A GENERALIZED CORRELATION ATTACK WITH A PROBABILISTIC CONSTRAINED EDIT DISTANCE

Jovan Dj. Golić and Slobodan V. Petrović

Institute of Applied Mathematics and Electronics, Belgrade
School of Electrical Engineering, University of Belgrade
Bulevar Revolucije 73, 11001 Beograd, Yugoslavia

Abstract: For a noisy clock-controlled shift register statistically optimal probabilistic constrained edit distance a recursive algorithm for its efficient computation are derived. corresponding generalized correlation attack is proposed.

1. Introduction

Consider the initial state reconstruction of a binary r clock-controlled shift register depicted in Fig. 1, see [1].

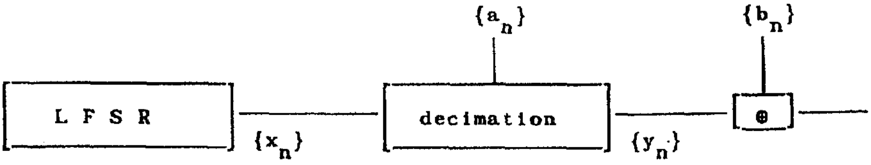

Fig. 1. A noisy clock-controlled shift register.

Linear feedback shift register (LFSR) produces a binary seq $\{x_n\}$ as usual. In a statistical model, the decimation sequence is regarded as a realization of a sequence of indepe identically distributed (i.i.d.) integer variables $\{A_n\}$ with a probability distribution $\Pr(A_n=k)=P_k$, $0\le k\le E$. A binary noise seq

This research was supported by the Science Fund of Se grant #0403, through Institute of Mathematics, Serbian Acade Arts and Sciences.

R.A. Rueppel (Ed.): Advances in Cryptology - EUROCRYPT '92, LNCS 658, pp. 472-476, 1993.

$\{b_n\}$ is a realization of a sequence of i.i.d. binary variables $\{B_n\}$ with $Pr(B_n=1)=p<0.5$. The output sequence $\{z_n\}$ is defined as the modulo 2 sum

$$z_n = x_{t_n} \oplus b_n, \qquad t_n = n + \sum_{j=1}^{n} a_j, \quad n \geq 1. \tag{1}$$

The objective is to reconstruct the LFSR initial state given the output segment $\{z_n\}_{n=1}^{N}$ along with the feedback polynomial and the noise probability. A solution to this problem is proposed in [1]. Essentially, it is a statistical procedure based on the minimum constrained Levenshtein distance (CLD) decision rule. However, as noted in [1], the problem remains to show how close the minimum CLD decision rule is to the maximum posterior probability one, which is optimal given the statistical model. For a slightly modified statistical model in which the LFSR sequence is assumed to be a realization of a sequence of balanced i.i.d. binary variables, we here derive the maximum posterior probability decision rule. Namely, we introduce the statistically optimal probabilistic constrained edit distance (PCED) and develop an efficient recursive algorithm for its computation.

2. Probabilistic Constrained Edit Distance

Let $U=\{u_i\}_{i=1}^{M}$ and $V=\{v_i\}_{i=1}^{N}$ denote two finite length sequences over a finite alphabet A. Let an edit transformation of U into V be defined as a series of edit operations of deletions and substitutions. It can be uniquely represented by a two-dimensional edit sequence $(U,V')=\{(u_i,v_i')\}_{i=1}^{M}$, where $V'=\{v_i'\}_{i=1}^{M}$ is a sequence over $\tilde{A} = A \cup \phi$, $\phi \notin A$, such that by deleting all the ϕ symbols from V' one obtains V. Namely, if $v_i'=\phi$, then u_i is deleted and if $v_i' \neq \phi$, then v_i' is substituted for u_i, $1 \leq i \leq M$. Let G_{UV} denote the set of all possible edit sequences that transform U into V subject to the constraint that there are no more than E consecutive ϕ symbols in V' and that the last symbol of V' is different from ϕ. It follows that $N \leq M \leq (E+1)N$.

In accordance with the noisy clock-controlled shift register statistical model in which the LFSR sequence is assumed to be a realization of a sequence of balanced i.i.d. binary variables, we now associate a probability distribution with the set of permitted edit sequences, that is, the union of G_{UV} over all U and V. To this end, for an arbitrary permitted edit sequence (U,V') define the decimation sequence $D(U,V')=\{d_i\}_{i=1}^{N}$ so that d_i is the length of the series of deletions between the (i-1)-th and the i-th substitution, $0 \leq d_i \leq E$, $1 \leq i \leq N$. An edit sequence (U,V') can then be regarded as a sequence of N blocks each composed of a substitution preceded by a series of deletions, allowing a series of zero deletions. Assume that the blocks are produced independently and that within each block the substitution and the series of deletions are also independent. Further, let $P_k A^{-k}$, A being the cardinality of **A**, denote the probability of a series of deletions of length k, $0 \leq k \leq E$, and let p(u,v) denote the probability of substituting v for u, $u,v \in \mathbf{A}$. Thus, P_k is the probability that a series of deletions has length k, $0 \leq k \leq E$. The probability of an edit sequence (U,V') is then given by

$$\Pr(U,V') = \prod_{i=1}^{N} A^{-d_i} P_{d_i} p(u_{i+\Sigma_{j=1}^{i} d_j}, v_i) = A^{N-M} \prod_{i=1}^{N} P_{d_i} p(u_{i+\Sigma_{j=1}^{i} d_j}, v_i). \quad (2)$$

The probability that a sequence U can be transformed into a sequence V is then given by

$$\Pr(U,V) = \sum_{(U,V')\in G_{UV}} \Pr(U,V'). \quad (3)$$

The problem is how to compute Pr(U,V) efficiently. To this end we define the partial probability $\Pr(U_{e+s},V_s)$ that a prefix $U_{e+s}=\{u_i\}_{i=1}^{e+s}$ of U can be transformed into a prefix $V_s=\{v_i\}_{i=1}^{s}$ of V, under the same constraints. Using an abbreviated notation $G_{es}=G_{U_{e+s}V_s}$ and $P(e,s)=\Pr(U_{e+s},V_s)$ we thus have

$$P(e,s) = \sum_{(U_{e+s},V_s')\in G_{es}} \prod_{i=1}^{s} A^{-d_i} P_{d_i} p(u_{i+\Sigma_{j=1}^{i} d_j}, v_i) \quad (4)$$

and

$$\Pr(U,V) = P(M-N,N). \tag{5}$$

The set of all the permitted values for (e,s) is clearly given by $1 \leq s \leq N$ and $0 \leq e \leq \min(M-N, sE)$.

Interestingly enough, using a similar technique as in [1] and bearing in mind the recursion [2] for unconstrained edit probability, it is not difficult to prove the following result which yields a recursion for P(e,s).

Theorem 1: The partial probability P(e,s) satisfies the recursion

$$P(e,s) = \sum_{e_1 \in \Psi_{es}} P(e-e_1, s-1) A^{-e_1} P_{e_1} p(u_{e+s}, v_s) \tag{6}$$

for $1 \leq s \leq N$ and $0 \leq e \leq \min(M-N, sE)$, where Ψ_{es} is the set of all e_1 such that

$$\max(0,\ e-\min(M-N,(s-1)E)) \leq e_1 \leq \min(e,\ E), \tag{7}$$

with the initial value P(0,0)=1 for s=0 and e=0.

Now, Theorem 1 together with (5) enables the efficient computation of the probability Pr(U,V), given U and V, which has the same space and time complexities as the recursive procedure [1] for the efficient computation of the CLD. Since the value of Pr(U,V) is very close to zero, for practical reasons we define the probabilistic constrained edit distance (PCED) as

$$D(U,V) = -\log \Pr(U,V). \tag{8}$$

Note that in the optimal statistical decision procedure about U given V, D(U,V) has to be minimized rather than maximized.

We now discuss the relation between the CLD and the PCED. Instead of the overall probability that a sequence U can be transformed into a sequence V, consider the probability of the most likely edit transformation of U into V. Then the negative logarithm of this probability reduces to a CLD with elementary edit distances defined as the negative logarithms of the corresponding probabilities. Unlike the definition in [1], this CLD is modified in a sense that the elementary edit distance is associated with a series of deletions rather than a single deletion.

3. Generalized Correlation Attack

Given the noisy clock-controlled shift register statistical model, Fig. 1, with the additional assumption that the LFSR sequence is a realization of a sequence of balanced i.i.d. binary variables, the probability that a LFSR sequence $U=\{x_n\}_{n=1}^{M}$ can be transformed into an output sequence $V=\{z_n\}_{n=1}^{N}$ is given by (3) and (2) with A=2, $P_k=Pr(A_n=k)$, $0\leq k\leq E$, and $p(u,v)=p$ if $u\neq v$ and $p(u,v)=1-p$ if $u=v$. Accordingly, applying Theorem 1 one can compute the probabilistic constrained edit distance, PCED, which is statistically optimal for a given model. Then, using the PCED instead of the constrained Levenshtein distance, CLD, the generalized correlation attack goes along the same lines as in [1]. Unlike the CLD procedure, the PCED procedure is statistically optimal and enables to take into account the probability distribution of the decimated sequence.

Comparative experimental analysis is underway. Preliminary results show a slight but clear advantage of the PCED procedure over the CLD one.

REFERENCES

[1] Jovan Dj. Golić, Miodrag J. Mihaljević, "A generalized correlation attack on a class of stream ciphers based on the Levenshtein distance," *J. Cryptology* vol. 3, pp. 201-212, 1991.

[2] P.A.V. Hall, G.R. Dowling, "Approximate string matching," *Computing surveys*, vol. 12, pp. 381-402, Dec. 1980.