J. Cryptology (1991) 3: 201–212

Journal of Cryptology
© 1991 International Association for Cryptologic Research

A Generalized Correlation Attack on a Class of Stream Ciphers Based on the Levenshtein Distance[1]

Jovan Dj. Golić and Miodrag J. Mihaljević
Institute of Applied Mathematics and Electronics, Belgrade,
Faculty of Electrical Engineering, University of Belgrade,
Bulevar Revolucije 73, 11001 Beograd, Yugoslavia

Abstract. A statistical approach to cryptanalysis of a memoryless function of clock-controlled shift registers is introduced. In the case of zero-order correlation immunity, an algorithm for a shift register initial state reconstruction based on the sequence comparison concept is proposed. A constrained Levenshtein distance relevant for the cryptanalysis is defined and a novel recursive procedure for its efficient computation is derived. Preliminary experimental results are given and open theoretic problems are discussed.

Key words. Cryptanalysis, Clock-controlled shift registers, Correlation attack, Sequence comparison, Levenshtein distance, Algorithms.

1. Introduction

Clock-controlled shift registers have become popular building blocks for key-stream generators. Schemes with clock-controlled shift registers are proposed that ensure large lower bounds on period and linear complexity, and possess no obvious flaws in statistical behavior. On the other hand, irregular clocking reduces the danger from correlation attacks. A review of the clock-controlled shift registers is presented in [3].

In this paper the security of a key-stream generator structure consisting of clock-controlled shift registers combined by a memoryless function is considered, see [2]. In the binary case, when the registers are clocked regularly, and the function is zero-order correlation immune[2] (its output is correlated to at least one input), Siegenthaler [12] introduced a ciphertext-only correlation attack based on the Hamming distance measure. He showed that it is possible to reconstruct the initial state of any register whose output is correlated to the generator output. However,

[1] Date received: March 25, 1990. Date revised: December 6, 1990.

[2] Following [11], a Boolean function $f(x_1, \ldots, x_n)$ is said to be mth-order correlation immune if m is the maximum integer such that the random variable $f(X_1, \ldots, X_n)$ is statistically independent of every set of m random variables chosen from the balanced and independent binary random variables $X_1, \ldots, X_n$.

when the registers are clocked irregularly the Hamming distance is useless and, hence, the correlation atack is no longer applicable. The main contribution of this paper is that we show that a statistical correlation attack is still feasible, but with an appropriately defined constrained Levenshtein distance instead of the Hamming one. The concept of the Levenshtein and related distances is known in the area of string editing (see, for example, [4] and [7]–[10]) with main applications in text correcting, decoding, and molecular biology.

A statistical model and the statement of the problem are given in Section 2. A basic idea for the generalized correlation attack is presented in Section 3, whereas the constrained Levenshtein distance and relevant probability distributions are, in general, defined in Section 4. In Section 5 we first provide a mathematically precise definition of the constrained Levenshtein distance and then establish a theorem which enables efficient recursive computation of it. The proof of the theorem is given in the Appendix. A complete algorithm for the initial state determination is proposed in Section 6 together with some illustrative numerical results. In Section 7 a summary of the results and a list of still open theoretic problems are given.

2. Statistical Model

In principle, the shift registers may be clocked arbitrarily. For simplicity we assume that a shift register whose output is correlated to the generator output is one–two clocked by another shift register. Without loss of generality we also assume that the shift registers have linear feedback. The corresponding statistical model is shown in Fig. 1.

Let $\{x_n\}$ be a binary sequence produced by a linear feedback shift register (LFSR) defined by

$$x_n = \sum_{l=1}^{L} c_l x_{n-l}, \qquad n = 0, 1, \ldots, \tag{1}$$

where $f(X) = \sum_{l=0}^{L} c_{L-l} X^l$, $c_0 = 1$, is the LFSR characteristic polynomial and $\mathbf{X}_0 = [x_{-l}]_{l=1}^{L}$ is a nonnull LFSR initial state. Let $\{a_n\}$ be the output of another linear feedback shift register. A decimation box output is for simplicity defined by

$$y_n = x_{f(n)}, \qquad f(n) = n + \sum_{j=1}^{n} a_j, \qquad n = 0, 1, 2, \ldots. \tag{2}$$

In the statistical model, $\{a_n\}$ is regarded as a realization of the sequence

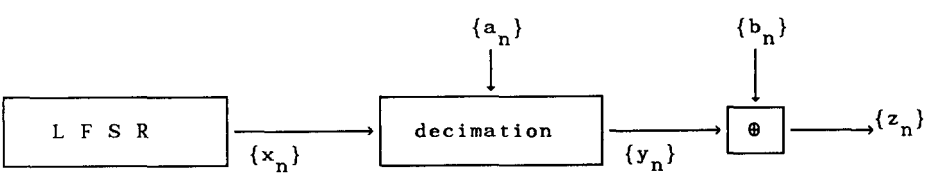

Fig. 1. A noisy clock-controlled shift register structure.

of independent identically distributed (i.i.d.) binary variables $\{A_n\}$ such that $\Pr(A_n = 1) = 0.5$ for every n. A binary noise sequence $\{b_n\}$ is a realization of a sequence of i.i.d. binary variables $\{B_n\}$ such that $\Pr(B_n = 1) = p \neq 0.5$ for every n, where p is the cross-correlation parameter, which may involve plaintext statistics as well [12]. Finally, a binary sequence $\{z_n\}$ is defined as the sum modulo 2 of the decimated and noise sequences

$$z_n = x_{f(n)} \oplus b_n, \qquad f(n) = n + \sum_{j=1}^{n} a_j, \qquad n = 0, 1, 2, \ldots. \tag{3}$$

In this statistical model we consider the problem of the initial state $(\mathbf{X}_0 = [x_{-l}]_{l=1}^{L})$ reconstruction assuming that $f(X)$, p, and a segment $\{z_n\}_{n=1}^{N}$ are known.

3. Generalized Correlation Attack

A correlation attack [12] is based on the Hamming distance between two binary sequences of the same length. Obviously, the same statistical approach cannot be applied here. However, suppose we defined a suitable distance measure d between two binary sequences of different length, which reflects the transformation of the LFSR sequence $\{x_n\}$ to the output sequence $\{z_n\}$ according to the model displayed in Fig. 1. Then we could proceed along essentially the same lines as in [12], thus establishing a statistical procedure which we call a generalized correlation attack.

By the assumed statistical model, each $\mathbf{X}_0$ gives rise to a conditional probability distribution on the set of all binary sequences $\{z_n\}_{n=1}^{N}$. We thus have a pattern recognition system with $2^L - 1$ classes corresponding to all the nonnull initial states of the LFSR. Given an observed segment $\{z_n\}_{n=1}^{N}$, the optimal decision strategy (yielding the minimum probability of decision error) is to decide on the initial state with the maximum posterior probability. When the LFSR is regularly clocked, as in [12], it is optimal to decide on the initial state $\hat{\mathbf{X}}_0$ such that the Hamming distance between $\{z_n\}_{n=1}^{N}$ and $\{\hat{x}_n\}_{n=1}^{N}$ is minimum (a sufficient statistic). However, when the LFSR is clocked irregularly it is not clear how to find an optimum decision rule. Anyway, given an appropriate distance measure, we can define a decision procedure that is close to being optimal.

Let $\{\hat{x}_n\}_{n=1}^{M}$ be the LSFR sequence corresponding to an initial state $\hat{\mathbf{X}}_0$. The choice of the length M, $N \leq M \leq 2N + 1$, is discussed later. Let d be the distance between $\{\hat{x}_n\}_{n=1}^{M}$ and $\{z_n\}_{n=1}^{N}$. The following two hypotheses are possible:

H_0: The observed sequence $\{z_n\}_{n=1}^{N}$ is produced by $\hat{\mathbf{X}}_0$.
H_1: The observed sequence $\{z_n\}_{n=1}^{N}$ is not produced by $\hat{\mathbf{X}}_0$.

Consequently, d can be considered as an outcome of a random variable D with two possible probability distributions (statistically averaged over the ensemble of all the initial states): $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$. How to determine or estimate these distributions is discussed in the next section. Suppose that they are known. Note that they depend on N, assuming that $M = M(N)$ is given as a function of N. First determine the threshold t and length N so as to achieve the given probabilities of

"the missing event" P_m and "the false alarm" P_f. As in [12], P_m is chosen close to zero (e.g., 10^{-3}) and P_f is picked very close to zero, $P_f \cong 2^{-L}$, so that the expected number of false alarms is very small ($\cong 1$). Then the decision procedure goes through the following steps, for every possible initial state $\hat{\mathbf{X}}_0$:

- generate $\{\hat{x}_n\}_{n=1}^M$;
- calculate the distance d between $\{\hat{x}_n\}_{n=1}^M$ and $\{z_n\}_{n=1}^N$;
- according to the threshold t accept H_0 or H_1.

The output of the procedure is the set of the most probable candidates for the true initial state.

4. Distance Measures and Relevant Probability Distributions

The distance measure should be defined so that it enables statistical distinction between the two cases: first, when $\{\hat{x}_n\}_{n=1}^M$ and $\{z_n\}_{n=1}^N$ are picked at random, uniformly and independently (which is a reasonable model for H_1), and, second, when $\{z_n\}_{n=1}^N$ is obtained from $\{\hat{x}_n\}_{n=1}^M$ according to the model shown in Fig. 1, that is, by the deletion of some bits subject to the decimation constraints and by complementation of the remaining ones with probability p. This problem is a special case of the comparison problem between two sequences when one sequence is obtained from the other by symbol substitution, deletion, and insertion, which is extensively studied in the literature. For example, the sequence-matching problem is considered in coding theory (see, for example, [4]) and text processing (see, for example, [8]). A review of the sequence-matching techniques and applications is presented in [10].

According to [10], one of the widely used distances is the Levenshtein distance [4]. Let the edit operations that transform one sequence into another be substitution, deletion, and insertion. Then the Levenshtein distance between two sequences is defined as the minimum number of edit operations required to transform one sequence into the other. The various extensions of the basic Levenshtein distance are proposed in the literature. For our problem, the constrained Levenshtein distance concept [7] is relevant because of the constraints inherent to the decimation function. In [7]–[9] an efficient algorithm for the constrained Levenshtein distance computation is proposed when the constraints relate to the total number of deletions, insertions, and substitutions, respectively.

We basically define the distance measure between $\{\hat{x}_n\}_{n=1}^M$ and $\{z_n\}_{n=1}^N$ as the minimum number of deletions, subject to the constraint on the maximum number of consecutive deletions, and complementations required to obtain $\{z_n\}_{n=1}^N$ from $\{\hat{x}_n\}_{n=1}^M$. Whether this distance is a sufficient statistic remains an open question, but it is reasonable to believe that this is approximately the case. With the Levenshtein distance defined thus, the problem is to determine the probability distributions $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$. Can it be done theoretically?

In many applications of sequence comparison it is important to decide whether two sequences are mutually dependent or independent, on the basis of a given distance measure. One approach is a nonparametric estimation of the relevant

probability distributions. The other is analytical consideration of some characteristics of the underlying probability distributions. The expected degree of similarity is the first element needed in statistical testing. Following [1], the tradition in this area is to evaluate sequence similarities in terms of the length of the longest common subsequence of the sequences processed (which is related to the Levenshtein distance). Regarding the analytical treatment of the probability distributions for random sequence matching, it is noted on p. 352 of [10] that the derivation of exact mathematical results seems difficult and many interesting questions remain unanswered: for example, one is a question concerning the variance which seems to grow surprisingly slowly with the length of sequences, though no mathematical results are known yet.

Consequently, we anticipate that the problem to determine $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$ for the constrained Levenshtein distance defined is very difficult. More promising is the question of how these two distributions (especially $\{\Pr(D|H_1)\}$) behave asymptotically when N and M tend to infinity. In this paper we therefore adopt a nonparametric method for estimating the relevant probability distributions.

5. Constrained Levenshtein Distance

In this section we introduce a general definition of the constrained Levenshtein distance relevant for the cryptanalysis and derive a recursive algorithm for its efficient computation.

Suppose we have two finite length discrete sequences $U = \{u_i\}_{i=1}^{M}$ and $V = \{v_i\}_{i=1}^{N}$ over a finite alphabet $\mathbf{A}$, and a nonnegative real function $d(u, v)$, $u \in \mathbf{A}$, $v \in \tilde{\mathbf{A}}$, $\tilde{\mathbf{A}} = \mathbf{A} \cup \theta$, where θ stands for the null symbol associated with the deletion operation. Accordingly, $d(u, \theta)$, $u \in \mathbf{A}$, denotes the elementary edit distance associated with deleting a symbol u and $d(u, v)$, $u, v \in \mathbf{A}$, denotes the elementary edit distance associated with substituting a symbol v for a symbol u. We consider a problem of transforming U to V using the edit operations of deletion and substitution.

Definition. The constrained Levenshtein distance (CLD) between $\{u_i\}_{i=1}^{M}$ and $\{v_i\}_{i=1}^{N}$ is the minimum sum of elementary edit distances associated with the edit operations of deletion and substitution required to obtain $\{v_i\}_{i=1}^{N}$ from $\{u_i\}_{i=1}^{M}$ under the constraint that the maximum number of consecutive deletions is E.

Note that the constraints on the total number of consecutive deletions $(M - N)$ and substitutions (N) are inherent to the definition. Also, a necessary condition to be satisfied is

$$N \leq M \leq (E + 1)N + E. \tag{4}$$

Let $V' = \{v_i'\}_{i=1}^{N}$ denote an arbitrary sequence over $\tilde{\mathbf{A}}$ such that by deleting all the null symbols from V' we get V. Accordingly, every edit transformation of U to V can be uniquely represented by the two-dimensional edit sequence $(U, V') = \{(u_i, v_i')\}_{i=1}^{M}$: if $v_i' = \theta$, then u_i is deleted and if $v_i' \neq \theta$, then v_i' is substituted for u_i, for any $i = 1, 2, \ldots, M$. Let G_{UV} be the set of all possible edit sequences that transform

U to V, subject to the condition that there are no more than E consecutive null symbols in V'. The CLD can then be expressed by

$$D(U, V) = \min\left\{\sum_{i=1}^{M} d(u_i, v_i'): (U, V') \in G_{UV}\right\}. \tag{5}$$

In order to derive an efficient CLD computation procedure, we introduce the partial CLD $W(e, s)$ as the CLD between a prefix $U_{e+s} = \{u_i\}_{i=1}^{e+s}$ of U and a prefix $V_s = \{v_i\}_{i=1}^{s}$ of V, under the same constraints. Using an abbreviated notation $G_{es} = G_{U_{e+s}V_s}$ we have

$$W(e, s) = \min\left\{\sum_{i=1}^{e+s} d(u_i, v_i'): (U_{e+s}, V_s') \in G_{es}\right\} \tag{6}$$

and

$$D(U, V) = W(M - N, N). \tag{7}$$

The set of all the permitted values for (e, s) is clearly given by

$$0 \le s \le N, \tag{8}$$

$$0 \le e \le \min\{M - N, (s + 1)E\}. \tag{9}$$

For simplicity, suppose that the elementary edit distance $d(u, \theta)$ is equal for all the symbols u from $\mathbf{A}$, and denote it by d_0.

Now we state a theorem yielding a recursive property of $W(e, s)$, which in view of (7) enables efficient computation of the CLD.

Theorem. *The partial CLD $W(e, s)$ satisfies the recursion*

$$W(e, s) = \min\{W(e - e_1, s - 1) + e_1 d_0 + d(u_{e+s-e_1}, v_s): \max\{0, e - \min\{M - N, sE\}\} \le e_1 \le \min\{e, E\}\} \tag{10}$$

for $1 \le s \le N$, $0 \le e \le \min\{M - N, (s + 1)E\}$, *and, for* $s = 0$ *and* $0 \le e \le \min\{M - N, E\}$,

$$W(e, 0) = ed_0. \tag{11}$$

The proof of the theorem is given in the Appendix. It basically relies on the principles used in [7], but has an important difference which reflects the specific constraints. As a consequence, unlike the dynamic programming expression from [7], the order of the recursion with regard to deletions is greater than one.

In this paper we consider a specific application where the alphabet is binary, the maximum number of consecutive deletions E is equal to one, and the decimation sequence is a realization of a sequence of balanced i.i.d. binary random variables. We also assume that the elementary edit distances associated with deletion and effective substitution are both equal to one. In that case, the constrained Levenshtein distance is reduced to the number of deletions and substitutions needed for the required transformation. Apart from that, given a string V of length N, the length M of the string U that actually produced V is not known, since the decimation

sequence is unknown. So, the question arises as to how to choose M given N. One possibility is to take M to be close to its expected value (in our case $\simeq 3N/2$), which minimizes the mean square error between the actual and assumed M. Another, more appropriate way is to modify the definition of the CLD so as to eliminate the constraint on the maximum number of consecutive deletions either at the beginning or the end of U, and, then, to specify M so that the probability of U being longer than M is close to zero (for example, we can assume the maximum value $(E+1)N+E$, which in our case is $2N+1$). We thus define the constrained edit distances CLD′ (CLD″) in the same way as the CLD with a difference that an arbitrary number of consecutive deletions is permitted at the beginning (end) of U. Fortunately, it appears that only a slight modification of the recursion (10) in the theorem suffices to obtain the CLD′ (CLD″). Namely, proceeding along essentially the same lines as in the proof of the theorem, we can verify that the CLD′ (CLD″) are both determined by (7), where $W(e, s)$ is given by the same theorem with the following modifications: for the CLD′, (11) holds for $0 \le e \le M - N$ and instead of (10) we have

$$W(e, s) = \min\{W(e - e_1, s - 1) + e_1 d_0 + d(u_{e+s-e_1}, v_s)\colon 0 \le e_1 \le \min\{e, E\}\}, \tag{12}$$

which holds for $1 \le s \le N$ and $0 \le e \le M - N$; and, for the CLD″, (10) holds for $1 \le s \le N - 1$ whereas, for $s = N$,

$$W(e, s) = \min\{W(e - 1, s) + d_0,\ W(e, s - 1) + d(u_{e+s}, v_s)\}, \tag{13}$$

which is Oommen's expression [7] without insertions.

In order to make a clearer distinction between the initial states that give rise to the close cyclic shifts ($\le E$) of the shift register sequence, we can impose an additional constraint that the first (for CLD, CLD″) or the last (for CLD, CLD′) edit operation is substitution. It is clear that the CLD in this case is reduced to the ordinary CLD between the original sequences without the symbols affected by the assumed substitution.

Having computed the partial CLD $W(e, s)$ for all the permitted values of (e, s), we obtain not only the CLD $W(M - N, N)$ but can also reconstruct an optimum edit sequence by backtracking through the matrix $W(e, s)$ starting from the element $W(M - N, N)$, see [7]. In general, an optimum edit sequence, which contains a reconstructed decimation sequence, is not unique.

For the cryptanalysis application we have adopted the constrained Levenshtein distance CLD′. In our case it reduces to the distance measure CLD* which can be computed recursively by the following procedure.

The Constrained Levenshtein Distance (CLD*) Computation Procedure.

1. Input: binary sequences $\{\hat{x}_n\}_{n=1}^{M}$ and $\{z_n\}_{n=1}^{N}$.
2. Initialization:

$$d(k, 0) = k, \qquad k = 0, 1, \ldots, M - N,$$

$$d(0, l) = d(0, l - 1) + (\hat{x}_l \oplus z_l), \qquad l = 1, 2, \ldots, N.$$

3. Recursive calculation for $M > N$:

$$d(k, l) = \min\{d(k-1, l-1) + (\hat{x}_{k+l-1} \oplus z_l) + 1, d(k, l-1) + (\hat{x}_{k+l} \oplus z_l)\},$$
$$l = 1, 2, \ldots, N, \quad k = \max\{1, M - 2N + l\}, \ldots, M - N.$$

4. Output: the CLD* between $\{\hat{x}_n\}_{n=1}^{M}$ and $\{z_n\}_{n=1}^{N}$: $d^* = d(M - N, N)$.

The time and space complexities of the procedure are both quadratic $O(N(M-N))$. Note that the recursion is calculated only for $k = \max\{1, M - 2N + l\}, \ldots, M - N$, instead of $k = 1, 2, \ldots, M - N$, because in order to obtain $d(M - N, N)$ we need not calculate $d(k, l)$ for all the permitted values of (k, l). On the other hand, the space complexity can be reduced to a linear one $O(M - N)$, since we need not memorize the whole matrix for $d(k, l)$, but only the vectors $(d(0, l), d(1, l), \ldots, d(M - N, l))$ and $(d(0, l + 1), d(1, l + 1), \ldots, d(M - N, l + 1))$ which are computed recursively for $l = 1, 2, \ldots, N - 1$.

6. Algorithm and Experimental Results

In this section we propose a cryptanalytic algorithm for the clock-controlled shift register initial state reconstruction which is based on the CLD* computation procedure. We also give some illustrative experimental results.

The cryptanalytic algorithm is essentially a combination of the decision procedure given in Section 3 and the CLD* computation procedure from Section 5. Since the underlying probability distributions $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$ are not known analytically, the threshold t and the sufficient length N are determined iteratively in a nonparametric manner. The length M is chosen as a function of N in a way described in Section 5 (for example, $M = 3N/2$, $M = 3N/2 + c\sqrt{N}$, or $M = 2N + 1$). The parameters assumed to be known are the shift register characteristic polynomial $f(X)$, the probability of "the missing event" P_m (e.g., $P_m = 10^{-3}$), the expected number of the solution candidates n_0 ($n_0 = 1 + (2^L - 2)P_f \cong 2^L P_f$, P_f being the probability of "the false alarm"), and the initial and increment values for N, N_0 and ΔN, respectively. The cross-correlation parameter p need not be known; without loss of generality it is only assumed that $p < 0.5$. The input to the procedure is the observed output segment $\{z_n\}$ of sufficient length. A basic form of the algorithm is as follows.

Algorithm.
Initialization: $N = N_0$ and Ω is the set of all the possible initial states.

Step 1. Generate the set of $\sim 1/P_m$ samples of $\{\Pr(D|H_0)\}$ repeating the following: Pick $\hat{\mathbf{X}}_0$ at random, generate $\{\hat{x}_n\}_{n=1}^{M}$ and $\{\hat{z}_n\}_{n=1}^{N}$ according to the assumed model, and calculate the distance applying the CLD* procedure. Choose the threshold t to be greater than the maximum distance value in the sample set. Set $n_f = 0$.

Step 2. Generate a new initial state from Ω, $\hat{\mathbf{X}}_0$, different from the previously generated ones after the last pass through Step 1. If Ω is empty, go to Step 6.

Step 3. For the assumed initial state, generate the LFSR sequence $\{\hat{x}_n\}_{n=1}^{M}$.
Step 4. Applying the CLD* procedure calculate the distance d^* between $\{\hat{x}_n\}_{n=1}^{M}$ and $\{z_n\}_{n=1}^{N}$.
Step 5. If $d^* > t$ exclude the initial state from Ω and go to Step 2.
If $d^* \leq t$ set $n_f = n_f + 1$ and then go to Step 2.
Step 6. If $n_f > n_0$ increase $N \to N + \Delta N$ and go to Step 1. Otherwise, stop the procedure.

Table 1. Estimations of $\{\Pr(D|H_i)\}$, $i = 0, 1$, for $\Delta = 10$, when $N = 4000$, $p = 0.25$.

D starting interval point	$\Pr(D\|H_0)$	$\Pr(D\|H_1)$
2210	0.000	0.000
2220	0.000	0.000
2230	0.000	0.000
2240	0.000	0.000
2250	0.000	0.000
2260	0.000	0.000
2270	0.000	0.000
2280	0.000	0.000
2290	0.000	0.000
2300	0.001	0.000
2310	0.004	0.000
2320	0.033	0.000
2330	0.103	0.000
2340	0.211	0.000
2350	0.276	0.000
2360	0.219	0.000
2370	0.109	0.000
2380	0.036	0.000
2390	0.006	0.005
2400	0.002	0.056
2410	0.000	0.213
2420	0.000	0.380
2430	0.000	0.277
2440	0.000	0.062
2450	0.000	0.007
2460	0.000	0.001
2470	0.000	0.000
2480	0.000	0.000
2490	0.000	0.000
2500	0.000	0.000
2510	0.000	0.000
2520	0.000	0.000
2530	0.000	0.000
2540	0.000	0.000
2550	0.000	0.000

Table 2. Estimations of $\{\Pr(D|H_i)\}$, $i = 0, 1$, for $\Delta = 10$, when $N = 5000$, $p = 0.25$.

D starting interval point	$\Pr(D\|H_0)$	$\Pr(D\|H_1)$
2810	0.000	0.000
2820	0.000	0.000
2830	0.000	0.000
2840	0.000	0.000
2850	0.000	0.000
2860	0.000	0.000
2870	0.000	0.000
2880	0.000	0.000
2890	0.006	0.000
2900	0.012	0.000
2910	0.036	0.000
2920	0.134	0.000
2930	0.231	0.000
2940	0.220	0.000
2950	0.210	0.000
2960	0.099	0.000
2970	0.034	0.000
2980	0.013	0.000
2990	0.003	0.001
3000	0.001	0.017
3010	0.000	0.131
3020	0.000	0.285
3030	0.000	0.296
3040	0.000	0.212
3050	0.000	0.057
3060	0.000	0.001
3070	0.000	0.000
3080	0.000	0.000
3090	0.000	0.000
3100	0.000	0.000
3110	0.000	0.000
3120	0.000	0.000
3130	0.000	0.000
3140	0.000	0.000
3150	0.000	0.000

Table 3. Estimations of $\{\Pr(D|H_i)\}$, $i = 0, 1$, for $\Delta = 10$, when $N = 10{,}000$, $p = 0.25$.

D starting interval point	$\Pr(D\|H_0)$	$\Pr(D\|H_1)$
5780	0.000	0.000
5790	0.000	0.000
5800	0.000	0.000
5810	0.000	0.000
5820	0.000	0.000
5830	0.000	0.000
5840	0.020	0.000
5850	0.080	0.000
5860	0.086	0.000
5870	0.133	0.000
5880	0.146	0.000
5890	0.260	0.000
5900	0.113	0.000
5910	0.093	0.000
5920	0.046	0.000
5930	0.020	0.000
5940	0.002	0.000
5950	0.000	0.000
5960	0.000	0.000
5970	0.000	0.000
5980	0.000	0.000
5990	0.000	0.000
6000	0.000	0.000
6010	0.000	0.003
6020	0.000	0.013
6030	0.000	0.064
6040	0.000	0.182
6050	0.000	0.231
6060	0.000	0.262
6070	0.000	0.187
6080	0.000	0.046
6090	0.000	0.013
6100	0.000	0.000
6110	0.000	0.000
6120	0.000	0.000
6130	0.000	0.000
6140	0.000	0.000
6150	0.000	0.000
6160	0.000	0.000

Table 4. Estimations of $\{\Pr(D|H_i)\}$, $i = 0, 1$, for $\Delta = 10$, when $N = 20{,}000$, $p = 0.25$.

D starting interval point	$\Pr(D\|H_0)$	$\Pr(D\|H_1)$
11670	0.000	0.000
11680	0.003	0.000
11690	0.011	0.000
11700	0.014	0.000
11710	0.037	0.000
11720	0.051	0.000
11730	0.077	0.000
11740	0.088	0.000
11750	0.125	0.000
11760	0.153	0.000
11770	0.128	0.000
11780	0.085	0.000
11790	0.072	0.000
11800	0.061	0.000
11810	0.043	0.000
11820	0.023	0.000
11830	0.019	0.000
11840	0.006	0.000
11850	0.002	0.000
11860	0.000	0.000
11870	0.000	0.000
⋮	⋮	⋮
12020	0.000	0.000
12030	0.000	0.000
12040	0.000	0.000
12050	0.000	0.000
12060	0.000	0.005
12070	0.000	0.064
12080	0.000	0.070
12090	0.000	0.123
12100	0.000	0.188
12110	0.000	0.241
12120	0.000	0.135
12130	0.000	0.082
12140	0.000	0.053
12150	0.000	0.029
12160	0.000	0.006
12170	0.000	0.000
12180	0.000	0.000

Note that the algorithm yields at most n_0 candidates for the solution. They can be arranged in order of the increasing value of the distance d^*, the smallest one corresponding to the most likely solution. It is important to observe that there will always be a certain number of possible candidates for the initial state being the close cyclic shifts of each other, meaning that the solution is inherently not quite unique.

The success of the algorithm is essentially based on the assumption, so far experimentally verified, that the increase of N gives rise to the increase of the separation between $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$.

The proposed algorithm has run successfully on a number of examples. Tables 1–4, which show the histogram estimators, with the length of elementary interval equal to Δ, illustrate the separation between $\{\Pr(D|H_0)\}$ and $\{\Pr(D|H_1)\}$.

7. Conclusion

In this paper we propose a generalized correlation attack on a zero-order correlation immune memoryless function of irregularly clocked shift registers, which is based on the sequence comparison approach. An appropriate constrained Levenshtein distance (CLD) relevant for the cryptanalysis is introduced and a novel recursive procedure for its efficient computation is derived. Some numerical examples are given to illustrate the chances of success of the algorithm.

Regarding the still open theoretic problems we emphasize the following:

- How close is the minimum CLD decision rule to the maximum posterior probability one?
- Determination or approximation of the relevant probability distributions and their asymptotic behavior when the length of the observed segment goes to infinity.
- For which decimation procedures and values of the cross-correlation parameter p and shift register length L is the statistical discrimination between H_0 and H_1, on the basis of the CLD, possible (at least asymptotically)?
- Construction of a feast generalized correlation attack on a noise clock-controlled linear feedback shift register, bearing in mind the fast correlation attacks [5], [6] corresponding to the ordinary correlation attack [12].

Appendix

Proof of the Theorem. For $s = 0$ the proof is immediate. Assume now that $s \geq 1$. Applying the dynamic programming principle [7] to (6), we also partition the set of all the permitted edit sequences $G_{e,s}$, but in an essentially different way. Namely, in order to deal with the specific constraints we divide $G_{e,s}$ according to the sequence of deletions after the last substitution. Thus, let $G_{es}^{e_1}$, $e_1 = 0, 1, \ldots, E$, be a subset of G_{es} that consists of all the edit sequences that end with exactly e_1 deletions. However, some of these subsets may be empty. It is easy to see that $G_{es}^{e_1}$ is empty if and only if the pair $(e - e_1, s - 1)$ is not permitted. Using (8) and (9) we then obtain that $G_{es}^{e_1}$ is not empty if and only if $\max\{0, e - \min\{M - N, sE\}\} \leq e_1 \leq \min\{e, E\}$.

Consequently, (6) can be put into the form

$$W(e, s) = \min\left\{\min\left\{\sum_{i=1}^{e+s} d(u_i, v_i')\colon (U_{e+s}, V_s') \in G_{es}^{e_1}\right\}\colon \right.$$
$$\left. \max\{0, e - \min\{M - N, sE\}\} \le e_1 \le \min\{e, E\}\right\}. \quad \text{(A.1)}$$

Considering that for $s \ge 1$ every edit sequence from $G_{es}^{e_1}$ ends in exactly e_1 deletions after a substitution, and that the constraint relates to the maximum number of consecutive deletions, it follows that (U_{e+s}, V_s') belongs to $G_{es}^{e_1}$ if and only if its prefix $(U_{e-e_1+s-1}, V_{s-1}')$ belongs to $G_{e-e_1,s-1}$, assuming that $(e - e_1, s - 1)$ is permitted. Hence

$$\min\left\{\sum_{i=1}^{e+s} d(u_i, v_i')\colon (U_{e+s}, V_s') \in G_{es}^{e_1}\right\}$$
$$= \min\left\{\sum_{i=1}^{e-e_1+s-1} d(u_i, v_s')\colon (U_{e-e_1+s-1}, V_{s-1}') \in G_{e-e_1,s-1}\right\} + e_1 d_0 + d(u_{e-e_1+s}, v_s)$$
$$= W(e - e_1, s - 1) + e_1 d_0 + d(u_{e+s-e_1}, v_s) \quad \text{(A.2)}$$

under the condition that $(e - e_1, s - 1)$ is permitted. Finally, (10) is a direct consequence of (A.1) and (A.2). □

References

[1] V. Chvatal, D. Sankoff, Longest common subsequences of two random sequences, *J. Appl. Probab.*, pp. 306–315, 1975.

[2] J. Dj. Golić, On the linear complexity of functions of periodic *GF(q)* sequences, *IEEE Trans. Inform. Theory*, vol. 35, pp. 69–75, Jan. 1989.

[3] D. Gollman, W. G. Chambers, Clock-controlled shift registers: a review, *IEEE J. Select. Areas Comm.*, vol. 7, pp. 525–533, May 1989.

[4] A. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Phys. Dokl.*, vol. 10, pp. 707–710, 1966.

[5] W. Meier, O. Staffelbach, Fast correlation attacks on certain stream ciphers, *J. Cryptology*, vol. 1, pp. 159–176, 1989.

[6] M. J. Mihaljević, J. Dj. Golić, A fast iterative algorithm for a shift register initial state reconstruction given the noisy output sequence, *Advances in Cryptology—AUSCRYPT '90*, Lecture Notes in Computer Science, vol. 453, pp. 165–175, Springer-Verlag, Berlin, 1990.

[7] B. J. Oommen, Constrained string editing, *Inform. Sci.*, vol. 40, pp. 267–284, 1986.

[8] B. J. Oommen, Recognition of noisy subsequences using constrained edit distance, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, pp. 676–685, Sept. 1987.

[9] B. J. Oommen, Correction to "Recognition of noisy subsequences using constrained edit distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, pp. 983–984, Nov. 1988.

[10] D. Sankoff, J. B. Kruskal, *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, Reading, MA, 1983.

[11] T. Siegenthaler, Correlation-immunity of nonlinear combining functions for cryptographic applications, *IEEE Trans. Inform. Theory*, vol. 30, pp. 776–780, Sept. 1984.

[12] T. Siegenthaler, Decrypting a class of stream ciphers using ciphertext only, *IEEE Trans. Comput.* vol. 34, pp. 81–85, Jan. 1985.