New long-term *glimpse* of RC4 stream cipher

Subhamoy Maitra and Sourav Sen Gupta*

Indian Statistical Institute, Kolkata, India
subho@isical.ac.in, sg.sourav@gmail.com

Abstract. In 1996, Jenkins pointed out a correlation between the hidden state and the output keystream of RC4, which is well known as the *Glimpse* theorem. With a permutation of size N-bytes, the probability of guessing one location by random association is $1/N$, whereas the existing correlations related to *glimpse* allow an adversary to guess a permutation location, using the knowledge of the keystream output bytes, with probability $2/N$. To date, this is the best known state-leakage based on glimpse. For the first time in RC4 literature, we show that there are certain events that leak state information with a probability of $3/N$, considerably higher than the existing results. Further, the new glimpse correlation that we observe is a *long-term* phenomenon; it remains valid at any stage of the evolution of RC4 Pseudo Random Generation Algorithm (PRGA). This new glimpse with a considerably higher probability of state-leakage may potentially have serious ramifications towards state-recovery attacks on RC4.

Keywords: stream cipher, RC4, glimpse, long-term, correlation

1 Introduction

Over the last three decades of research in stream ciphers, several designs have been proposed and analyzed by the community. The RC4 stream cipher, 'allegedly' designed by Rivest in 1987, has sustained to be one of the most popular ciphers in this category for more than 25 years. The cipher has continued gaining its fabled popularity for its intriguing simplicity that has made it widely accepted in the community for various software and web applications.

The cipher consists of two major components, the Key Scheduling Algorithm (KSA) and the Pseudo-Random Generation Algorithm (PRGA). The internal permutation of RC4 is of N bytes, and so is the key K. The original secret key is of length typically between 5 to 32 bytes, and is repeated to form the final key K. The KSA produces the initial permutation of RC4 by scrambling an identity permutation using key K. The initial permutation S produced by the KSA acts as an input to the next procedure PRGA that generates the output keystream. The RC4 algorithm is as shown in Fig. 1.

* Supported by DRDO sponsored project Centre of Excellence in Cryptology (CoEC), under MOC ERIP/ER/1009002/M/01/1319/788/D(R&D) of ER&IPR, DRDO.

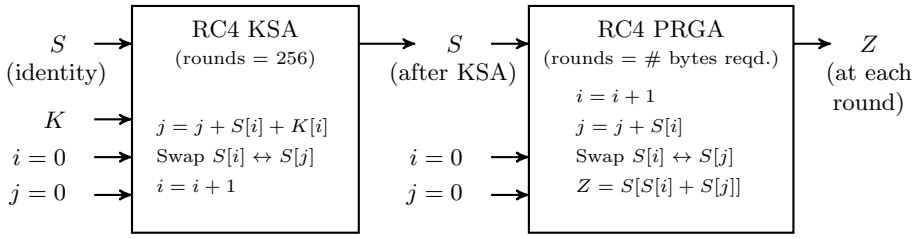

Fig. 1. Key-Scheduling Algorithm and Pseudo-Random Generation Algorithm of RC4.

1.1 Notation and Assumptions

For round $r \geq 1$ of RC4 PRGA, we denote the indices by i_r, j_r, the output byte by Z_r, the index location of output Z_r as t_r, and the permutations before and after the swap by S_{r-1} and S_r respectively. Thus, round r of RC4 is defined by the operations

$$i_r = i_{r-1} + 1; \quad j_r = j_{r-1} + S_{r-1}[i_r]; \quad \text{Swap } S_{r-1}[i_r] \leftrightarrow S_{r-1}[j_r];$$
$$t_r = S_r[i_r] + S_r[j_r]; \quad Z_r = S_r[t_r].$$

The initial permutation of PRGA is denoted by S_0, and all arithmetic operations in the context of RC4 are to be considered modulo N.

During the course of this paper, we shall assume uniform randomness of certain events for the proofs. In most of these cases, the randomness assumption will be based on natural pseudo-randomness assumptions of the RC4 stream cipher, as appropriate. In some cases, we shall assume randomness based on experimental evidences, run over atleast a billion trials of RC4 with random keys. For all such assumptions, a random association probability $1/N$ will be assumed if there is no significant bias, of the order $1/N$ or similar. Some of these events may have prominent biases when treated conditionally with certain other events, but we shall only treat them in their unconditional forms, where they exhibit no significant biases. We shall state, and justify if required, the randomness assumptions as and when required in this paper.

1.2 Motivation and Contribution

In 1996, Jenkins [4] pointed out that the RC4 keystream provides a glimpse of the RC4 state as follows, which is known as Glimpse theorem or Jenkins' correlation. We present the complete proof of the theorem for clarity.

Theorem 1 (Glimpse theorem). *After the r-th round of RC4 PRGA, for $r \geq 1$, we have*

$$\Pr(S_r[j_r] = i_r - Z_r) \;=\; \Pr(S_r[i_r] = j_r - Z_r) \;\approx\; \frac{2}{N}.$$

Proof. To prove this result, one needs to use the paths $i_r = S_r[i_r] + S_r[j_r]$ and $j_r = S_r[i_r] + S_r[j_r]$ respectively. Note that

$$i_r = S_r[i_r] + S_r[j_r] \;\Rightarrow\; Z_r = S_r[i_r] = i_r - S_r[j_r], \text{ and}$$
$$j_r = S_r[i_r] + S_r[j_r] \;\Rightarrow\; Z_r = S_r[j_r] = j_r - S_r[i_r].$$

Thus, one may evaluate $\Pr(S_r[j_r] = i_r - Z_r)$ as

$$\begin{aligned}&\Pr(Z_r = i_r - S_r[j_r] \mid i_r = S_r[i_r] + S_r[j_r]) \cdot \Pr(i_r = S_r[i_r] + S_r[j_r])\\ &+ \Pr(Z_r = i_r - S_r[j_r] \mid i_r \neq S_r[i_r] + S_r[j_r]) \cdot \Pr(i_r \neq S_r[i_r] + S_r[j_r])\\ &\approx 1 \cdot 1/N + 1/N \cdot (1 - 1/N) \approx 2/N,\end{aligned}$$

where it is assumed that the desired event $(S_r[j_r] = i_r - Z_r)$ occurs with the probability of random association $1/N$ if $i_r \neq S_r[i_r] + S_r[j_r]$. One may prove the bias in $(S_r[i_r] = j_r - Z_r)$ similarly. □

One may note that this glimpse correlation can be observed at any point of the RC4 keystream. Later, in Asiacrypt 2005, Mantin [6] has also explored a general set of similar events in this direction that leak state information with probability more than that of random association. There exist several related works that look only at the initial keystream bytes of RC4 to obtain information regarding the state and eventually the secret keys (a few recent examples are in [13]). However, these observations never work in the long term scenario.

The question that we ask here is:

> *"Can one discover a correlation between the RC4 keystream and the state that offers a glimpse with a probability significantly more than $2/N$ in long term evolution of the cipher?"*

We answer to this question affirmatively. We prove the following: given that two consecutive bytes Z_r, Z_{r+1} of RC4 are equal to the specific value $(r + 2)$ during the consecutive two rounds r and $r + 1$ (modulo N), the probability that the $(r + 1)$-th location of the state array during round r (denoted as $S_r[r + 1]$ as per our notation) will be equal to $(N - 1)$ is $3/N$, significantly higher than the probability of random association $1/N$. The result is presented in Section 2.

2 Long-term *glimpse* of RC4

We start with our most important observation which we made while trying to obtain the scenario where the S array comes back to the same permutation after two consecutive rounds.

2.1 The main observation motivating our result

As one may note in Fig. 2, if in the r-th round, $j_r = i_r + 1$ and $S_r[j_r] = N - 1$, then the two places swapped in round $(r + 1)$ will be restored in round $(r + 2)$. That is, we shall have S_{r+2} identical to S_r in such a case. This motivated our first result, as in Theorem 2.

Theorem 2. *After the r-th round ($r \geq 1$) of RC4 PRGA, we have*

$$\Pr(S_r[r+1] = N-1 \mid Z_{r+1} = Z_r) \;\approx\; \frac{2}{N}.$$

Proof. We shall first prove $\Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1) \approx 2/N$, and then apply Bayes' theorem to get the desired result. The condition $S_r[r+1] = N-1$, and the path $j_r = r+1$ results in $j_{r+1} = j_r + S_r[r+1] = r+1+N-1 = r$, which eventually gives

$$\begin{aligned} t_{r+1} &= S_{r+1}[i_{r+1}] + S_{r+1}[j_{r+1}] \\ &= S_r[j_{r+1}] + S_r[i_{r+1}] \\ &= S_r[r] + S_r[r+1] \\ &= S_r[i_r] + S_r[j_r] = t_r. \end{aligned}$$

Thus, $Z_{r+1} = S_{r+1}[t_{r+1}] = S_{r+1}[t_r]$ is equal to $Z_r = S_r[t_r]$ in almost all cases, except when t_r equals either i_{r+1} or j_{r+1}, the only two locations that get swapped in transition from S_r to S_{r+1}. Thus,

$$\Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1 \;\wedge\; j_r = r+1) \approx 1.$$

This scenario is as illustrated in Fig. 2.

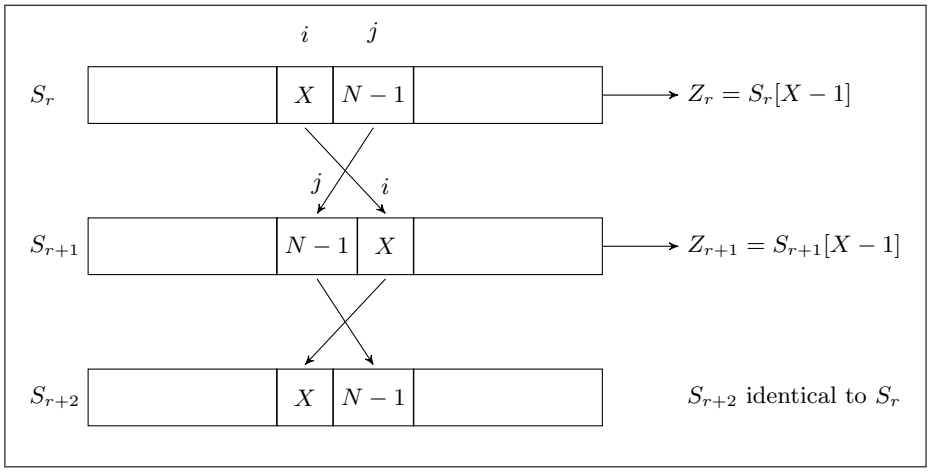

Fig. 2. The scenario for $(Z_{r+1} = Z_r \mid S_r[r+1] = N-1 \;\wedge\; j_r = r+1)$.

We may now evaluate $\Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1)$ as

$$\begin{aligned} &\Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1 \;\wedge\; j_r = r+1) \cdot \Pr(j_r = r+1) \\ &+ \Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1 \;\wedge\; j_r \neq r+1) \cdot \Pr(j_r \neq r+1) \\ &\approx 1 \cdot 1/N + 1/N \cdot (1 - 1/N) \approx 2/N. \end{aligned}$$

Applying Bayes' theorem to the above result, we obtain

$$\begin{aligned}&\Pr(S_r[r+1] = N-1 \mid Z_{r+1} = Z_r) \cdot \Pr(Z_{r+1} = Z_r)\\ &= \Pr(Z_{r+1} = Z_r \mid S_r[r+1] = N-1) \cdot \Pr(S_r[r+1] = N-1)\\ &\approx 2/N \cdot 1/N.\end{aligned}$$

Assuming pseudo-randomness of RC4 keystream bytes, we may write $\Pr(Z_{r+1} = Z_r) \approx 1/N$ (experimentally verified over a billion trials). This gives $\Pr(S_r[r+1] = N-1 \mid Z_{r+1} = Z_r) \approx 2/N$. □

Thus the event $(Z_{r+1} = Z_r)$ leaks the information of a single permutation location with probability twice that of random association.

2.2 Corollary of the Glimpse theorem from [4]

Before proceeding further, we would like to point out a simple corollary of the Glimpse theorem (Theorem 1) that leaks the information of a permutation location with the same probability.

Corollary 1. *After the r-th round of RC4 PRGA, for $r \geq 1$, we have*

$$\Pr(S_r[r+1] = N-1 \mid Z_{r+1} = r+2) \;\approx\; \frac{2}{N}.$$

Proof. In RC4 transition between rounds r and $r+1$, we have $i_{r+1} = r+1$, and $S_{r+1}[j_{r+1}] = S_r[i_{r+1}] = S_r[r+1]$, due to the swap in round r. Thus, by the Glimpse theorem (Theorem 1), we have

$$\Pr(S_r[r+1] = r+1-Z_{r+1}) \;\approx\; 2/N.$$

In case of $Z_{r+1} = r+2$, we get the desired conditional result. □

2.3 The main result of this paper

In the scenarios presented in Theorem 2 and Corollary 1, we find two different cases that leak the value of a specific location in the S array, namely $S_r[r+1]$, with probability $2/N$ in each case. Moreover, the two events seem to be unrelated, or atleast not completely dependent. Thus, it is quite natural to expect that considering the events together, one may have better confidence about the value in that specific location $S_r[r+1]$. In this direction, we present our main result of this paper in the form of Theorem 3.

Theorem 3. *After the r-th round ($r \geq 1$) of RC4 PRGA, we have*

$$\Pr(S_r[r+1] = N-1 \mid Z_{r+1} = Z_r \;\wedge\; Z_{r+1} = r+2) \;\approx\; \frac{3}{N}.$$

Proof. Let us define the main events as follows:

$$A := (S_r[r+1] = N-1),\ B := (Z_{r+1} = Z_r),\ C := (Z_{r+1} = r+2).$$

The result requires $\Pr(A|B \wedge C)$, and it seems that a naive composition of Theorem 2 (which gives $\Pr(A|B)$) and Theorem 1 (which gives $\Pr(A|C)$) will produce the desired result. However, this is not the case. If we try to compute $\Pr(A \wedge B \wedge C)$ as $\Pr(B \wedge C|A)\Pr(A)$, then the first part is not easily computable as events B and C, conditional to event A, are not independent (verified experimentally over a billion trials). Hence we try the following route.

$$\Pr(A \wedge B \wedge C) = \Pr(C|B \wedge A) \cdot \Pr(B|A) \cdot \Pr(A).$$

Still there remains a problem with the first part, as event C occurs simultaneously with the occurrence of Z_{r+1} in event B. This is easy to observe experimentally, but not so easy to prove in theory.

To avoid the aforesaid problem in computing $\Pr(C|B \wedge A)$, we rewrite the problem definition slightly, and try to prove

$$\Pr(S_r[r+1] = N-1 \mid Z_r = r+2 \ \wedge \ Z_{r+1} = r+2) \ \approx \ 3/N.$$

We compute this as $\Pr(A \wedge B' \wedge C) = \Pr(C|B' \wedge A) \cdot \Pr(B'|A) \cdot \Pr(A)$, where $A := (S_r[r+1] = N-1)$, $C := (Z_{r+1} = r+2)$ as before, and $B' := (Z_r = r+2)$. Now we may compute $\Pr(C|B' \wedge A)$ easily, as event C occurs after completion of both the events A and B'.

Computing $\Pr(C|B' \wedge A)$*:* Note that event $A := (S_r[r+1] = N-1)$ implies $Z_{r+1} = S_{r+1}[S_{r+1}[r+1] + S_r[r+1]] = S_{r+1}[S_{r+1}[r+1] - 1]$. And of course, event $B' := (Z_r = r+2)$ implies $Z_r = S_r[t_r] = r+2$. We consider the following paths for the proof.

- *Case I:* $(S_{r+1}[r+1] = r+2)$. In case of this path, we shall have $Z_{r+1} = S_{r+1}[(r+2)-1] = S_{r+1}[r+1] = r+2$, with probability of occurrence 1.
- *Case II:* $(S_{r+1}[r+1] = t_r + 1)$. In case of this path, we shall have $Z_{r+1} = S_{r+1}[(t_r+1)-1] = S_{r+1}[t_r] = S_r[t_r] = r+2$, with probability of occurrence approximately 1, disregarding the two cases when t_r may be equal to either i_{r+1} or j_{r+1}.

In almost all other cases, we may assume that $C := (Z_{r+1} = r+2)$ happens with probability of random association $1/N$ (verified experimentally over a billion trials). We compute $\Pr(C|B' \wedge A)$ as

$$\begin{aligned}
&\Pr(C|B' \wedge A \wedge (S_{r+1}[r+1] = r+2)) \cdot \Pr(S_{r+1}[r+1] = r+2) \\
&+ \Pr(C|B' \wedge A \wedge (S_{r+1}[r+1] = r+2)) \cdot \Pr(S_{r+1}[r+1] = t_r + 1) \\
&+ \sum_{\substack{X \neq r+2 \\ X \neq t_r+1}} \Pr(C|B' \wedge A \wedge (S_{r+1}[r+1] = X)) \cdot \Pr(S_{r+1}[r+1] = X) \\
&\approx 1 \cdot 1/N + 1 \cdot 1/N + (1 - 2/N) \cdot 1/N \approx 3/N.
\end{aligned}$$

Computing $\Pr(A|B' \wedge C)$*:* As no glimpse-like connection has been found between $S_r[r+1]$ and Z_r in the literature to date, we may assume $\Pr(B'|A) \approx 1/N$ (verified experimentally over a billion trials), and we may of course take $\Pr(A) \approx 1/N$ as per natural pseudo-randomness assumptions of RC4. Thus,

$$\Pr(A \wedge B' \wedge C) = \Pr(C|B' \wedge A) \cdot \Pr(B'|A) \cdot \Pr(A) \approx 3/N \cdot 1/N \cdot 1/N.$$

We may assume $\Pr(B' \wedge C) = \Pr(B') \cdot \Pr(C) \approx 1/N \cdot 1/N$ (verified experimentally over a billion trials), and this produces the desired conditional result $\Pr(A|B \wedge C) = \Pr(A|B' \wedge C) \approx 3/N$. □

2.4 Experimental results

We have performed extensive experiments to obtain accurate practical estimates of each of the results presented in this paper. Each correlation reported in this paper is of order $1/N$ with respect to a base event of probability $1/N$. Thus, $O(N^3)$ trials are sufficient to identify the biases with considerable probability of success (refer to [5, 9] for detailed explanation on the complexity).

The experimental results presented in this section are based on an average of N^4 trials of RC4, in each case, with keys chosen uniformly at random. The experiments were carried out using GCC-compiled C-code on a Unix machine with 3.34 GHz processor and 8 GB of memory. Table 2.4 lists the theoretical estimates against the experimental values for each of the results presented in Section 2.

Table 1. Experimental values and theoretical estimates pertaining to our results, where $A := (S_r[r+1] = N-1)$, $B := (Z_{r+1} = Z_r)$ and $C := (Z_{r+1} = r+2)$.

Biased Event	Probability (experimental value)	Probability (theoretical estimate)	Result (as in Section 2)
$(A \mid B)$	0.0077881670	$2/N = 0.0078125$	Theorem 2
$(A \mid C)$	0.0078166422	$2/N = 0.0078125$	Corollary 1
$(A \mid B \wedge C)$	0.0117323766	$3/N = 0.01171875$	Theorem 3

The values presented in Table 2.4 testify that our theoretical estimates for the higher-order glimpse correlation and associated results closely match their respective experimental values. Slight deviations, if any, are due to marginal gaps of order $1/N^2$ or less, which we have purposefully disregarded in case of the theoretical results.

2.5 Discussion of our results

The glimpse correlations have been quite well studied in RC4 literature, as they provide practical leaks into the state permutation of the cipher from the knowledge of the output keystream. Glimpse correlations can be exploited towards

state-recovery and key-recovery attacks on RC4. One may find some important results in state-recovery attacks on RC4 in [7, 3], and a few attacks along the lines of RC4 key-recovery from the permutation in [8, 2, 1].

Although glimpse biases provide practical cryptanalytic tools against RC4, not many have been identified over the last two decades of analysis. Jenkins [4] was the first to report a glimpse into RC4 state from the keystream with probability $2/N$, and it has since been the best one that persists in the long-term evolution of the PRGA. Later in 2001 and 2005, Mantin [5, 6] generalized the glimpse correlations into 'useful states' of RC4, which included Jenkins' correlations as a special case. These biases were again of magnitude $2/N$, and persisted in the long-term evolution of PRGA. In recent times, several correlations between the state permutation and keystream have been observed, mainly by Sepehrdad et al [12, 13], and later proved by Sen Gupta et al [10, 11]. Although these correlations are larger in magnitude, none persist in the long-term evolution of RC4 PRGA, and only pertain to the initial bytes of the output.

Our result in this paper provides the following.

Strong long-term glimpse correlation: It provides a long-term glimpse correlation of magnitude $3/N$, the best to date. It is interesting to note that no long-term glimpse bias of magnitude more than $2/N$ has been reported in the literature over the last 15 years, since the first one [4] in 1996.

Guessing single permutation location using two output bytes: The long-term glimpse correlations reported in the literature to date generally relate a keystream output byte to a single location of the state permutation, typically at a specific round of RC4. Thus, simultaneous knowledge of two or more keystream bytes may help in guessing two or more permutation locations, but does not always provide additional benefits in guessing a single location of the permutation over any one of them. Our result combines the knowledge of two consecutive output bytes Z_r, Z_{r+1} to obtain a significant advantage in guessing a single permutation location $S_r[r+1]$. To the best of our knowledge, such a correlation has never been proposed in the literature.

3 Conclusion

In this paper we have shown that there exist long term correlations during the evolution of RC4 PRGA, even with a higher magnitude compared to the existing Jenkins' correlations [4], leaking information (providing a glimpse) about certain locations in the S array from the knowledge of the keystream output bytes.

The new glimpse association that we prove occurs with a probability of $3/N$, which is considerably higher than the probability of random association $1/N$, as well as higher in magnitude compared to the best known existing glimpse correlation probability $2/N$, as in the current literature [4].

Acknowledgments. The authors would like to thank the anonymous reviewers for their valuable comments that helped improve the quality of the paper.

References

1. M. Akgün, P. Kavak, and H. Demirci, "New Results on the Key Scheduling Algorithm of RC4," in *INDOCRYPT '08*, vol. 5365 of *Lecture Notes in Computer Science*, pp. 40–52, 2008.
2. E. Biham and Y. Carmeli, "Efficient Reconstruction of RC4 Keys from Internal States," in *FSE '08*, vol. 5086 of *Lecture Notes in Computer Science*, pp. 270–288, 2008.
3. J. D. Golic and G. Morgari, "Iterative Probabilistic Reconstruction of RC4 Internal States," *IACR Cryptology ePrint Archive*, Report 2008/348, 2008. Available at `http://eprint.iacr.org/2008/348`.
4. R. J. Jenkins, "ISAAC and RC4," 1996. Published on the Internet at `http://burtleburtle.net/bob/rand/isaac.html` [last accessed on December 28, 2012].
5. I. Mantin, "Analysis of the stream cipher RC4," Master's thesis, The Weizmann Institute of Science, Israel, 2001. Available at `http://www.wisdom.weizmann.ac.il/~itsik/RC4/rc4.html`.
6. I. Mantin. A Practical Attack on the Fixed RC4 in the WEP Mode. ASIACRYPT 2005, LNCS, Springer-Verlag, Vol. 3788, pp. 395–411, 2005.
7. A. Maximov and D. Khovratovich, "New State Recovery Attack on RC4," in *CRYPTO '08*, vol. 5157 of *Lecture Notes in Computer Science*, pp. 297–316, 2008.
8. G. Paul and S. Maitra, "Permutation After RC4 Key Scheduling Reveals the Secret Key," in *SAC '07*, vol. 4876 of *Lecture Notes in Computer Science*, pp. 360–377, 2007.
9. G. Paul and S. Maitra. RC4 Stream Cipher and Its Variants. CRC Press, 1st Edition, November 16, 2011.
10. S. Sen Gupta, S. Maitra, G. Paul and S. Sarkar. "Proof of Empirical RC4 Biases and New Key Correlations," in *SAC '11*, vol. 7118 of *Lecture Notes in Computer Science*, pp. 151–168, 2011.
11. S. Sen Gupta, S. Maitra, G. Paul and S. Sarkar. "(Non-)Random Sequences from (Non-)Random Permutations - Analysis of RC4 stream cipher". To appear in Journal of Cryptology (Springer), accepted November 3, 2012.
12. P. Sepehrdad, S. Vaudenay, and M. Vuagnoux, "Discovery and Exploitation of New Biases in RC4," in *SAC '10*, vol. 6544 of *Lecture Notes in Computer Science*, pp. 74–91, 2011.
13. P. Sepehrdad, S. Vaudenay, and M. Vuagnoux, "Statistical Attack on RC4 - Distinguishing WPA," in *EUROCRYPT '11*, vol. 6632 of *Lecture Notes in Computer Science*, pp. 343–363, 2011.