A NOISY CLOCK-CONTROLLED SHIFT REGISTER CRYPTANALYSIS CONCEPT BASED ON SEQUENCE COMPARISON APPROACH

Jovan Dj. Golić
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 cryptanalysis method for the initial state reconstruction of a noisy clock-controlled shift register using the noisy output sequence only, is proposed. The method is based on the sequence comparison approach.

1. PROBLEM STATEMENT

A review of clock-controlled shift registers is presented in [1]. A statistical model of the clock-controlled shift register structure, which is under consideration in this correspondence, is shown in Fig.1. For simplicity, we assume that the shift register whose output is correlated with the generator output is one-two clocked.

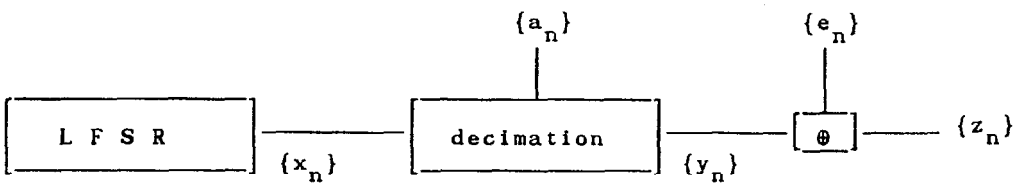

Fig.1. A model of the clock-controlled shift register structure.

A binary sequence $\{x_n\}$ is the output of a linear feedback shift register (LFSR) with characteristic polynomial $f(X)=\Sigma_{\ell=0}^{L} c_{L-\ell} X^{\ell}$, $c_0=1$, and $X_0=[x_{-\ell}]_{\ell=1}^{L}$ is the LFSR initial state. For example, a decimation sequence $\{a_n\}$ is the output of another binary shift register. The decimation box output is defined by $y_n = x_{f(n)}$, $f(n) = n + \Sigma_{j=1}^{n} a_j$, $n=0,1,2,\ldots$.

I.B. Damgard (Ed.): Advances in Cryptology - EUROCRYPT '90, LNCS 473, pp. 487-491, 1991.

In the statistical model, $\{a_n\}$ is regarded as a realization of the sequence of i.i.d. binary variables $\{A_n\}$ such that $Pr(A_n=1) = 0.5$ for every n. A binary sequence $\{e_n\}$ is a realization of a sequence of i.i.d. binary variables $\{E_n\}$ such that $Pr(E_n=1) = p < 0.5$ for every n, where p is the cross-correlation parameter, which may involve the plaintext statistics as well, [2]. Finally, a binary sequence $\{z_n\}$ is defined by

$$z_n = x_{f(n)} \oplus e_n \quad , \quad f(n) = n + \Sigma_{j=1}^{n} a_j \quad , \quad n=0,1,2,\ldots \quad . \qquad (1)$$

In this correspondence, the problem of the initial state $(X_0=[x_{-\ell}]_{\ell=1}^{L})$ reconstruction when $f(X)$, p, and a segment $\{z_n\}_{n=1}^{N}$ are known, is considered.

2. INITIAL STATE RECONSTRUCTION

A correlation attack [2] is based on the Hamming distance between two binary sequences of the same length. Obviously, the same statistical approach can not 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 $\{x_n\}$ into the output sequence $\{z_n\}$ according to the model displayed in Fig.1. Then, we could proceed along essentially the same lines as in [2], thus establishing a statistical procedure which we call a generalized correlation attack.

Due to 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 classes corresponding to all the initial states of the LFSR. Given an observed segment $\{z_n\}_{n=1}^{N}$, an optimal decision strategy (yielding the minimum probability of decission error) is to decide on the initial state with maximum posterior probability. When the LFSR is regularly clocked, as in [2], it is optimal to decide on

the initial state $\hat{X}_0$ such that the Hamming distance between $\{z_n\}_{n=1}^N$ and $\{\hat{x}_n\}_{n=1}^N$ is minimum (a sufficient statistics). 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 minimum distance decision procedure which may be close to optimal.

Let $\{\hat{x}_n\}_{n=1}^M$ be an LFSR sequence corresponding to the initial state $\hat{X}_0$ (typically, $M \cong 3N/2$). Let d be the distance between $\{\hat{x}_n\}_{n=1}^M$ and $\{z_n\}_{n=1}^N$. Two cases-hypotheses are possible:

H_0 : the observed sequence $\{z_n\}_{n=1}^N$ is produced by $\hat{X}_0$;

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

Consequently, d is a realization of a random variable D with two possible probability distributions (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 will be discussed in the next Section. Suppose that they are known. Note that they depend on N, assuming that $M = M(N)$. 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 [2], P_m is chosen close to zero (f.e., 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 $\hat{X}_0$:

Step 1: generate $\{\hat{x}_n\}_{n=1}^M$.

Step 2: calculate the distance d between $\{\hat{x}_n\}_{n=1}^M$ and $\{z_n\}_{n=1}^N$.

Step 3: 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. The computational complexity is proportional to the number of possible initial states (for example, 2^L).

3. A DISTANCE MEASURE AND RELEVANT PROBABILITY DISTRIBUTIONS

A distance measure should be defined so that it enables statistical

discrimination 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, and second, when $\{z_n\}_{n=1}^{N}$ is obtained from $\{\hat{x}_n\}_{n=1}^{M}$, according to the model in Fig.1, that is, by the deletion of some bits subject to the decimation constraints and by the 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 [3], for example) and text processing (see [5], for example). A review of the sequence matching techniques and applications is presented in [4].

According to [4], one of the widely used distances is the Levenshtein distance [3]. 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 (CLD) concept [7] is relevant, because the constraints are inherent to the decimation function (see relation (1)). In [5], [6], 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 define CLD^*, the distance measure between $\{\hat{x}_n\}_{n=1}^{M}$ and $\{z_n\}_{n=1}^{N}$ as the minimum number of deletions and complementations required to obtain $\{z_n\}_{n=1}^{N}$ from $\{\hat{x}_n\}_{n=1}^{M}$ subject to the assumed constraint on the number of consecutive deletions. Whether this distance is a sufficient statistics remains an open question, but it is reasonable to believe that this is approximately the case.

With the CLD^* so defined, a problem is to determine the probability distributions $\{Pr(D|H_0)\}$ and $\{Pr(D|H_1)\}$. According to the literature the problem appears very difficult. One approach is a nonparametric estimation .

Another problem is to define a procedure for efficient

computation of the defined distance measure. Following the main ideas from [7], a novel dynamic programming algorithm can be derived that computes the desired distance measure CLD^* in the following way.

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: $d(k,0) = k$, $k=0,1,\ldots,M-N$,
$d(0,\ell) = d(0,\ell-1) + (\hat{x}_\ell \oplus z_\ell)$, $\ell=1,2,\ldots,N$.

3. Recursive calculation for $M > N$:
$d(k,\ell) = \min\{ d(k-1,\ell-1)+(\hat{x}_{k+\ell-1}\oplus z_\ell)+1 , d(k,\ell-1)+(\hat{x}_{k+\ell}\oplus z_\ell) \}$,
$\ell=1,2,\ldots,N$, $k=\max\{1, M-2N+\ell\},\ldots,M-N$.

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

The computational complexity of the procedure is quadratic $O(N(M-N))$.

Note that an arbitrary number of initial deletions is allowed, since the length M of $\{\hat{x}_n\}_{n=1}^{M}$ that actually produced $\{z_n\}_{n=1}^{N}$ is not known (therefore, one can assume that $M = 2N+1$).

4. REFERENCES

[1] D.Gollman, W.G.Chambers, "Clock-controlled shift registers: A review", IEEE Journal on Selected Areas in Communications, vol. SAC-7, May 1989., pp.525-533.

[2] T.Siegenthaler, "Decrypting a class of stream ciphers using ciphertext only", IEEE Trans. Comput. vol. C-34, Jan. 1985. pp.81-85.

[3] A.Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", Sov. Phy. Dokl., vol.10, pp.707-710, 1966.

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

[5] B.J.Oommen, "Recognition of noisy subsequences using constrained edit distance", IEEE Trans. Pattern Analysis Mach. Intell., vol. PAMI-9, Sep. 1987., pp.676-685.

[6] B.J.Oommen, "Correction to 'Recognition of noisy subsequences using constrained edit distance'", IEEE Trans. Pattern Analysis Mach. Intell., vol. PAMI-10, Nov. 1988., pp.983-984.

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