Embedding and Probabilistic Correlation Attacks on Clock-Controlled Shift Registers

Jovan Dj. Golić[1]

Information Security Research Centre, Queensland University of Technology,
GPO Box 2434, Brisbane, Q 4001, Australia
School of Electrical Engineering, University of Belgrade

Luke O'Connor[2]

Distributed Systems Technology Centre
Information Security Research Centre, Queensland University of Technology

Abstract

Embedding and probabilistic correlation attacks on clock-controlled shift registers that are clocked at least once per output symbol are defined in general and are analyzed in the unconstrained case, with an arbitrary number of deletions at a time, and in the constrained case, with at most d deletions at a time. It is proved that the unconstrained embedding attack is successful if and only if the deletion rate is smaller than one half and if the length of the observed keystream sequence is greater than a value linear in the shift register length r. It is shown how to compute recursively the joint probability which is a basis for the unconstrained probabilistic attack with independent deletions. The efficiency of the attack is characterized in terms of the capacity of the corresponding communication channel with independent deletions and it is concluded that the probabilistic attack is successful for any deletion rate smaller than one if the given keystream sequence is sufficiently long, also linearly in r. It is proved that the constrained embedding attack is successful for any d and the minimum necessary length of the known output sequence is shown to be linear in r, and at least exponential and at most superexponential in d. This demonstrates that making d large can not ensure the theoretical security against the attack, but can considerably improve the practical security.

1 Introduction

Irregularly decimated linear recurring sequences produced by clock-controlled shift registers generally possess nice cryptographic properties such as long periods, high linear complexities (see [9, 3, 8], for example), and immunity to fast

[1]This research was supported in part by the Science Fund of Serbia, grant #0403, through the Institute of Mathematics, Serbian Academy of Arts and Sciences.

[2]The work reported in this paper has been funded in part by the Cooperative Research Centres program through the Department of the Prime Minister and Cabinet of Australia.

correlation attacks. Let $X = \{x_t\}_{t=1}^{\infty}$ denote the output sequence of a regularly clocked binary shift register with not necessarily linear feedback. Let a decimation sequence be defined in terms of its increments, that is, as a non-negative integer sequence $D = \{d_t\}_{t=1}^{\infty}$. In practice, D is produced by a finite-state machine, called a clock-control generator, and is therefore ultimately periodic. It is assumed that the secret key controls the initial states of both the shift register and the clock-control generator. The output sequence $Y = \{y_t\}_{t=1}^{\infty}$ of the clock-controlled shift register (CCSR) is defined as a decimated sequence (see [8], for example)

$$y_t = x\left(\sum_{i=1}^{t} d_i\right), \quad t \geq 1. \tag{1}$$

Note that the decimation operation actually means that in order to obtain the next output symbol y_t, after producing y_{t-1}, one has to delete $d_t - 1$ consecutive symbols from X if $d_t \geq 1$ or has to repeat y_{t-1} if $d_t = 0$. The objective of correlation attacks is to reconstruct the initial state of the clock-controlled shift register based on a given segment of the output sequence, without knowing the decimation sequence. Let $\mathcal{D}$ denote the range of D, that is, the set of all values achievable by D. If $\mathcal{D} = \{k, m\}$, then a CCSR is said to be $\{k, m\}$-clocked [10] and in particular if $\mathcal{D} = \{0, 1\}$, then it is called stop-and-go [1]. Cascades of stop-and-go and $\{k, m\}$-clocked shift registers have been cryptanalyzed in [1] and [10], respectively, based on a specific lock-in effect. Recently, a correlation-like attack on cascades of stop-and-go linear feedback shift registers has been proposed in [14, 13]. In [17], a constrained embedding attack on the initial state of a $\{1, 2\}$-clocked shift register has been developed. If $\mathcal{D} = [1, k] = \{1, 2, \ldots, k\}$, then a CCSR is said to be $[1, k]$-clocked. A divide and conquer correlation attack on the initial state of a noisy $[1, k]$-clocked shift register has been introduced in [6] using a generalization of the Levenshtein distance, whereas a probabilistic correlation attack on the same scheme has been devised in [7]. A correlation attack based on a variation of the unconstrained Levenshtein distance has been proposed in [15].

In general, for any $\mathcal{D}$, say that a CCSR is $\mathcal{D}$-clocked. A notion of a constrained embedding attack [17] can be extended as follows. Say that a given binary string $Y^n = \{y_i\}_{i=1}^{n}$ of length n can be $\mathcal{D}$-embedded into a given binary string $X^m = \{x_i\}_{i=1}^{m}$ of length m, $m \geq n$, if there exists a decimation string $D^n = \{d_i\}_{i=1}^{n}$ of length n such that $d_i \in \mathcal{D}$ and $y_i = x\left(\sum_{j=1}^{i} d_j\right)$, $1 \leq i \leq n$. For simplicity, we will use the notation X, Y, and D instead of X^m, Y^n, and D^n, respectively. In a $\mathcal{D}$-embedding correlation attack, given a segment of the keystream sequence Y of length n, the objective is to find all the initial states of the shift register that under regular clocking result in a sequence of length m, $m > n$, into which X can be $\mathcal{D}$-embedded, where n and m should be large enough. The attack is successful if there are only a few, preferably only one, candidate initial states. Note that a $\mathcal{D}$-embedding attack can be applied to any $\mathcal{D}'$-clocked shift register such that $\mathcal{D}' \subseteq \mathcal{D}$, because in this case $\mathcal{D}'$-embedding

implies $\mathcal{D}$-embedding. Define a $\mathcal{D}$-embedding probability $P_{\mathcal{D},Y}(n,m)$ as the probability that a given binary string Y of length n can be $\mathcal{D}$-embedded into a purely random (uniformly distributed) binary string X of length m. The length $m = m(n)$ should be chosen appropriately so that the probability of missing event is small enough, where the decimation sequence, which is not known to the cryptanalyst, is assumed to be random. It is clear that the initial state reconstruction is possible only if $P_{\mathcal{D},Y}(n,m(n))$ tends to zero when n increases, for all Y or for a purely random Y. If this happens exponentially, then the minimum necessary length of the keystream sequence for the successful reconstruction is linear in the shift register length. The only known result in the literature about $\mathcal{D}$-embedding probabilities is given in [17] for $\mathcal{D} = \{1,2\}$. More precisely, an upper bound on $P_{\{1,2\},Y}(n,2n)$ for all Y which is exponentially small in n is derived in [17].

Our main objective in this paper is to consider the unconstrained case where $\mathcal{D} = Z^+$, the set of positive integers. The corresponding unconstrained embedding attack is applicable to an arbitrary clock-controlled shift register that is clocked at least once per output symbol. In Section 2, we introduce a probabilistic model for a $\mathcal{D}$-clocked shift register assuming that the decimation sequence and the regularly clocked shift register sequence are independent random sequences. We also define the deletion rate p_d as the relative expected number of deleted symbols in a $\mathcal{D}$-clocked shift register. In Section 3, we derive the unconstrained embedding probability $P_{Z^+,Y}(n,m)$ for arbitrary n, m, and Y and determine its asymptotic behaviour. We show that when n increases the probability exponentially tends to zero if $p_d < 0.5$, tends to 0.5 if $p_d = 0.5$, and tends to 1 if $0.5 < p_d \leq 1$. Therefore we prove that the unconstrained embedding attack is successful only if $p_d < 0.5$. Compare this with the $\{1,2\}$-embedding attack [17] where $0 < p_d < 0.5$ and typically $p_d = 1/3$.

In Section 4, we describe a statistically optimal correlation attack which instead of the embedding possibility uses the joint probability in the assumed probabilistic model for a $\mathcal{D}$-clocked shift register. For the unconstrained case ($\mathcal{D} = Z^+$) with independent deletions, we show that the corresponding joint probability can be computed recursively. We then point out a necessary and sufficient condition for a successful attack in terms of the capacity of the corresponding communication channel with independent deletion synchronization errors. In light of the results presented in [4], obtained by systematic computer simulations, we conclude that the unconstrained probabilistic attack with independent deletions is successful for any $0 \leq p_d < 1$.

Our second objective is to analyze the constrained $\mathcal{D}$-embedding attack where $\mathcal{D} = [1, d+1]$ and thus extend the result [17] from $d = 1$ to an arbitrary positive integer d, which is the noiseless instance of the general problem defined in [6]. Note that the technique from [17] is based on direct counting and hence can not be used in the general case. In Section 5, an exponentially small upper bound on the constrained embedding probability for a random string is obtained based on the corresponding upper bound for a constant string that is derived by using regular languages and generating functions. Consequently, it turns

out that by making d large one can not achieve the theoretical security against the constrained embedding attack, but can significantly increase the practical security.

2 Probabilistic Model

Assume that $\tilde{X} = \{\tilde{x}_t\}_{t=1}^{\infty}$ is a purely random binary sequence, that is, a sequence of balanced i.i.d. binary random variables. Also assume that a random decimation sequence $\tilde{D} = \{\tilde{d}_t\}_{t=1}^{\infty}$ is a sequence of i.i.d. non-negative integer random variables that is independent of $\tilde{X}$. Let $\mathcal{P} = \{P(d)\}_{d \in \mathcal{D}}$ denote the probability distribution of $\tilde{d}_t$, for any $t \geq 1$, where $\mathcal{D}$ is the set of values with positive probability. The random sequences $\tilde{X}$ and $\tilde{D}$ are combined by a decimation equation (1) in the output random sequence $\tilde{Y} = \{\tilde{y}_t\}_{t=1}^{\infty}$. Clearly, $\tilde{Y}$ is a purely random binary sequence itself provided that $0 \notin \mathcal{D}$. One can then define the joint probability distribution $P(X, Y)$ for all pairs of binary strings $X = \{x_t\}_{t=1}^{m}$ and $Y = \{y_t\}_{t=1}^{n}$, for any $m \geq n$, which is a basis for the statistically optimal attack employing the maximum posterior probability decision rule. Note that efficient computation of $P(X, Y)$ generally presents a problem, which is discussed in Section 4.

The deletion rate p_d for an arbitrary decimation probability distribution $\mathcal{P}$ is defined as

$$p_d = 1 - \frac{1}{\bar{d}}, \quad \bar{d} = \sum_{d \in \mathcal{D}} d\, P(d), \tag{2}$$

which is essentially the relative expected number of deleted symbols from the input sequence $\tilde{X}$ needed to obtain the output sequence $\tilde{Y}$ for the assumed probabilistic model. The deletion rate is needed in order to control the missing event probability in a $\mathcal{D}$-embedding attack and also to characterize the efficiency of the unconstrained embedding attack, see Section 3.

Another, more general way of looking at the decimation probabilistic model is to view it as a specific communication channel with deletion errors. To this end one should assume that the input random sequence is a sequence of i.i.d., not necessarily balanced, binary random variables. Interestingly, one can also define the capacity of such a channel as the maximum mutual information between the input and output, taken over all the input probability distributions. Furthermore, an analogue of the Shannon's coding theorem for noisy channels also holds for the channels with synchronization errors, see [4]. More precisely, the transmission of information through this channel with arbitrarily small probability of error is possible if and only if the rate of the code is smaller than the capacity. Unfortunately, the problem of deriving the capacity of such channels seems to be intractable. There are very few theoretical results in the literature, giving only some lower and upper bounds on the capacity (see [4], for example). This problem will be addressed in more detail in Section 4, including its relation to the embedding and probabilistic correlation attacks on the clock-controlled shift registers.

3 Unconstrained Embedding Attack

The basic lines of the $\mathcal{D}$-embedding attack are already explained in Section 1. Given a set of non-negative integers $\mathcal{D}$, say that a binary string $Y = \{y_i\}_{i=1}^n$ of length n can be $\mathcal{D}$-embedded into a binary string $X = \{x_i\}_{i=1}^m$ of length m if there exists a non-negative integer string $D = \{d_i\}_{i=1}^n$ of length n such that $d_i \in \mathcal{D}$ and $y_i = x\left(\sum_{j=1}^i d_j\right)$, $1 \le i \le n$. To check whether Y can be $\mathcal{D}$-embedded into X, one can generally use the direct matching algorithm which is for $\mathcal{D} = \{1, 2\}$ given in [17]. Namely, first find all the matching positions in X for the first symbol of Y, using the constraints given by $\mathcal{D}$. Then proceed iteratively: after finding all the matching positions in X for y_{t-1}, find all the matching positions in X for the next symbol y_t according to $\mathcal{D}$. The computational complexity is $O(nm)$. Another approach is to use the constrained Levenshtein distance algorithm [6] for $\mathcal{D} = [1, k]$ or the unconstrained one [15] for $\mathcal{D} = Z^+$. The embedding is possible if and only if the distance is equal to the minimum value which is, basically, the difference of the string lengths. The computational complexity is $O(n(m-n))$.

In a $\mathcal{D}$-embedding attack one checks whether a given segment of the keystream sequence Y of length n can be $\mathcal{D}$-embedded into a regularly clocked shift register sequence X of length $m(n)$, for all possible shift register initial states. The length $m(n)$ of X should be chosen so that the probability $P\left(\sum_{i=1}^n d_i > m(n)\right)$ is equal to an upper bound P_m on the missing event probability which should be close to zero. It is clear that one may choose $m(n) = n/(1-p_d) + c\sqrt{n}$, where p_d is the deletion rate (2) and c is a constant depending on P_m. On the other hand, the success of the attack can be measured by the false alarm probability P_f which can be approximately expressed in terms of the $\mathcal{D}$-embedding probability $P_{\mathcal{D},Y}(n, m(n))$ as $P_f = 1 - (1 - P_{\mathcal{D},Y}(n, m(n)))^{2^r - 1}$. Recall that $P_{\mathcal{D},Y}(n, m)$ is defined as the probability that a binary string Y of length n can be $\mathcal{D}$-embedded into a purely random binary string X of length m. The criterion $P_f \approx 0$ is well approximated by

$$2^r P_{\mathcal{D},Y}(n, m(n)) \le 1, \tag{3}$$

which yields the minimum necessary length n of the observed keystream sequence, provided that $P_{\mathcal{D},Y}(n, m(n))$ decreases sufficiently fast as n increases. The minimum length is linear in r if the embedding probability decreases exponentially with n.

The problem of deriving the $\mathcal{D}$-embedding probability for a general decimation set $\mathcal{D}$ appears to be very difficult. In this section, we will consider the unconstrained embedding case where $\mathcal{D} = Z^+$, the set of positive integers, emphasizing that the unconstrained embedding, that is, Z^+-embedding attack applies to an arbitrary $\mathcal{D}$-clocked shift register such that $\mathcal{D} \subseteq Z^+$. Note that a binary string $Y = \{y_i\}_{i=1}^n$ can be Z^+-embedded into a binary string $X = \{x_i\}_{i=1}^m$ if there exists a positive integer decimation string $D = \{d_i\}_{i=1}^n$ such that $y_i = x\left(\sum_{j=1}^i d_j\right)$, $1 \le i \le n$. We will derive an analytical expression for the unconstrained embedding probability and examine its asymptotic behaviour. Then we will show

that the efficiency of the unconstrained embedding attack can be characterized in terms of the deletion rate of a $\mathcal{D}$-clocked shift register, for any $\mathcal{D} \subseteq Z^+$. For simplicity, denote the unconstrained embedding probability $P_{Z^+,Y}(n,m)$ by $P_Y(n,m)$.

Theorem 3.1 For an arbitrary binary string Y of length n, the unconstrained embedding probability is given by

$$P_Y(n,m) = P_+(n,m) \stackrel{\text{def}}{=} \sum_{k=0}^{m-n} \binom{n-1+k}{k} 2^{-n-k} \tag{4}$$

$$= 1 - 2^{-m} \sum_{k=0}^{n-1} \binom{m}{k}. \tag{5}$$

□

Proof. Assume that Y of length n can be Z^+-embedded into X of length m. We prove that there exists a decimation string $D^* = \{d_i^*\}_{i=1}^n$ that is minimal in a sense that each of its elements is minimal over the set of all permissible decimation strings, given Y and X. We prove it by iterative construction. Let d_0^* be equal to the minimal positive integer j such that $y_1 = x_j$. Then, iteratively for $2 \leq i \leq n$, let d_i^* be equal to the minimal positive integer j such that $y_i = x\left(\sum_{k=1}^{i-1} d_k^* + j\right)$. It is straightforward to show that the so-obtained D^* is minimal. Note that D^* is unique by definition.

Clearly, $P_Y(n,m) = A_Y(n,m)/2^m$ where $A_Y(n,m)$ is the number of binary strings X of length m into which Y of length n can be Z^+-embedded. The uniqueness property of D^* ensures that different D^* give rise to different X. Also, it is easy to see that each D^* such that $\sum_{i=1}^n d_i = m - k$, where $0 \leq k \leq m-n$, gives rise to exactly 2^k different X. Since for any $0 \leq k \leq m-n$, there are exactly $\binom{m-k-1}{n-1}$ such D^*, independent of Y, then (4) follows easily. Equation (5) is directly obtained by considering a constant string Y. □

We proceed by analyzing the asymptotic properties of $P_+(n,m)$. Suppose that $\{m(n)\}_{n=1}^\infty$ is a positive integer sequence such that

$$\lim_{n\to\infty} \frac{n}{m(n)} = 1 - \lambda, \quad 0 \leq \lambda \leq 1. \tag{6}$$

Then by employing the well-known inequality

$$\frac{1}{\sqrt{8k(n-k)/n}} 2^{nH(k/n)} \leq \sum_{i=0}^{k} \binom{n}{i} \leq 2^{nH(k/n)} \tag{7}$$

which holds for $k \leq n/2$, where $H(x) = -x\log x - (1-x)\log(1-x)$ is the binary entropy function (logarithm is to the base two throughout), it is easy to see that (5) results in

Corollary 3.1 If $\{m(n)\}_{n=1}^{\infty}$ satisfies (6), then

$$\lim_{n\to\infty} -\frac{\log P_+(n, m(n))}{n} = \frac{1-H(\lambda)}{1-\lambda}, \quad 0 \le \lambda < 0.5 \tag{8}$$

$$\lim_{n\to\infty} P_+(n, m(n)) = \begin{cases} 0.5 & \lambda = 0.5 \\ 1 & 0.5 < \lambda \le 1 \end{cases}. \tag{9}$$

□

Let us now turn to the unconstrained embedding attack on a $\mathcal{D}$-clocked shift register, $\mathcal{D} \subseteq Z^+$. It follows that in order to keep the upper bound P_m on the missing event probability away from one it is necessary and sufficient that $\lambda \ge p_d$ where p_d is the deletion rate (2) for a $\mathcal{D}$-clocked shift register. Of course, as noted earlier, one may choose $m(n) = n/(1-p_d)+c\sqrt{n}$ in which case $\lambda = p_d$. If $\lambda < p_d$, that is, if P_m tends to one, then the true initial state is indistinguishable from the others. Combining this with Corollary 3.1 and the criterion (3) for a successful reconstruction we obtain

Theorem 3.2 For any $\mathcal{D} \subseteq Z^+$, the unconstrained embedding attack on a $\mathcal{D}$-clocked shift register with deletion rate p_d is successful if $p_d < 0.5$ and if the length n of the observed keystream sequence satisfies

$$n \ge r\,\frac{1-p_d}{1-H(p_d)}. \tag{10}$$

If $p_d \ge 0.5$, then the attack is not successful. □

Theorem 3.2 shows that a $\mathcal{D}$-clocked shift register, for any $\mathcal{D} \subseteq Z^+$, is theoretically secure against the unconstrained embedding attack if the deletion rate p_d is greater than or equal to one half. Of course, this does not mean that it is also secure against a more suited $\mathcal{D}$-embedding attack if $\mathcal{D} \ne Z^+$ and $p_d \ge 0.5$. However, if $\mathcal{D} = Z^+$ it follows that a Z^+-clocked shift register is theoretically secure against the embedding attacks whatsoever if $p_d \ge 0.5$. One may be tempted to conclude that this also holds for arbitrary correlation attacks as well. In the next section this case is examined in more detail.

4 Unconstrained Probabilistic Attack

Embedding attacks make no use of the probability distribution of the decimation sequence. Therefore, they are not optimal in general. For the assumed probabilistic model, the probabilistic attack based on the joint probability of the original and decimated sequences is statistically optimal. The problem of efficient computation of this probability is solved in [7] for the constrained case $\mathcal{D} = [1, k]$ by using a result [11] dealing with string matching in the unconstrained case with independent deletions, insertions, and substitutions.

Consider now a special case, which we call the unconstrained case with independent deletions. Namely, consider a Z^+-clocked shift register with the decimation probability distribution $P(d) = p^{d-1}(1-p)$, $d \in Z^+$. It follows that

$p_d = p$. It is easy to show that the probabilistic model is then equivalent to the model in which the output random binary sequence $\tilde{Y}$ is obtained from the input random binary sequence $\tilde{X}$ by a random binary sequence of independent deletions of symbols from $\tilde{X}$ with the probability p. This model turns out to be a special case of the model considered in [11], since there are no insertions and effective substitutions (when one symbol is replaced by a different one). Accordingly, the desired joint probability distribution of input and output sequences in this model can be computed either by a recursive algorithm [11] or by a more efficient recursive algorithm which is now described. The objective is to determine the joint probability $P(X, Y)$ for the described probabilistic model, for arbitrary input and output binary strings $X = \{x_t\}_{t=1}^{m}$ and $Y = \{y_t\}_{t=1}^{n}$, $m \geq n$, respectively. Let $P(e, s)$ denote the partial joint probability for the prefix $X^{e+s} = \{x_t\}_{t=1}^{e+s}$ of X of length $e+s$ and the prefix $Y^s = \{y_t\}_{t=1}^{s}$ of Y of length s, for any $1 \leq s \leq n$ and $0 \leq e \leq m-n$. Let $\delta(x, y)$ denote the substitution probability defined to be equal to 0.5 if x and y are equal and to zero otherwise. Then using a similar technique as in [6, 7] one can prove

Theorem 4.1 The partial probability satisfies the recursion

$$P(e, s) = P(e-1, s)p + P(e, s-1)(1-p)\delta(x_{e+s}, y_s) \tag{11}$$

for $1 \leq s \leq n$ and $0 \leq e \leq m-n$, with the initial values $P(e, 0) = p^e, 0 \leq e \leq m-n$, and $P(-1, s) = 0, 1 \leq s \leq n$. □

Finally, $P(X, Y) = P(m-n, n)$. The computational complexity is $O(n(m-n))$. Consequently, in the unconstrained probabilistic attack on a Z^+-clocked shift register with independent deletions, by the above algorithm one computes the joint probability for a given segment Y of the keystream sequence of length n and a regularly clocked shift register sequence X of length $m(n)$, for all possible initial states, and then decides on the initial state with maximum joint probability. The length $m(n)$ should be chosen so that $\lim_{n\to\infty} n/m(n) = 1-p$, for example, $m(n) = n/(1-p)$. The attack is also applicable to an arbitrary $\mathcal{D}$-clocked shift register, in which case one sets $p = p_d$. Of course, in this case a constrained probabilistic attack, such as the one described in [7], is better suited.

We are interested to determine the conditions under which the unconstrained probabilistic attack is successful. In principle, this could be done by analyzing the asymptotic properties of the joint probability obtained by the recursion (11). One should examine the two cases: first, when X and Y are generated according to the assumed model and, second, when X and Y are independent purely random strings. This appears to be very difficult. Instead, we propose another approach related to the capacity of a communication channel with independent deletion errors to which the probabilistic model under consideration becomes equivalent if one allows an arbitrary input distribution. Assume that the regularly clocked shift register sequences behave like random codewords, for different initial states. The same assumption underlies the criterion (3) for a success of the embedding attacks. Then in light of the analogue of the Shannon's coding theorem for communication channels with synchronization errors it follows that

the statistically optimal decoding procedure or, equivalently, the unconstrained probabilistic attack with independent deletions is successful if and only if

$$\frac{r}{m} < C \iff n > r\frac{1-p}{C} \tag{12}$$

where C is the capacity of the channel, r is the length of the shift register, m is the length of the codewords, and $n = m(1-p)$ is the expected length of the received codewords. For non-optimal decision procedures, such as the unconstrained embedding attack, this condition is necessary but in general not sufficient. It is then clear that Theorem 3.2 essentially yields a lower bound on the capacity $\underline{C}(p) = 1 - H(p)$, $0 \leq p \leq 0.5$, and $\underline{C}(p) = 0$, $0.5 \leq p \leq 1$. This improves on a lower bound $1 - H(p) - p$ which has been analytically derived in [4].

Furthermore, the theoretically established upper bounds [4] on the capacity and the presented experimental results for the capacity itself, obtained by extensive computer simulations, clearly indicate that the capacity of the considered channel with independent deletions is greater than zero for any $0 \leq p < 1$ and is equal or very close to the upper bound

$$\bar{C}(p) = (1 - \frac{p}{2})\log(2-p) + \frac{p}{2}\log p. \tag{13}$$

This means that the unconstrained probabilistic attack with independent deletions is successful for any $0 \leq p < 1$ provided that the length n of the observed keystream sequence satisfies (12), where $C \approx \bar{C}$.

For illustration, consider the recently proposed 'shrinking' generator [2] that consists of two linear feedback shift registers one of which irregularly clocks the other, see also [15]. It can be very well approximated by a Z^+-clocked shift register with independent deletions that occur with the probability $p = 0.5$. Theorem 3.2 proves that the generator is theoretically secure against any embedding correlation attacks on the irregularly clocked shift register. However, it turns out that the unconstrained probabilistic attack is successful if the length of the observed keystream sequence is greater than approximately $0.5/\bar{C}(0.5) \approx 3$ lengths of the irregularly clocked shift register.

For an arbitrary $\mathcal{D}$-clocked shift register in general, our conjecture is that there might exist the conditions under which the embedding correlation attacks can not work, but the statistically optimal probabilistic correlation attack always works if the length of the observed keystream sequence is greater than a value linear in the length of the shift register.

5 Constrained Embedding Attack

In this section, we consider a constrained $\mathcal{D}$-embedding attack where $\mathcal{D} = [1, d+1]$ for an arbitrary positive integer d. As noted before, it applies to any $\mathcal{D}'$-clocked shift register such that $\mathcal{D}' \subseteq [1, d+1]$. The corresponding $\mathcal{D}$-embedding is for simplicity called d-embedding. Equivalently, a binary string $Y = \{y_i\}_{i=1}^n$ can

be d-embedded into a binary string $X = \{x_i\}_{i=1}^m$ if Y can be obtained from a prefix of X by deleting no more than d consecutive bits before each bit of Y. If also the prefix of X coincides with X and the first bit of X is not deleted, then Y is said to ***strictly*** d-embed into X. The missing event probability of the d-embedding attack is exactly equal to zero if the length of X is chosen to be maximum possible $m(n) = (n+1)d + d$. The corresponding d-embedding probability $P_{[1,d+1],Y}(n, m(n))$ is denoted by $P_{d,Y}(n)$. Ideally, we would like to determine this probability for each Y or for a purely random Y. However, this appears to be a very difficult combinatorial problem which is even not solved for $d = 1$ in [17]. Therefore, our objective here is to obtain a suitable exponentially small upper bound which holds for a purely random string. To this end, define $P_{d,Y}(n,k)$ as the probability that Y can be strictly d-embedded into a purely random string X of length $n+k$, $0 \le k \le nd$. The corresponding upper bound then follows directly

$$P_{d,Y}(n) \;\le\; P^*_{d,Y}(n) \stackrel{\text{def}}{=} \sum_{k=0}^{nd} P_{d,Y}(n,k). \tag{14}$$

Clearly, an upper bound for all Y is then $P^*_d(n) \stackrel{\text{def}}{=} \sum_{k=0}^{nd} P_d(n,k)$ where $P_d(n,k)$ denotes the maximum of $P_{d,Y}(n,k)$ over all Y of length n. Since it appears very difficult to derive an analytical expression for $P_d(n,k)$ even for $d = 1$, we take an approach based on the following property which enables us to obtain upper bounds for concatenations of strings.

Lemma 5.1 Let $Y = Y_1Y_2$ denote the concatenation of Y_1 and Y_2 of lengths n_1 and n_2, respectively, where Y has length $n = n_1 + n_2$. Then

$$P^*_{d,Y}(n) \;\le\; P^*_{d,Y_1}(n_1)\, P^*_{d,Y_2}(n_2). \tag{15}$$

□

Note that unlike the upper bound, a lower bound on $P_{d,Y}(n)$ is easily obtained by considering d-embeddings that match each bit of Y as soon as possible (least index) in X.

Lemma 5.2 For all Y, $P_{d,Y}(n) \ge \left(1 - \frac{1}{2^{d+1}}\right)^n$. □

The approach that we propose is based on Lemma 5.1 and consists of the two stages. Let $P_{d,c}(n,k)$ and $P^*_{d,c}(n)$ denote $P_{d,Y}(n,k)$ and $P^*_{d,Y}(n)$ for a constant string Y, respectively, where a constant binary string Y of length n is denoted as 0^n or 1^n. In the first stage we will analytically determine an exponential upper bound on $P_{d,c}(n,k)$ and the corresponding upper bound on $P^*_{d,c}(n)$. Then in the second stage, by using Lemma 5.1 and the fact that any string Y can be divided into constant substrings, we will establish an exponential upper bound on the d-embedding probability that holds with probability arbitrarily close to one for a sufficiently long purely random string Y.

An upper bound on $P_{d,c}(n,k)$ is based on the following observation. If the constant string $Y = 0^n$ can be strictly d-embedded into a string X of length

$m \geq n$, then X does not contain the substring 1^{d+1} (with the analogous property holding for 1^n and the substring 0^{d+1}). Accordingly, by enumerating all binary strings that possess this property we can obtain an upper bound on $P_{d,c}(n,k)$. We will do this by using regular expressions from the theory of formal languages [12], and generating functions from combinatorial theory [16]. The set of binary strings that begin with 0 and do not contain $1^l, l = d+1$, as a substring is a regular language [12] for fixed d, which we will denote by L_l^c. Equivalently, there is a deterministic finite automata (DFA) which recognizes (or accepts) the members of L_l^c. The DFA for L_l^c is not unique, but for a given DFA that accepts L_l^c, a regular expression for L_l^c can be determined. Of all such regular expressions consider

$$L_l^c = (\mathbf{0} + \mathbf{01} + \cdots + \mathbf{0}\underbrace{\mathbf{11}\cdots\mathbf{11}}_{l-1 \text{ times}})^* \tag{16}$$

meaning that each string $X \in L_l^c$ can be obtained by repeated concatenation of strings from the set $\{0\} \cup \{01\} \cup \cdots \cup \{011\cdots11\}$, since the $*$ operator means 'select zero or more times'. The empty string ϵ of length zero is also included. It is crucial to note that each $X \in L_l^c$ can be uniquely decomposed into the substrings $0, 01, 011, \cdots, 011\cdots11$ that define the regular expression (16) for L_l^c. The unique decomposition property allows L_l^c to be enumerated using the generating function $\frac{1}{1-z} = \sum_{i \geq 0} z^i$, and several other basic results for generating functions [16].

Lemma 5.3 Let $C_l(n)$ denote the number of strings from L_l^c of length $n \geq 0$, $l = d+1$. Then $C_l(n)$ is equal to the nth coefficient $[z^n]$ of the generating function

$$G_l(z) = \frac{1}{1 - (z + z^2 + \cdots + z^{l-1} + z^l)} = \frac{1}{1 - \sum_{i=1}^{l} z^i}. \tag{17}$$

□

The embedding probability $P_{d,c}(n,k)$ is then upper-bounded by

$$P_{d,c}(n,k) \leq 2^{-(n+k)}\, C_{d+1}(n+k). \tag{18}$$

It is well-known (see [16]) that one can write an explicit expression for $C_l(n)$ in terms of the roots of the reciprocal polynomial $P_l(z) = z^l - \sum_{i=0}^{l-1} z^i$ for $G_l(z)$. For $l > 2$, it is not possible to come up with an analytical expression for the roots of $P_l(z)$, and we must therefore use numerical approximations to obtain an upper bound on $C_l(n)$ which holds for all values of n. By ordinary functional analysis, it is easy to show that for $l \geq 2$, the polynomial $P_l(z)$ has a positive real root $\beta_l < 2 - 1/2^l$. A careful examination then reveals that for all $n \geq 0$ and $l \geq 2$

$$\frac{C_l(n)}{2^n} \leq \left(\frac{\beta_l}{2}\right)^{n-l+1} < \left(1 - \frac{1}{2^{l+1}}\right)^{n-l+1}. \tag{19}$$

Finally, (19), (18), and (14) combined result in

Theorem 5.1 For all $n \geq 1$,

$$P^*_{d,c}(n) \quad < \quad 2^{d+2}\left(1 - \frac{1}{2^{d+2}}\right)^{n-d}. \tag{20}$$

□

The bound is greater than one for small values of n, depending on d, but for large n it tends to zero exponentially fast. It is relatively close to the lower bound from Lemma 5.2. It is easy to see that if $n \geq (d+2)2^{d+2}$, then the bound in (20) is smaller than one. Consequently, we divide a binary string Y of length n into the runs of ones and zeros. Only the runs of length at least $(d+2)2^{d+2}$ count. If Y is purely random and n is large enough, then with probability arbitrarily close to one (see [5]), there are approximately $n/2^i$ constant runs of length at least $i, i \geq 1$. Therefore it follows that the number of bits contained in the runs of length at least $i, i \geq 1$, is $(i+1)/2^i$. Combining Lemma 5.1 with Theorem 5.1 it is then simple to prove

Theorem 5.2 For a purely random string Y of length n, for large enough n, with probability arbitrarily close to 1

$$P^*_{d,Y}(n) \quad < \quad \left(\left(1 - \frac{1}{2^{d+2}}\right)^{2^{-(d+2)2^{d+2}}}\right)^n. \tag{21}$$

□

Theorem 5.2 essentially asserts that given a random string Y, the probability that Y can be d-embedded into a random string X exponentially tends to zero with the string length. The result is applicable to the embedding divide and conquer attack on a clock-controlled shift register because its output sequence behaves like a random sequence. The corresponding minimum length of the observed sequence needed for a successful reconstruction is then approximately

$$n \quad \geq \quad r\, 2^{(d+2)(1+2^{d+2})} \ln 2 \tag{22}$$

which is linear in r but superexponential in d. This is a consequence of our theoretical approach, but in practice, experiments indicate that the minimum necessary length is linear in r and exponential in d. Note that from the lower bound in Lemma 5.2 it follows that if the length of the observed sequence satisfies approximately

$$n \quad < \quad r\, 2^{d+1} \ln 2, \tag{23}$$

then a successful initial state reconstruction is not possible.

6 Conclusion

In this paper, we define embedding and probabilistic correlation attacks on irregularly clocked shift registers and analyze them in two particular, unconstrained

and constrained cases. The objective is to identify the initial state of the shift register based on the known keystream sequence, without knowing the decimation/clocking sequence. All the attacks apply to an arbitrary binary clock-controlled shift register with not necessarily linear feedback that is clocked at least once per output symbol, and imply the exhaustive search over all possible shift register initial states. In the unconstrained embedding attack one allows an arbitrary number of deletions per output symbol, whereas in the unconstrained probabilistic attack one also assumes that the deletions take place independently with a given probability. The decimation sequence is assumed to be random and the corresponding deletion rate is defined as the relative expected number of deleted symbols. An analytical expression for the unconstrained embedding probability is derived using some combinatorial arguments. It is proved accordingly that the unconstrained embedding attack is successful only if the deletion rate is smaller than one half, in which case the minimum necessary length of the observed keystream sequence is shown to be linear in the shift register length r. It is then demonstrated how to compute recursively the joint probability needed for the unconstrained probabilistic attack with independent deletions. The attack is then analyzed in terms of the capacity [4] of the corresponding communication channel with independent deletions. It is thus shown that the unconstrained probabilistic attack is successful for any deletion rate smaller than one if the length of the known keystream sequence is greater than a minimum value linear in r. Apart from that, a lower bound on the capacity derived in [4] is improved by using the asymptotic properties of the unconstrained embedding probability. The results are then applied to the recently proposed 'shrinking' generator [2], which is basically an unconstrained clock-controlled shift register [15] with independent deletions with probability one half.

In the constrained embedding attack one allows at most d deletions per output symbol, for an arbitrary positive integer d. By using finite automata theory and generating functions, an upper bound on the constrained embedding probability for a constant string is derived and then employed to obtain an exponentially small upper bound on the constrained string embedding probability for a purely random string. Also, an exponential lower bound on the constrained embedding probability for any string is established as well. The results show that for any d the constrained embedding attack is successful if the length of the output sequence is greater than a value linear in r and superexponential in d, and is not successful if this length is smaller than a value linear in r and exponential in d. Consequently, by making d large one can not achieve the theoretical security against the constrained embedding attack, but can significantly improve the practical security.

References

[1] **W. G. Chambers and D. Gollmann. Lock-in effect in cascades of clock-controlled shift registers.** ***Advances in Cryptology, EUROCRYPT '88, Lecture Notes in Computer Science, vol. 330, C. G. Günther ed., Springer-Verlag,*** **pages 331–342, 1988.**

[2] D. Coppersmith, H. Krawczyk, and Y. Mansour. The shrinking generator. *Pre-Proceedings of CRYPTO '93*, pages 3.1–3.11, 1993.

[3] C. Ding, G. Xiao, and W. Shan. *The Stability Theory of Stream Ciphers.* Lecture Notes in Computer Science, vol. 561, Berlin: Springer–Verlag, 1991.

[4] A. S. Dolgopolov. Capacity bounds for a channel with synchronization errors. *Prob. Peredachi Inform. (in russian)*, 26:27–37, 1990.

[5] W. Feller. *An Introduction to Probability Theory and its Applications.* New York: Wiley, 3rd edition, Volume 1, 1968.

[6] J. Dj. Golić and M. J. Mihaljević. A generalized correlation attack on a class of stream ciphers based on the Levenshtein distance. *Journal of Cryptology*, 3(3):201–212, 1991.

[7] J. Dj. Golić and S. V. Petrović. A generalized correlation attack with a probabilistic constrained edit distance. *Advances in Cryptology, EUROCRYPT '92, Lecture Notes in Computer Science, vol. 658, R. A. Rueppel ed., Springer-Verlag*, pages 472–476, 1992.

[8] J. Dj. Golić and M. V. Živković. On the linear complexity of nonuniformly decimated PN-sequences. *IEEE Transactions on Information Theory*, 34:1077–1079, Sept. 1988.

[9] D. Gollmann and W. G. Chambers. Clock controlled shift registers: a review. *IEEE Journal on Selected Areas in Communications*, 7(4):525–533, 1989.

[10] D. Gollmann and W. G. Chambers. A cryptanalysis of $\text{step}_{k,m}$–cascades. *Advances in Cryptology, EUROCRYPT '89, Lecture Notes in Computer Science, vol. 434, J.-J. Quisquater, J. Vandewalle eds., Springer-Verlag*, pages 680–687, 1990.

[11] P. A. V. Hall and G. R. Dowling. Approximate string matching. *Computing Surveys*, 12:381–402, Dec. 1980.

[12] J. Hopcroft and J. Ullman. *An Introduction to Automata, Languages and Computation.* Reading, MA: Addison Wesley, 1979.

[13] R. Menicocci. Short Gollmann cascade generators are insecure. *Abstracts of the Fourth IMA Conference on Coding and Cryptography*, Cirencester, 1993.

[14] R. Menicocci. Cryptanalysis of a two-stage Gollmann cascade generator. *Proceedings of SPRC '93*, Rome, pages 62–69, 1993.

[15] M. J. Mihaljević. An approach to the initial state reconstruction of a clock-controlled shift register based on a novel distance measure. *Advances in Cryptology, AUSCRYPT '92, Lecture Notes in Computer Science, vol. 718, J. Seberry and Y. Zheng eds., Spinger-Verlag*, pages 349–356, 1993.

[16] F. Roberts. *Applied Combinatorics.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[17] M. V. Živković. An algorithm for the initial state reconstruction of the clock-controlled shift register. *IEEE Transactions on Information Theory*, 37:1488–1490, Sept. 1991.