Journal of the Indian Statistical Association
Vol.42, 2, 2004, 159-178

REPETITION CODES CRYPTANALYSIS OF BLOCK CIPHERS

Eric Filiol
Army Signals Academy
Virology and Cryptology Laboratory, France

Abstract

This paper presents a new theoretical model of block cipher cryptanalysis based on the use of a well-known error-correcting code: the repetition codes. We first demonstrate how to describe a block cipher with such a code before explaining how to design a new ciphertext only cryptanalysis of these cryptosystems on the assumption that plaintext belongs to a particular class. We then show how known plaintext linear cryptanalysis can be generalized by a more general repetition approach. Two cryptanalysis algorithms are presented. The first one uses a single repetition code while the second one uses concatenated codes whose outer and inner codes are repetition codes. We compare the two algorithms and prove that the first one is more efficient than the second one. Open problems and technical parameters are finally given as well as improvements on Matsui's DES cryptanalysis.

Key Words: *Block cipher, Data Encryption Standard, cryptanalysis, coding theory, repetition codes*

1 Introduction

In October 2000, the NIST has selected the Rijndaël block cipher as the *Advanced Encryption Standard* (AES) to replace the DES block cipher and extent it to a massive world-wide usage. The growing dependence of the commercial community on block ciphers -for its data security functions- make it desirable to keep under review the strength of this kind of encryption systems.

The evaluation of the AES, as well as for the other finalists [1], has been essentially based on the the former cryptanalysis or their variant forms: differential cryptanalysis [2], linear cryptanalysis [10], ... and no significative results were likely to question their strength. Finally we must admit that security consideration as a key point in the final choice was not so relevant as we could have imagined since all of the finalists offer a suitable high security. To quote Adi Shamir [14], "*any new real life cryptanalysis which may appear in the future will equally challenge the finalists*".

On the other hand, the future seems to favour block encryption, at least on the trade level. Few stream ciphers are known or proposed whereas meanwhile many block systems are proposed (17 block cipher systems for only 5 stream ciphers have been suggested for the *New European Schemes for Signature, Integrity and Encryption* (NESSIE) project [13]). As for the AES, only block ciphers were requested. Though we can strongly affirm that a very consequent theory for stream encryption exists, the block encryption theory does not provide more than a few cryptanalytic techniques and results on the constituent primitives at the round level. A rigorous and global description of formalization of a whole system, including a combinatorial approach in particular, is still to come. In other words, who can affirm that hiding a trap, for example, is totally impossible without being detected (this has still

been more or less an open question for the DES); and what about the existence of particular global mask values on input and output which could drastically improve linear cryptanalysis techniques. The authors of AES acknowledge this second fact [4, Chap 7 and paragraph 2 of page 124]aesbk, which moreover is also relevant for any cryptosystem.

Actually, most of cryptanalysis capacity depends on the ability of detecting these high correlations if there are some. In real-life cryptanalysis it is not so much the maximum average correlation potential that is relevant but the maximum correlation potential corresponding to the given key under attack [4]. Our experience in cryptanalysis shows us that very often it is more interesting and efficient to consider this potential when considering a particular class of plaintext. In case of block ciphers, this approach is particularly efficient since plaintext represents an active part in the production of the block cipher. This fact has recently been pointed out by the statistical analysis of the Algebraic Normal Form of Boolean functions modeling a block cipher [6].

In this paper we intend to introduce a new theoretical model of block cipher cryptanalysis related to this approach. On the assumption that a given subset of plaintext space has been encrypted and that consequently, particular, higher correlation properties exist between only the resulting subset of ciphertexts and any key of the key space, we design an attack using repetition codes on ciphertext blocks only. This cryptanalysis is called *Plaintext-dependent Repetition Codes Cryptanalysis* (PDRC attack for short). It differs from a classical chosen-plaintext attack as we do not have to choose or even know any of the plaintext blocks. Moreover, a PDRC attack uses only ciphertext blocks. Thus the difficulty is to find suitable properties that leaks information about the key from the ciphertext. We then show that it is possible to generalize linear cryptanalysis when considering the plaintext blocks too. In this case, we consider *Repetition Codes Cryptanalysis* (RC cryptanaly-

sis for short) which is itself a generalization of the PDRC attack. It is shown that DES linear cryptanalysis can be significantly improved with RC attack.

This paper is organized as follows. Section 2 presents theoretical preliminaries and notation. Then Section 3 details the formal model of the new cryptanalysis based on repetition codes. In particular we give a combinatorial resistance criterion against PDRC attack and formulate open problems relatively to the PDRC attack. Section 4 illustrates this approach by considering an theoretical block cipher and explaining how to practically implement PDRC attack. Section 5 presents the RC attack as the generalization of both the linear cryptanalysis and the PDRC attack. Section 6 concludes while presenting open problems and future studies.

2 Background Theory and Notation

2.1 Repetition Codes

Let us consider a *Binary Symmetric Channel* (BSC) of parameter p used to transmit messages over a binary alphabet. Its transition probability matrix is the square matrix of order 2 whose coefficients are given by $a_{i,j} = q$ whenever $i \neq j$ and $a_{i,j} = p = 1 - q$ otherwise.

In other words, if an emitter sends bit b_t then $\hat{b}_t = b_t \oplus e_t$ will be effectively received with probability p (channel error probability). To recover from transmission errors one uses error-correcting codes and in particular linear codes. A binary linear code $[n, k, d]$ is a vector subspace of $\mathbb{F}_2^n$, of dimension k. Its *minimal distance* d is the minimum Hamming weight of all non zero codewords (that is to say the n-bit vectors). In other words $d = \min_{x \in \mathbb{F}_2^n} \{wt(x)\}$ where $wt(x)$ denotes the number of non zero positions in $x = (x_1, \ldots, x_n)$. Then a well-known result [9] defines the number of errors on a codeword that can be cor-

rected by a code of minimal distance d as $\frac{d-1}{2}$.

A n-repetition code, on a set of two symbols, is a $[n, 1, n]$ linear code and consists of two codewords, each one of them is made up of n identical symbols. Whenever $q > p$, maximum likelihood decoding (MLD) amounts to find out in the received vector which symbol is repeated most. The vector will be decoded as 0 if its Hamming distance to null vector is less than its distance to vector $(1, 1, 1, \cdots, 1)$, otherwise it is decoded as 1. Thus MLD reduces to majority decoding.

Example 1 : Let us consider the message 01100 and a 3-repetition code. Then the sequence 000 111 111 000 000 is transmitted. The sequence 010111101110100 is received and decoded as 01110. There is one residual error.

These codes are the most easily decodable among codes ensuring a high protection. Moreover, repetition codes are the most efficient ones when dealing with high noise probability p [10].

Proposition 1 [12] Let $n = 2s + 1$. Then the n repetition code is correcting at most s errors and is a perfect code. Its bit error probability (residual decoding error) is given by

$$P_{err} = \sum_{i=s+1}^{n} \binom{n}{p} p^i \cdot q^{n-i}. \tag{1}$$

The term *perfect* means that every words in the "ambient" space $\mathbb{F}_2^n$ is decodable for maximum likelihood as in a perfect block code. Finally the probability of successful decoding is given by

$$P_{succ} = 1 - P_{err}.$$

It is worth noticing that if $p < \frac{1}{2}$ the $P_{err,2s+1}$ tends towards 0 as $s \to \infty$.

2.2 Block Ciphers and Linear Cryptanalysis

A block cipher working on m-bit plaintext blocks P_i with a n-bit secret key K ((m, n)-block cipher for short) is a mapping from $\mathbb{F}_2^m \times \mathbb{F}_2^n$

to $\mathbb{F}_2^m$. Each time a given key K is chosen, the resulting mapping restriction is a permutation over $\mathbb{F}_2^m$. A block cipher is thus a set of 2^n permutations over $\mathbb{F}_2^m$. Note that it represents a very small subset of all these permutations ($(2^m)!$ in total).

Linear cryptanalysis [10] of block ciphers is a known plaintext attack in which a very large number of plaintext-ciphertext pairs are used to determine the value of a subset of key bits, thus greatly reducing the exhaustive search part.

A condition for applying linear cryptanalysis to such a block scheme is to find "effective", probabilistic linear expressions between any plaintext block P_i, any ciphertext block C_i and any key K of the form:

$$< P_i, u > \oplus < C_i, w > \overset{p}{=} < K, v > \tag{2}$$

where $< ., . >$ denotes the usual scalar product over $\mathbb{F}_2^m$. If this equation holds with a probability $p \neq \frac{1}{2}$ then by checking the left-hand side of Equation (2) for a large number N of plaintext-ciphertext pairs, the right-hand side of this equation may be guessed by a simple maximum likelihood decoding. A single information bit about the key is obtained. This cryptanalysis is effective if the deviation $|p - \frac{1}{2}|$ is large enough. In [10], it is shown that the probability of unsuccessful guessing is very small as soon as $N > |p - \frac{1}{2}|^{-2}$.

Generally the linear approximation described by Equation (2) is obtained by "chaining" single-round linear approximations obtained by considering statistical biases in the constituent primitives. This implies that other, possibly higher correlations that are depending on the global structure of the systems are out of analysis capabilities [4, Chap 7 and paragraph 2 of page 124].

3 Repetition Codes Cryptanalysis of Block Ciphers

3.1 Block Ciphers and Repetition Codes

Let us consider a given property $\mathcal{I}$ and let us denote $P_{\mathcal{E}}[\mathcal{I}]$ the probability of $\mathcal{I}$ to be satisfied on set $\mathcal{E}$. Then a block cipher can be broken if we have, for some $\mathcal{I}$, $P_{\mathbb{F}_2^{m+n}}[\mathcal{I}] \neq \frac{1}{2}$.

Each key K in the key space $\mathcal{K} = \mathbb{F}_2^n$ selects a corresponding permutation over $\mathbb{F}_2^m$. Thus K may be recover if $P_{\mathbb{F}_2^m}[\mathcal{I}_K] \neq \frac{1}{2}$ where $\mathcal{I}_K$ denotes the property $\mathcal{I}$ related to the key K. Then we may dispose of an attack if we can exhibit such a property verified for any $K \in \mathcal{K}$ (denoted $\mathcal{I}_{\mathcal{K}}$). For linear cryptanalysis, $\mathcal{I}_{\mathcal{K}}$ is a particular linear probabilistic equation.

Let us now consider the plaintext space $\mathcal{P} = \mathbb{F}_2^m$ and a partition $(\mathcal{P}_i)_{i \leq 2^k}$ of $\mathcal{P}$ for some $k \in \mathcal{N}$. Without loss of generality we suppose that $|\mathcal{P}_i| = 2^{m-k}$ for all i. Now suppose there exists (possibly many) $\mathcal{P}_i$ such that $\mathcal{I}_{\mathcal{P}_i}[\mathcal{I}_{\mathcal{K}}] = p_i \neq \frac{1}{2}$. Since the encryption key $K \in \mathcal{K}$ remains the same for all the plaintext blocks, we may compare the encryption process as a Binary Symmetric Channel (BSC) with parameter p_i where the noise is produced by the plaintext blocks from $\mathcal{P}_i$ (see Figure 1). The BSC is directly and closely determined by $\mathcal{P}_i$. The noisy version $\widehat{\mathcal{I}_K}$

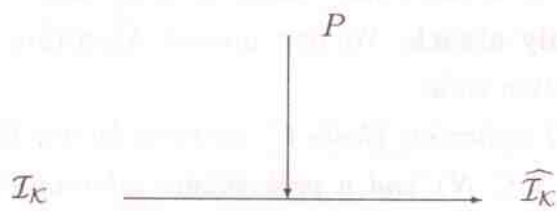

Figure 1: Block Cipher and Binary Symmetric Channel

of $\mathcal{I}_K$ is a (possibly complex) function $f(C)$ of ciphertext blocks C. In

other words encrypting N plaintext blocks $P \in \mathcal{P}_i$ may be equivalently defined as transmitting $\mathcal{I}_K$ by means of a N repetition code through a BSC of parameter p_i. From Figure 1, it means that over $\mathcal{C}_i$ we have $P[\mathcal{I}_K = \widehat{\mathcal{I}_K}] = 1 - p_i$.

The aim of the designer is to obtain a set of permutations over $\mathcal{C}$ such that no obvious properties $\mathcal{I}$ leaks information about the key. But the situation is likely to be very different when considering a restriction to a subset $\mathcal{C}_i \subseteq \mathcal{C}$. If we have

$$P_{\mathcal{C}}[\mathcal{I}] = \sum_{i=0}^{2^k} P_{\mathcal{C}_i}[\mathcal{I}] \cdot P[\mathcal{C}_i] = \frac{1}{2}$$

we however may have many $P_{\mathcal{C}_i}[\mathcal{I}]$ different from $\frac{1}{2}$ (it suffices that $\sum_i \epsilon_i = \sum_i (p_i - \frac{1}{2}) = 0$). This fact seems to be partly explained by the fact that the actual number of permutations over $\mathcal{C}$ effectively represented by a block cipher is extraordinary negligible compared of the total number of permutations over the same plaintext space.

3.2 Description of the PDRC Attack

With the setting defined in the previous section, we now can describe the plaintext-dependent repetition code cryptanalysis, very simply. Note, once again, that local independance from the plaintext (due to the restriction to a particular subset $\mathcal{C}_i \subseteq \mathcal{C}$) allows us to design a **ciphertext only attack**. We first present Algorithm A.1 which uses only one repetition code.

Input: *N odd) ciphertext blocks C_j encryted by key K from plaintext $P_j \in \mathcal{C}_i$ $(1 \leq j \leq N)$ and a probabilistic information $\mathcal{I}_{\mathcal{K}}$ such that $\mathcal{I}_{\mathcal{K}} \overset{p_i}{\cong} f(C_j)$ for some f and for all j.*

Output: *Exact value $\mathcal{I}(K)$ for the actual key.*

1. *Initialize counter* $ct \leftarrow 0$.

2. *For each of the N ciphertext blocks C_j*

   (a) *Compute $f(C_i)$.*

   (b) *If $f(C_i) = 1$ then $ct++$.*

3. *end for*

4. *If $ct \geq \frac{N+1}{2}$ then $\mathcal{I}(K) = 1$ else $\mathcal{I}(K) = 0$.*

Complexity of algorithm A.1 is easy to evaluate. It performs only N evaluations of f. Thus complexity is $\mathcal{O}(N)$. Since N is the length of the repetition code, according to Section 2.1, it depends only on p_i and p_{succ}, the probability of successful guessing for $\mathcal{I}(K)$.

To the knowledge of the author there does not exist a general formula for N directly from parameters p_i and p_{succ}. We can only tabulate results for fixed values of them. It is a well-known fact that for a fixed p_i, p_{succ} increases with N.

Example 2. Let us consider $p_i = 0.49999$. Then $p_{\text{succ}} = 0.501784$ for $N = 49999$ while $p_{\text{succ}} = 0.5025$ for $N = 99999$.

In order to obtain a as high as possible probability of success, we designed a second algorithm A.2 which uses *concatenated repetition codes*. The concatenation codes have been in introduced by Forney in 1966 [8] and generalized by Zinov'ev in 1976 [15]. The principle is to use two codes as depicted in Figure 2.

The combination of inner encoder, channel and outer decoder can be thought of as forming a new channel (called a *superchannel*). The aim is to improve the correcting capacity of the inner code by use of a second code. When transmitting over a very noisy channel, repetition codes are suitable outer codes in classical concatenated codes.

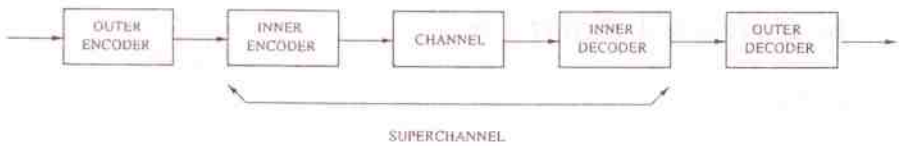

Figure 2: A Concatenated Code

In our cryptanalytic case, the superchannel is a BSC with parameter $p' = 1 - P_{\text{succ}}$ produced by the inner decoding residual error. We then iterate the decoding process on this superchannel with an outer repetition code. Here is the algorithm A.2 whose complexity is in $\mathcal{O}(N_1 \cdot N_2)$:

Input: *$N_1 \cdot N_2$ (N_1, N_2 odd) ciphertext blocks C_j encryted by key K from plaintext $P_j \in C_i (1 \leq j \leq N)$ and a probabilistic information $\mathcal{I}_K$ such that $\mathcal{I}_K \overset{p_i}{\cong} f(C_j)$ for some g and for all j.*

Output: *Exact value $\mathcal{I}(K)$ for the actual key.*

1. *Initialize counter* $ct1 \leftarrow 0$.
2. *For* $1 \leq k \leq N_1$
   (a) *Initialize counter* $ct2 \leftarrow 0$.
   (b) *For each of the N_2 ciphertext blocks C_j (k-th set)*
      i. *Compute* $f(C_i)$.
      ii. *If* $f(C_i) = 1$ *then* $ct2++$.
   (c) *end for*
   (d) *if* $ct2 \geq \frac{N_2+1}{2}$ *then* $\mathcal{I}(K) = 1$ *else* $\mathcal{I}(K) = 0$.
3. *If* $\mathcal{I}(K) = 1$ *then* $ct1++$.

4. *end for*

5. *If* $ct1 \geq \frac{N_1+1}{2}$ *then* $\mathcal{I}(K) = 1$ *else* $\mathcal{I}(K) = 0$.

While generally concatenated codes yield a better probability of success, it is not the case when the outer and inner codes are both repetition codes.

Proposition 2. Let N an odd number of ciphertext blocks. Algorithm A.2 has a higher probability of success than Algorithm A.1.

The proof is given in Appendix A. However the concatenated code approach allow us to compute a lower bound of A.1 success probability. The general Formula (1) cannot be computed directly as soon as N is too large.

3.3 Resistance Criterion against PDRC Attack

PDRC attack is possible if and only if there exists a subset $\mathcal{C}_i \subset \mathcal{C}$ such that $P_{\mathcal{C}_i}[\mathcal{I}_\mathcal{K}] \neq \frac{1}{2}$ for some property $\mathcal{I}$. This allow us to formulate the following resistance criterion against PDRC Attack.

Proposition 3 : Let S be a (m, n) block cipher and let us consider a property $\mathcal{I}$ about the key bits relatively to the ciphertext bits. S is immune against the PDRC attack relatively to property $\mathcal{I}$ if and only if $\forall j \in \mathcal{N}$ the partition $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_j)$ of $\mathcal{C}$ is such that

$$\forall k \leq j, \quad P_{\mathcal{C}_k}[\mathcal{I}] = \frac{1}{2}.$$

The cryptanalyst's work is to find a exploitable property $\mathcal{I}$ and a particular subset of "meaningful" plaintext blocks in order to conduct PDRC attack on S. On cryptographer's side things may be far more difficult. This difficulty is summarized with the four open problems here following.