۵

A New Statistical Test Based on Linear Complexity Profile (LCP)

M. Dakhil-Alian
Ph.D.
Department of Electrical and Computer Engineering , Isfahan University of Technology

M. R. Aref
Professor
Department of Electrical Engineering , Sharif University of Technology

B. Sadeghian
Assistant Professor
Department of Computer Engineering , Amirkabir University of Technology

M. Modarres
Ph.D. Student
Department of Electrical Engineering , Sharif University of Technology

Abstract

Pseudo-random sequences are widely used in many applications such as stream ciphers. To evaluate the randomness of sequences, statistical tests are usually applied. In this paper we explain the linear complexity behavior of binary i.i.d random sequences. Linear complexity profile (LCP) of truly random sequence typically looks like an irregular staircase. Accordingly, we give a probabilistic model for height of stairs in the LCP. By using the model we present a new chi-square statistical test that is easily implemented and in comparison with standard statistical tests, has a good performance.

KeyWords

Statistical test, linear complexity, binary random variables, stream cipher.

Introduction

Stream cipher utilizes pseudo-random sequences (running key sequences) to encipher messages. The running key sequences have good statistical properties. In other word, every section of the sequences looks like as a sequence generated by Binary Symmetric Source (BSS). A BSS is a source that independently generates zero or one with a same probability. According to typical behavior of a random sequence, that is generated by BSS or fair coin tossing, Golomb proposed three requirements to measure the randomness of a periodic binary sequence [1]. Every sequence which satisfies the requirements is called pseudo-random (PN) sequence.

Linear Feedback Shift Registers (LFSR) with primitive polynomials can generate PN sequences. Such sequence have good statistical properties, but they are highly predictable. To reduce this defect, the running key generators employ nonlinear transformations. A useful measure unpredictability is provided by associated linear complexity. Thus running key sequence must have high linear complexity (necessary condition). In fact the employment of nonlinear transformations increase linear complexity and unpredictability. But high linear complexity is not sufficient condition. For example linear complexity of the following sequence with n bits is equal to n, but it has no good statistical behavior to be used in stream cipher systems.

$$S^n = 0, 0, 0, 0, \ldots, 0, 0, 1 \qquad (1)$$

Moreover of high linear complexity, the linear complexity must irregularly increase with respect to the length of sequence. In theory, a good random sequence should have a linear complexity profile (LCP) which follows closely, but irregularity the n/2 line (where n is the number of sequence bits) [2].

In practice, many different tests are carried on the sequences to evaluate its randomness. The tests divide into two groups, i.e., complexity tests and statistical tests. Complexity tests evaluate how long of generated sequence is required to reconstruct the whole sequence. The statistical tests evaluate whether the sequences, generated by running key generator, performs according to a sepecific probabilistic model. If it does, it is evaluated as a good generator. For thorough discussion of these models, the interested reader is referred to [3] and [4].

This paper is an attempt to design a practical statistical test based on the idea of irregularity of linear complexity profile. To explain our test, we will present behavior of linear complexity of a sequence $S^n = S_0, S_1, \ldots, S_{n-1}$ of independent and identically distributed (i.i.d) binary random variables. Accordingly, we derive a specific probabilistic model for the sequences and define our new test.

2 - Linear Complexity of Binary Random Sequences

For an i.i.d random sequence. it is impossible to predict one bit from all previous bits. An approach to definition of randomness in these sequences is based on unpredictability [5]. In this approach, a finite sequence is described by the length of the shortest Turing machine program that could generate the sequence. In another approach, instead of computational model such as Turing machine, we use a LFSR model and measure the unpredictability of a finite sequence by length of the shortest LFSR that is able to generate the sequence [6]. The length of shortest LFSR is referred to linear complexity of the sequence. the LFSR may be found by Berlekamp-Massey algorithm [7].

Let $S^n = S_0, S_1, \ldots, S_{n-1}$ denote a sequence of i.i.d binary random variables and let $\Lambda(S^n)$ be its linear complexity value. To evaluate the linear complexity of a random sequence, Berlekamp Massey algorithm can be used to make LCP of the sequence. Fig. 1 shows the LCP of a fair coin tossing sequence, which is derived by the algorithm. The typical dynamic behavior of i.i.d random sequence also resembles this figure.

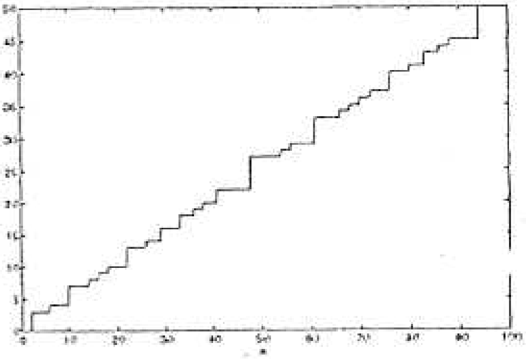

Fig. (1) LCP of the fair coin tossing sequence

To express the behavior of i.i.d binary random sequences, the following theorem were proved by R.A. Ruppel.

Theorem 1: The expected value and variance of the linear complexity of a sequence $S^n = S_0, S_1, ..., S_{n-1}$ of n i.i.d binary random variables is given by [2]:

$$E(\Lambda(S^n)] = \frac{n}{2} + \frac{4 + R_2(n)}{18} - 2^{-n}\left(\frac{n}{3} + \frac{2}{9}\right) \qquad (2)$$

and,

$$Var(\Lambda(S^n)] = \frac{86}{81} - 2^{-n}\left(\frac{14 - R_2(n)}{27}n + \frac{82 - 2R_2(n)}{81}\right.-$$

$$\left.2^{-2n}\left(\frac{n^2}{9} + \frac{4n}{27} + \frac{4}{81}\right)\right) \qquad (3)$$

Where $R_2(n)$ denotes the remainder when n is divide by 2.

From viewpoint of this theorem we expect a typical random sequence to have associated a typical linear complexity profile closely the n/2 line. Moreover for large n, the variance of several random sequences such as S^n approach to 86/81.

There are sequences whose LCP are very close to the n/2 line but they do not have their desired statistical properties. For example sequences such as $S^{\infty} = S_0, S_1, S_2, ...$ generated by Eq. (4) have perfect linear complexity profile with undesired statistical properties [8].

$$\begin{cases} S_0 = 1 \\ S_{2i} = S_{2i-1} + S_{i-1} \end{cases} \qquad (4)$$

The sequence S^n = 101000100000001... is one of these sequences whose LCP is plotted in Fig. 2.

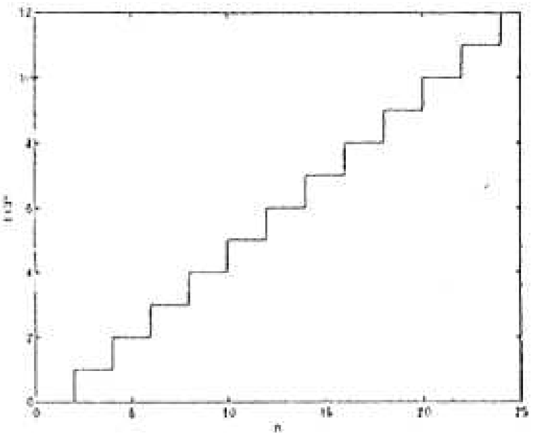

Fig. (2) The LCP of S^n = 10100010000001...

Theroem2: If $S^{\infty} = S_0, S_1, S_2, ...$ denotes of i.i.d binary random variables and if $\Lambda(S^n)$ = L, average number of sequence bits that must be processed until the length change occurs is given by [2]:

$$E[W/\Lambda(S^n) = L] = \begin{cases} 2 & \text{if} \quad L \leq \frac{n}{2} \\ 2 + 2L - n & \text{if} \quad L \leq \frac{n}{2} \end{cases} \qquad (5)$$

Moreover the average length change is:

$$E[H/\Lambda(S^n) = L] = \begin{cases} 2 & \text{if} \quad L \geq \frac{n}{2} \\ 2 - 2L + 2 & \text{if} \quad L < \frac{n}{2} \end{cases} \qquad (6)$$

This theorem says that the LCP of an i.i.d binary random sequence will resemble to irregular staircase with average height and length which agree to Eq. (5) and Eq. In next section we derive a probabilistic model for ideal LCP and present a new test based on the model.

3- Probabilistic Model for LCP

Let $S^{\infty} = S_0, S_1, S_2, ...$ denote a sequence of i.i.d binary random variables. From Eq. (2), the expected value of linear complexity of S^{∞} is close to the n/2 line. On the other hand, the LCP of the sequence typically looks like an irregular staircase. If $H^{\infty}=h_0, h_1, h_2, \ldots$ denote the sequence of height stairs in LCP of S^{∞}, then $h_i (i = 0, 1, 2, \ldots)$ will be a random variable that statisfies the following conditional probability [9]:

$$P(h = m/h \neq 0) = \frac{1}{2^m} \quad m = 1, 2, 3, \ldots \tag{7}$$

By using Eq. (7), we can find the expected value of non zero h as:

$$E(h/h \neq 0) = \sum_{m=1}^{\infty} mP(h = m/h \neq 0) = 2 \tag{8}$$

To derive a probabilistic model for LCP of random sequence, we will prove the following theorem:

Theorem 3: If $S^{\infty}=S_0, S_1, S_2, \ldots$ denote a sequence of i.i.d binary random variables, then the probabilisitic model for the height of the stairs in the LCP of S^{∞} is:

$$P(h= m) = \begin{cases} \frac{3}{4} & \text{if} \quad m = 0 \\ \frac{1}{2^{m+2}} & \text{if} \quad m = 1, 2, 3, \ldots \end{cases} \tag{9}$$

Proof:

Let $H^{m+n}=h_0, h_1, \ldots, h_{n-1}, h_n, \ldots, h_{n+m-1}$ be the LCP height stairs sequence of $S^{m+n}=S_0, S_1, \ldots, S_{n-1}, S_n, \ldots, S_{n+m-1}$. The expected value of linear complexity of S^{m+n} is

$$E[\Lambda(S^{m+n})] = E(h_0 + h_1 + \ldots + h_{n-1} + h_n + \ldots + h_{n+m-1}) \tag{10}$$

or equivalently,

$$E[\Lambda(S^{m+n})] = E[\Lambda(S^n)] E(h_0 + h_{n+1} + \ldots + h_{n+m-1}) \tag{11}$$

(where $S^n = S_0, S_1, S_2, \ldots, S_{n-1}$.)

For large n and using Eq. (2), we obtain:

$$\frac{m+n}{2} + \frac{4+R_2(m+n)}{18} - \varepsilon_{m+n} = \frac{n}{2} + \frac{4+R_2}{18} - \varepsilon_n + mE(h) \tag{12}$$

Where $\varepsilon_{m+n}, \varepsilon_n$, are equal to $2^{-m-n}(\frac{m+n}{3} - \frac{2}{9})$ and $2^{-n}(\frac{n}{3} - \frac{2}{9})$ respectively.

By using equation $E(h) = E(h/h \neq 0)\, p(h \neq 0)$, we have:

$$P(h \neq 0) = \frac{1}{4} + \frac{R_2(m+n) - R_2(n) - \varepsilon_{m+n} - \varepsilon_n}{36m} \tag{13}$$

Thus for large m and n, we can write $p(h \neq 0)=1/4$ and therefore:

$$P(h = 0) = \frac{3}{4} \tag{14}$$

On the other hand by using Eq. (7) and (14), we complete the proof, i.e.,

$$P(h = m) = P(h = m / h \neq 0)\, P(h \neq 0) + P(h = m/h = 0)\, P(h = 0) \quad m = 1, 2, 3, \ldots \tag{15}$$

and therefore,

$$P(h = m) = \frac{1}{2^{m+2}}, \quad m = 1, 2, 3, \ldots \tag{16}$$

According to this theorem, if H^n be the LCP height stairs sequence of the S^n, then we expect that about 3/4 of element of H^n equal to zero, 1/8 of them equal to one, 1/16 of them equal to two and so forth. Thus a sequence has the desired behavior when the statistics of stairs in it's LCP are close to Eq. (9).

4-Goodness of Fit Test

Let $S^n = S_0, S_1, S_2, \ldots, S_{n-1}$ denote a sequence of binary random variables with identically probability function as:

$$P(s_i = a_j) = P_j \quad i = 0, 1, \dots, n-1, \; j = 0, 1, \dots, m-1 \tag{17}$$

To evaluate whether a sequence such as S" conform with the probability function in Eq. (17), we apply chi- square test. The test parameter X^2 is defined as

$$\chi^2 = \sum_{j=0}^{m-1} \frac{(N_j - np_j)}{np_j} \tag{18}$$

Where Nj denote the number of aj in the S".

When n approaches to infinite values , the probability distribution function of χ^2 will be independent of P_j and in this case we have [10]:

$$P(\chi^2 \le k) = \int_0^k \frac{2^{-\frac{(m-1)}{2}} y^{-\frac{(m-3)}{2}} e^{\frac{y}{2}}}{\Gamma(\frac{m-1}{2})} dy \tag{19}$$

where (.) denotes the gamma function.

In chi-square test, we have a probabilistic model such as Eq. (17) and a sample sequence which we want to evaluate the compatibility of it with the model. For implement of the test we must determine significant level (or confidence interval). In the test, first the occurrence number of a_j (j=0, 1, .., m-1) in the sequence is counted (i.e., N_j) and χ^2 is computed by using Eq. (18). If χ^2 value is no greater than the threshold that is determined by significant level and degree of freedom, the sequence is passed the test. This means that the sequence is compatible with the model. The threshold value is extract from Chi-square table with m-1 degree of freedom. It is note worthy that for implementation of the test, the following relation should be satisfied [10]:

$$\forall i : np_i > 5 \tag{20}$$

5 - The New Statistical Test

For a given sequence such as S^n, the sequence H^n is found by Belekamp-Massey algorithm. For large n, we expect that the statistical behavior of the H^n, is approximated by Eq. (9). Since all of hi (i=0, 1, .., n-1) have the same distribution, to evaluate whether the H^n is generated according to the probability model presented in Eq. (9), we can apply a Chi-square test. As in section 4, for this purpose we can use Eq. (18) i.e.,

$$\chi^2 = \frac{(Nh_0 - \frac{3n}{4})^2}{\frac{3n}{4}} + \sum_{i=1}^{m} \frac{(Nh_i - \frac{n}{2^{i+2}})^2}{\frac{n}{2^{i+2}}} \tag{21}$$

Where Nh_i (i=0, 1, .., m) denote the number of i in the H^n.

χ^2 in Eq. (21)has m degree of freedom.

For a specific significant level, We can find the threshold from Chi-square table. Thus for a given sequence, by using Eq. (21) we compute χ^2 value and compare it to the threshold. If χ^2 is less than the threshold, the sequence is passed the test and otherwise is rejected. Since length of the sequence is considered n bits, in practice, according to Eq. (20), m should be:

$$m < \log_2 \left(\frac{n}{20}\right) \tag{22}$$

Example: Let $S^L = S_0, S_1, \dots, S_{L-1}$ be the sequence that generated by:

$$S_n = \begin{cases} 1 & \text{if } n = 2^j - 1 \\ 0 & \text{otherwise} \end{cases} \quad j = 0, 1, 2, \dots \tag{23}$$

This sequence has a perfect linear complexity profile but no good statistical behavior. To perform the test at first we must find Nhj in the H'. For simplification suppose n be equal to 2^L. Since the sequence has perfect LCP, then we have

$$Nh_i = \begin{cases} 2^{L-1} & \text{if } i = 0.1 \\ 0 & \text{otherwise} \end{cases} \qquad (24)$$

By using Eq. (21), and (22), we obtain

$$\chi^2 = 2^{L-2}\left(\frac{5}{6} + \sum_{i=2}^{L-5} \frac{1}{2^i}\right) > L-5 \qquad (25)$$

We consider, by %95 confidence interval (%5 significant level), the χ^2 is always greater than the threshold. Therefore the sequence is rejected by the test. This result is desirable.

We applied our test to evaluate the sequences that were generated by pseudorandom generators [11]. The results of applying this statistical test, on these sequences show that our test gives a measure to evaluate the randomness of sequences, while the measure can exclude patterned sequences passed the LCP test.

Conclusions

In this paper we have presented a new statistical test that evaluates the sequences from viewpoint of linear complexity profile of i.i.d binary random sequences. For this purpose, the behavior of random sequences has been considered and we have derived a probabilistic model for height stairs in LCP of the sequence. In Eq . (9) we have shown that P)h=0)=3/4. It is impossible to predict the next bit of the sequence with probability greater than 1/2, since the minimal length polynomial for a binary sequence is not unique. When we apply the Berlekamp-Massey algorithm we obtain one of the minimal length polynomials for the sequence. thus if we use this polynomial of a LFSR, the LFSR may or may not generate the next bit of the sequence. In other word, the ambiguity in next bit is equivalent to ambiguity of minimal polynomial.

To implement the test, we have used Eq. (9) to define the Chi-square test. Our experiments show that this test has good performance and it can easily indictae the defect of undesirable sequences. Results of many experiments also show that when a sequence passes the new test, it also passes standard statistical tests such as frequency, serial, poker and run test.

References

[1] S. W. Golomb. Shift Register Sequence, Aegean Park Press, Laguna Hill, Californian, (1982).

[2] R. A. Rueppel, Analysis and Design of Stream Cipher, Springer-Verlag, Berlin, (1986).

[3] H. Baker and F. Paper, Cipher Systems: The Protection of Communication, London, Northwood Books, (1982).

[4] U. Maurer, "A Universal Statistical Test for Random bit Generator", Journal of Cryptology, Vol. 5, No. 2, pp. 89-105, (1992).

[5] A. N. Kolmogrov, "Three Approaches to the Quantitative Definition of Information" Problemy Predachi Inform. Transmission, Vol. 1, No. 1, pp. 3-11, (1965).

[6] A. Lempel, J. Ziv," On the complexity of Finite Sequences", IEEE, Trans. On Information Theory, Vol. IT-22, Jan (1967).

[7] J.L. Massey", Shift Register and BCH Decoding", IEEE, Trans. On Information Theory, Vol. 15, No. 1, pp. 122-127, (1969).

[8] H. Niederriter," *Sequence with Almost Perfect* Linear Complexity Profile", Springer-Verlag, Advance in Cryptology, Eourocrypt 87, pp. 37-52, (1987).

[9] H. Niederriter," The probabilistic theory of Linear Complexity Profile", Springer-Verlag. Advance in Cryptology, Eourocrypt 88, pp. 191-210 (1988).

[10] G. H. Larson, Introduction to Probability and Statistical Inference, John-wiley, (1974).

[11] M. Dakhil-alian, "Evaluation and Design of Pseudo-random Sequences and Chaotic Generators, Ph.D. thesis, Isfahan University of Tecnology, Isfahan Iran (1999).