denote the matrix obtained from the transition matrix R of the **reversed graph** by deleting the edges between **2** and its predecessor vertices.

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Notice that, since both of the edges to the vertex **2** are deleted, the third column of T is all zero, that is there is no edge pointing to **2**. The third row, namely $0\ 1\ 0\ 0\ 0\ 1\ 0\ 0$ indicates that the only paths from 2 are to 1 and to 5. Moreover the rank of T which can be computed as the number of linearly independent columns of T is 4. Consider T, T^2 and T^3

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

$$T^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$T^3 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Each entry 1 in the third rows that corresponds to **2**, of all of these three matrices above is in one to one correspondence with a path having the vertex **2** as a starting point. More clearly,

- Two paths $2,1$ and $2,5$ of length 1 are represented by the two 1's in the third row of matrix T.
- Three paths $2,1,0$, $2,1,4$ and $2,5,6$ of length 2 are represented by the three 1's in the third row of matrix T^2.
- Five paths $2,1,0,0$; $2,1,0,4$; $2,1,4,6$; $2,5,6,3$ and $2,5,6,7$ of length 3 are represented by the five 1's in the third row of matrix T^3.

We claim that this is not a coincidence. In fact,

Theorem 2 *Consider the reversed graph with edges to **2** deleted. Then $(T^k)_{ij} = T^k_{ij}$, that is the $(i,j)^{th}$ entry of the matrix T^k, is equal to the number of all paths in this graph starting from the vertex i to vertex j of length k.*

Proof: We will use mathematical induction. For $k = 1$, it is true by the definition of matrix T. Assume the statement is true for k, and consider T^{k+1}. Recall the matrix multiplication: To obtain T^{k+1}_{ij} we multiply the i^{th} row of T with j^{th} column of T^k. That is;

$$T^{k+1}_{ij} = \sum_{r=0} T_{ir} \cdot T^k_{rj}$$

Here T_{ir} is the number of paths from vertex i to vertex r of length 1, and T^k_{rj} is, by the induction hypothesis, the number of paths from vertex r to vertex j of length k and hence the product $T_{ir} \cdot T^k_{rj}$ summed over the index r gives the total number of paths of length $k+1$ from vertex i to vertex j. □

Recall that aim of this paper is to count the number of all paths of length k, starting from a fixed vertex, as an example from vertex 2. In other words, to find the sum of all entries in the 3^{rd} row of T^k. For this reason we want to find T^k, or sum of all its entries in a row, in a practical way; for example in a recursive manner. We first illustrate this idea of computing sum of all elements in a row of a matrix using recursion, by an example. First of all we need to obtain a polynomial satisfied by the matrix.

Recall that eigen values of a square matrix $n \times n$ matrix A are roots of the characteristic polynomial of the matrix, defined by $det(A - \lambda I)$ where I denote the $n \times n$ identity matrix. Trace of a square matrix is defined as the sum of diagonal elements. Equivalently, it is equal to sum of the eigenvalues of the matrix. Similarly determinant is equal to the product of eigenvalues.

Example 1 *Consider two bit patterns **0** $= 00$, **1** $= 01$, **2** $= 10$, and **3** $= 11$. The two possible successors of each of these four patterns are;*
$0 \to 0,1 \qquad 1 \to 2,3 \qquad 2 \to 0,1 \qquad 3 \to 2,3$
and hence, data for the reversed graph is;

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

$0 \to 0, 2 \qquad 1 \to 0, 2 \qquad 2 \to 1, 3 \qquad 3 \to 1, 3.$

In other words, the transition matrix T of the corresponding reversed graph is

$$T = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Recall that trace of a square matrix is defined as the sum of diagonal elements. Equivalently, it is equal to sum of the eigenvalues of the matrix. Similarly determinant is equal to the product of eigenvalues. Eigen values of a square matrix $n \times n$ matrix A are roots of the characteristic polynomial of the matrix, defined by $det(A-\lambda I)$ where I denote the $n \times n$ identity matrix.

Notice that this matrix has trace equal to 2 and determinant equal to 0. We can compute powers of this matrix easily and get

$$T^2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = J,\ T^3 = 2J,$$

$$T^4 = 4J,\ T^5 = 8J$$

and hence for $k \geq 4$, by induction we have, $T^k = 2T^{k-1}$. From this observation we see that matrix T satisfies the equation

$$x^3(x-2) = 0.$$

Therefore for $k \geq 4$, the minimal polynomial of T^k is $(x-2)$.

In the general case, T satisfies the polynomial

$$x^{n+1} = trace(T)x^n - det(T)x^{n-1}.$$

This polynomial equation defines a recursion that that can be used to compute powers of the matrix T easily: a linear combination of 1 and 2 less powers of the same matrix gives the power of the matrix. Moreover, the sum of entries in the ith row of T^k can be computed easily by means of this recursion, and hence the total number of paths from vertex i can be obtained. As an example, if A is a 2×2 matrix, using the notation

$$A^n = \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix},$$

and the recursion $x^{n+1} = trace(T)x^n - det(T)x^{n-1}$, the following equations can be written

$$\begin{aligned} c_{n+1} &= trace(T)c_n - det(T)c_{n-1} \\ d_{n+1} &= trace(T)d_n - det(T)d_{n-1} \end{aligned}$$

Notice that the same linear recurrence relation is also satisfied by a sum of entries of the matrix in any fixed row or fixed column of the matrix. example if

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \text{then } A^2 = \begin{bmatrix} 7 & 9 \\ 15 & 22 \end{bmatrix}$$

and hence, to compute

$$A^3 = \begin{bmatrix} a_3 & b_3 \\ c_3 & d_3 \end{bmatrix},$$

using the formula $c_3 = trace(A)c_2 - det(A)c_1$, one obtains $c_3 = 5(15) - (-2)3 = 81$. Similarly we can easily compute d_3 and hence $c_3 + d_3$.

This way a recursion to compute the sum of all entries in a fixed row of A^k is obtained. This sum, in the case A is the reverse of transition matrix, equal to the total number of paths of length k starting from a certain vertex defined by the row. The degree of the recursion is the same as the degree of the characteristic equation of A.

Now, turning back to the 8×8 matrix T with characteristic polynomial $x^5(x^3 - 2x^2 + x - 1)$, one can can write

$$l_{n+3} = 2l_{n+2} - l_{n+1} + l_n$$

as a recursion satisfied by l_n and hence the following theorem can be stated:

Theorem 3 *Let l_n denote the number of all paths of length n starting from the vertex **2** and never coming back to **2**. Then l_n satisfies the following recursion relation*

$$l_{n+3} = 2l_{n+2} - l_{n+1} + l_n.$$

By convention $l_0 = 1$, and with simple counting $l_1 = 2,\ l_2 = 3$ and $l_3 = 5$ (and hence $l_4 = 9,\ l_5 = 16, \ldots$ and so on).

3. Probability Computations of Overlapping Blocks

Recall that on the reverse graph, the total number of paths that starts with **2** and the pattern **2** never appears again, of length k, is denoted by l_k and l_k satisfies certain recursion. In other words, l_k is the cardinality of the set

$$\{(0, 1, 0, b_1, b_2, \ldots, b_k) \in Z_2^{k+3} : \\ (b_i, b_{i+1}, b_{i+2}) \neq (0, 1, 0) = \mathbf{2}\}$$

This means that, l_k of all 2^{k+3} possible sequences, of bit length $k + 3$ satisfies the condition: of being of length k and the pattern **2** does not appear. This means that

- For sequences of bit-length $4 : (r_0r_1r_2r_3)$. There are 16 of them and each contains 2 patterns of length 3: $r_0r_1r_2$ and $r_1r_2r_3$. Pattern length is 2 and exactly $l_1 = 2$ of them starts with $(\mathbf{010})$, and does not reach to $(010)_2 = \mathbf{2}$ again. They are; **010**0 and **010**1. Hence the probability is $P_4 = \frac{l_1}{2^4} = \frac{2}{16} = \frac{1}{8}$.
- For sequences of bit-length $5 : (r_0r_1r_2r_3r_4)$. There are 32 of them and each contains 3 patterns of length 3: $r_0r_1r_2,\ r_1r_2r_3$ and $r_2r_3r_4$ and hence pattern length is 3. Exactly $l_2 = 3$ of them starts with (010), and does not reach to **2** again. These are; **010**$00,$ **010**01 and **010**11. Hence $P_5 = \frac{l_2}{2^5} = \frac{3}{32}$.
- Similarly, $P_6 = \frac{l_3}{2^6} = \frac{5}{64}$, for sequences of length 7 is $P_7 = \frac{l_4}{2^7} = \frac{9}{128}$, for sequences of length 8 is $P_8 = \frac{l_5}{2^8} = \frac{16}{256}$, and so on.

Consider random binary sequence of length $i + 3$ and let P_i denotes the probability that this sequence starts with **2** and never comes back to **2**. We have seen that

$$P_i = \frac{l_{i-3}}{2^i}$$

Defining $l_0 = 1$ for convention, (and hence $P_3 = \frac{1}{8}$), notice that

$$P_3+P_4+P_5+\cdots = \frac{1}{8}+\frac{1}{8}+\frac{3}{32}+\frac{5}{64}+\frac{9}{128}+\cdots = 1.$$

Now we can define a randomness test, that is χ-square goodness of fit test, by defining the five boxes as;

- **2** appears for the first time at position 0 or 1: Probability of this is $\frac{1}{8} + \frac{1}{8} = \frac{1}{4} = 0.25$.
- **2** appears for the first time at position 2 or 3 or 4: Probability of this is $\frac{3}{32} + \frac{5}{64} + \frac{9}{128} = \frac{31}{128} = 0.242188$.
- **2** appears for the first time at position 5 or 6 or 7 or 8 or 9: Probability of this is $\frac{16}{256} + \frac{28}{512} + \frac{49}{1024} + \frac{86}{2048} + \frac{151}{4096} = 0.243896$.
- **2** appears for the first time at position between 10 and 35: Probability of this is $\frac{265}{8192} + \frac{465}{16384} + \cdots + \frac{20330163}{8589934592} = 0.246972054$
- **2** appears for the first time at position after 35: Probability of this is 0.016944.

Notice that in order to be able to use this bin values in χ square goodness of fit test, the number of

overlapping blocks of length three should be at least $5 \times \frac{1}{0.016944} \approx 295$, and hence length of the sequence should be at least 297.

4. Other patterns of length 3

Notice that all definitions, explanations, theorems to this point are valid for the pattern $\mathbf{2} = (101)$. Now we will consider the other patterns of length $l = 3$. There are eight such patterns: **0**,**1**,**2**,**3**,**4**,**5**,**6**,**7**. Lets denote the transition matrix of the reverse graph with edges between i and its predecessor vertices deleted by $T_{(i)}$. Then the matrix T above in this new notation is $T_{(2)}$.

$$T_{(0)} = \begin{bmatrix} \mathbf{0} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{0} & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3-x^2-x-1)$.

$$T_{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3 - 2x^2 + 1)$.

$$T_{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3-2x^2+x-1)$.

$$T_{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3 - 2x^2 + 1)$.

$$T_{(4)} = \begin{bmatrix} 1 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3 - 2x^2 + 1)$.

$$T_{(5)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

and its characteristic equation is $x^5(x^3-2x^2+x-1)$.

$$T_{(6)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \mathbf{0} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \mathbf{0} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and its characteristic equation is $x^5(x^3-2x^2+1)$.

$$T_{(7)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \mathbf{0} \end{bmatrix}$$

and its characteristic equation is $x^5(x^3-x^2-x-1)$.

Thus

- $T_{(0)} = T_{(7)}$: and the recursion is

  $l_{n+3} = l_{n+2}+l_{n+1}+l_n$ with $l_0 = 1,\ l_1 = 1,\ l_2 = 2$.

- $T_{(1)} = T_{(3)} = T_{(4)} = T_{(6)}$: and the recursion is

  $l_{n+3} = 2l_{n+2} - l_n$ with $l_0 = 1,\ l_1 = 2,\ l_2 = 4$.

- $T_{(2)} = T_{(5)}$: and the recursion is

  $l_{n+3} = 2l_{n+2}-l_{n+1}+l_n$ with $l_0 = 1,\ l_1 = 2,\ l_2 = 3$.

As for the pattern 010, for each of the other patterns of length 3, using these recursions obtained above, corresponding randomness tests can be defined.

5. Blocks of length bigger than 3

As mentioned above, for longer fixed patterns of length 3 or more, the same arguments work. As an example, for the pattern 0000 of length 4, the transition matrix and its characteristic polynomial and similarly characteristic polynomials of all other patterns of length 4 are given below. Using these polynomials, one can easily obtain recursions as above and hence compute probability and expected values for each of these patterns.

$T_{(0)} =$

$$\begin{bmatrix} \mathbf{0} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \mathbf{0} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

and its characteristic equation is $x^{16} - x^{15} - x^{14} - x^{13} - x^{12} = x^{12}(x^4 - x^3 - x^2 - x - 1)$

Similarly we obtain the following characteristic polynomials:

$$\begin{aligned} C_P(T_{(0)}) &= x^{12}(x^4 - x^3 - x^2 - x - 1) \\ &= C_P(T_{(15)}) \\ C_P(T_{(1)}) &= C_P(T_{(14)}) = C_P(T_{(7)}) = C_P(T_{(8)}) \\ &= (x^4 - 2x^3 + 1)x^{12} \\ C_P(T_{(2)}) &= C_P(T_{(13)}) = C_P(T_{(4)}) = C_P(T_{(11)}) \\ &= x^{12}(x^4 - 2x^3 + x - 1) \\ C_P(T_{(3)}) &= C_P(T_{(12)}) = x^{12}(x^4 - 2x^3 + 1) \\ C_P(T_{(5)}) &= (x^4 - 2x^3 + x^2 - 2x + 1)x^{12} \\ &= C_P(T_{(10)}) \\ C_P(T_{(6)}) &= C_P(T_{(9)}) = x^{12}(x^4 - 2x^3 + x - 1) \end{aligned}$$

INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
M. Uğuz, Vol.12, No.2, pp.42-52.
https://doi.org/10.55859/ijiss.1288854

6. Expected Values

Here we will derive the expected value formula for the pattern 101. Using the recursions obtained above, it is straightforward to derive corresponding formulas for the other patterns.

Recall that if a random binary sequence generator stops the generation as soon as $\mathbf{2} = (010)$ appears and if P_i denotes the probability that length of this sequence is i, we have

$$P_i = \frac{l_{i-3}}{2^i}.$$

Let E denotes the expected value of the length of such a sequence. Then $E = \sum_{i=3}^{\infty} i \cdot P_i$.

Consider the generating function of the sequence $\{P_i\}$, say $F(x) = \sum P_i x^i$. Then $F(1)$ is the sum of all probabilities and hence $F(1) = 1$. Moreover, since $F'(x) = \sum_{i=1}^{\infty} i \cdot P_i \cdot x^{i-1}$, we have $F'(1) = E$.

Now recall that for the pattern 101, the sequence $\{l_i\}$ satisfies the recursion $l_{n+3} = 2l_{n+2} - l_{n+1} + l_n$ where $l_0 = 1,\ l_1 = 2,\ l_2 = 3,\ l_3 = 5, \ldots$. Moreover $P_0 = 0,\ P_1 = \frac{l_0}{2^3},\ P_2 = \frac{l_1}{2^4}, \ldots$. Thus, using this recursion, we can write

$$\begin{aligned} F(x) &= P_1 x + P_2 x^2 + P_3 x^3 + \cdots + P_n x^n + \cdots \\ &= \frac{l_0}{2^3}x + \frac{l_1}{2^4}x^2 + \frac{l_2}{2^5}x^3 \\ &\quad + \frac{2l_2 - l_1 + l_0}{2^6}x^4 + \frac{2l_3 - l_2 + l_1}{2^7}x^5 + \cdots \end{aligned}$$

Hence, $F(x)$ can be expressed as rational function:

$$\begin{aligned} F(x) &= \frac{1}{8}x + \frac{2}{16}x^2 + \frac{3}{32}x^3 \\ &\quad + \left(\frac{l_2}{2^5}x^3 + \frac{l_3}{2^6}x^4 + \ldots\right)x \\ &\quad - \left(\frac{l_1}{2^4}x^2 + \frac{l_2}{2^5}x^3 + \ldots\right)x^2 \\ &\quad + \left(\frac{l_0}{2^3}x + \frac{l_1}{2^4}x^2 + \ldots\right)x^3 \end{aligned}$$

Substituting $l_0 = 1,\ l_1 = 2,\ l_2 = 3$, we obtain,

$$\begin{aligned} F(x) = \frac{x}{8} + \frac{x^2}{8} + \frac{3}{32}x^3 &+ \left[F(x) - \frac{1}{8}x^2 - \frac{1}{8}x\right]x \\ &- \frac{1}{4}\left[F(x) - \frac{1}{8}x\right]x^2 \\ &+ \frac{1}{8}\left[F(x)\right]x^3 \end{aligned}$$

Rearrangement of the terms leads to

$$\begin{aligned} &\left(1 - x + \frac{x^2}{4} - \frac{x^3}{8}\right)F(x) \\ &= \frac{x}{8} + \frac{x^2}{8} + \frac{3x^3}{32} - \frac{x^3}{8} - \frac{x^2}{8} + \frac{x^3}{32} = \frac{4x}{32} \end{aligned}$$

and, finally we obtain

$$F(x) = \frac{x}{8 - 8x + 2x^2 - x^3}.$$

Taking derivative, we obtain

$$F'(x) = \frac{(8 - 8x + 2x^2 - x^3) - x(-8 + 4x - 3x^2)}{(8 - 8x + 2x^2 - x^3)^2}.$$

Thus the expected value of the index at which the pattern **2** appears for the first time is

$$F'(1) = 8.$$

Using this expected value, a statistical test can be defined to judge whether the first appearance of the pattern, say 101 in the sequence under consideration is too late or too early or as expected.

7. Conclusion

In this work we introduced a new approach to randomness test based on the overlapping blocks, using graph theory. We give all details, including box bounds for χ-square goodness of fit test, for the pattern 010 and for the other patters, explained how to generalize. Finally we computed the expected value again for the pattern 010, and explained how to generalize to other patterns. As the theorems proven in this paper can easily be generalized to patterns of longer size, as a future work, we plan to extent this study and define randomness tests for longer patters.

References

[1] M. Uğuz, "Kirptografide rastgelelik," in *Siber Güvenlik ve Savunma: Blokzincir ve Kriptoloji*, Ş. Sağıroğlu and S. Akleylek, Eds. Nobel Akademik Yayıncılık Eğitim Danışmanlık Tic. Ltd. Şti, 2021, vol. 5, pp. 311–346.

[2] M. Uğuz, A. Doğanaksoy, F. Sulak, and O. Koçak, "R-2 composition tests: a family of statistical randomness tests for a collection of binary sequences," *Cryptography and Communications*, vol. 11, pp. 921–949, 2019.

[3] F. Sulak, "New statistical randomness tests: 4-bit template matching tests," *Turkish Journal of Mathematics*, vol. 41, no. 1, pp. 80–95, 2017.

[4] P. M. Alcover, A. Guillamón, and M. d. C. Ruiz, "A new randomness test for bit sequences," *Informatica*, vol. 24, no. 3, pp. 339–356, 2013.

[5] K. Hamano and H. Yamamoto, "A randomness test based on t-codes," in *2008 International Symposium on Information Theory and Its Applications*. IEEE, 2008, pp. 1–6.

[6] K. Hamano, F. Sato, and H. Yamamoto, "A new randomness test based on linear complexity profile," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 92, no. 1, pp. 166–172, 2009.

[7] K. Hamano and H. Yamamoto, "A randomness test based on t-complexity," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 93, no. 7, pp. 1346–1354, 2010.

[8] V. Katos, "A randomness test for block ciphers," *Applied mathematics and computation*, vol. 162, no. 1, pp. 29–35, 2005.

[9] U. M. Maurer, "A universal statistical test for random bit generators," *Journal of cryptology*, vol. 5, pp. 89–105, 1992.

[10] D. E. Knuth, *The art of computer programming*. Pearson Education, 1997, vol. 3.

[11] A. Ruhkin, "Testing randomness: A suite of statistical procedures," *Theory of Probability & Its Applications*, vol. 45, no. 1, pp. 111–132, 2001.

[12] G. Marsaglia, "Random number cdrom including the diehard battery of tests of randomness," Accessed June. 28, 2023, 1995.

[13] C. William, "Cryptx package documentation," Accessed June. 28, 2023, 1992. [Online]. Available: https://metacpan.org/dist/CryptX

[14] P. L'ecuyer and R. Simard, "Testu01: Ac library for empirical testing of random number generators," *ACM Transactions on Mathematical Software (TOMS)*, vol. 33, no. 4, pp. 1–40, 2007.

[15] L. Bassham, A. Rukhin, J. Soto, J. Nechvatal, M. Smid, S. Leigh, M. Levenson, M. Vangel, N. Heckert, and D. Banks, "A statistical test suite for random and pseudorandom number generators for cryptographic applications," Accessed June. 28, 2023, 2010-09-16 2010. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=906762

[16] J. Soto and L. Bassham, "Randomness testing of the advanced encryption standard finalist candidates," Accessed June. 28, 2023, 2000. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir6483.pdf

[17] F. Sulak, A. Doğanaksoy, M. Uğuz, and O. Koçak, "Periodic template tests: A family of statistical randomness tests for a collection of binary sequences," *Discrete Applied Mathematics*, vol. 271, pp. 191–204, 2019.