LIKELIHOOD ESTIMATION FOR BLOCK CIPHER KEYS

Sean Murphy*, Fred Piper, Michael Walker(†), Peter Wild

Information Security Group
Royal Holloway
University of London
Egham
Surrey TW20 0EX, U.K.

(†)Vodafone Limited
The Courtyard
2-4 London Road
Newbury
Berkshire RG13 1JL, U.K.

May 31, 1995

Abstract

In this paper, we give a general framework for the analysis of block ciphers using the statistical technique of likelihood estimation. We show how various recent successful cryptanalyses of block ciphers can be regarded in this framework. By analysing the SAFER block cipher in this framework we expose a cryptographic weakness of that cipher.

Key Words. Statistical Inference, Likelihood Estimation, Block Ciphers, DES, SAFER, Cryptanalysis, Differential Cryptanalysis, Linear Cryptanalysis.

*This author acknowledges the support of the Nuffield Foundation.

1 Introduction

In this paper we set up a general framework for analysing block ciphers. In this framework the plaintext and ciphertext spaces are partitioned into a number of classes. We consider the probabilities of a plaintext in a given plaintext class being encrypted to a ciphertext in a given ciphertext class under different keys. For a judicious choice of partitions of plaintext and ciphertext spaces, these probabilities give a partition of the key space into key classes which allows the technique of likelihood estimation to be used to find the key class of the true key. We explain this idea more fully in Section 2 and show how it applies to iterated block ciphers in Section 3. In the rest of the paper we show how the recent cryptanalytic techniques of S-box pairs analysis, linear crytanalysis, differential cryptanalysis and linear structures fit naturally into this framework, so providing a comparison of these techniques. Finally, in Section 8, we see how this method can be used to show a cryptographic weakness of the SAFER [14] block cipher.

2 Statistical Estimation of the Key

We consider a block cipher to consist of a finite set of plaintexts M, a finite set of ciphertexts C and a finite set $\{f_k | k \in K\}$ of invertible functions from M onto C indexed by a set of keys K. The plaintext set M, the ciphertext set C, the key set K and the set of invertible functions $\{f_k : M \to C\}$ are all public. The ciphertext c corresponding to the plaintext m under a particular private key k^* is then

$$c = f_{k^*}(m).$$

The cryptanalyst's task is to find the particular key k^* given some information about a number of corresponding plaintext–ciphertext pairs. In particular, if the plaintexts are known, then we have a known plaintext attack, and if the plaintexts can be chosen by the cryptanalyst then we have a chosen plaintext attack.

We use the following framework to model the cryptanalyst's knowledge of the plaintexts and corresponding ciphertexts. Let $\phi : M \to X$ be a function from the plaintext set M onto a set X and $\psi : C \to Y$ a function from

the ciphertext set C onto a set Y. We call ϕ and ψ *partition functions*. The functions ϕ and ψ partition the plaintext and ciphertext spaces into equivalence classes indexed by the elements of X and Y respectively, and it is convenient to think of X and Y as sets of equivalence classes of M and C. Suppose that the cryptanalyst observes a pair (x, y), where $x = \phi(m)$ is the result of applying ϕ to some plaintext m and $y = \psi(c)$ is the result of applying ψ to the ciphertext $c = f_{k^*}(m)$ obtained by enciphering m under the true key k^*. Thus (x, y) is a pair of plaintext–ciphertext equivalence classes, which the cryptanalyst uses to estimate the true key k^*.

We show how the true key k^* can be estimated using the statistical technique of maximum likelihood estimation, as described in any textbook on statistical inference, for example Silvey [23]. Let $P_k[(x, y)]$ denote the probability that plaintext class x and ciphertext class y occur with key k. $P_k[(.,.)]$ defines a probability mass function on the set of possible plaintext and ciphertext classes $X \times Y$ parameterised by elements of the key set K. However we can regard $P_.[(x, y)]$ as a function on the key set K which is parameterised by elements of the set of possible plaintext and ciphertext classes $X \times Y$. This function is known as the *likelihood function* $L[(x, y); k]$ of the key k corresponding to the data (x, y), and it is a measure of the plausibility that k is the true key k^* after we have observed the data (x, y). Thus we have

$$\begin{gathered} L[(x, y); k] = P_k[(x, y)], \\ \text{and } \mathcal{L}[(x, y); k] = \log(L[(x, y); k]) = \log(P_k[(x, y)]), \end{gathered}$$

where $\mathcal{L}[(x, y); k]$ is the *log–likelihood function*. For any fixed k, we can think of the likelihood function $L[(x, y); k]$ as a random variable whose distribution is determined by the true distribution on the plaintext–ciphertext classes (given by the true key k^*). Thus we can define the expected value of the log–likelihood function at key k as

$$\theta(k) = \mathbf{E}\left\{\mathcal{L}[(x, y); k]\right\} = \mathbf{E}\left\{\log P_k\left[(x, y)\right]\right\}.$$

The following theorem, which we state in terms of plaintext and ciphertext classes, is a standard result for log–likelihood functions.

Theorem 1 [23]: For all keys $k \in K$, $\theta(k) \leq \theta(k^*)$ with equality if and only if $P_k[(x, y)] = P_{k^*}[(x, y)]$ for every plaintext–ciphertext class pair (x, y).

Theorem 1 states that $\theta(k)$ attains its maximum at k^*, and if the distributions corresponding to different keys on the plaintext–ciphertext classes are different, then k^* is unique. In any case, we can define an equivalence relation on the key space K in which two keys are equivalent if they have the same distribution on the plaintext–ciphertext classes, and then estimate the unique key class which maximises the likelihood. Let $\sigma : K \to Z$ be a *partition function* from the key space K onto a set Z such that $\sigma(k) = \sigma(k')$ if and only if $L[(x,y);k] = L[(x,y);k']$ for all (x,y). Z can be regarded as a set of key equivalence classes induced by the likelihood function. We can thus define the likelihood function $L[(x,y);z]$ of the key class z corresponding to the data (x,y). The expected value of the log–likelihood function at key class z is then given by

$$\theta(z) = \mathbf{E}\left\{\mathcal{L}[(x,y);z]\right\} = \mathbf{E}\left\{\log(P_z\left[(x,y)\right])\right\},$$

where $P_z[(x,y)] = P_k[(x,y)]$ for any $k \in K$ for which $\sigma(k) = z$. Theorem 1 shows that θ is *uniquely* maximised by $z^* = \sigma(k^*)$, the true key class.

Suppose now that we have N pairs of plaintext $\mathbf{m} = (m_1, \cdots, m_N)$ with corresponding ciphertexts $\mathbf{c} = (c_1, \cdots, c_N)$ that give plaintext classes $\mathbf{x} = (x_1, \cdots, x_N)$ and ciphertext classes $\mathbf{y} = (y_1, \cdots, y_N)$ respectively. The joint likelihood function is given by

$$L[(\mathbf{x},\mathbf{y});k] = \prod_{i=1}^{N} P_k[(x_i,y_i)],$$

so the joint log–likelihood function is given by

$$\mathcal{L}[(\mathbf{x},\mathbf{y});k] = \sum_{i=1}^{N} \mathcal{L}[(x_i,y_i);k] = \sum_{i=1}^{N} \log(P_k[(x_i,y_i)]).$$

We propose to estimate the true key k^* from the data $(\mathbf{x},\mathbf{y})$ by the method of maximum likelihood, so we have the following definition.

Definition: A *maximum likelihood estimate* (MLE) of the true key k^* is any $k \in K$ for which $L[(\mathbf{x},\mathbf{y});k]$ or equivalently $\mathcal{L}[(\mathbf{x},\mathbf{y});k]$ is maximal.

We can express $\theta(k)$ in terms of expected value of the joint log–likelihood function since

$$\mathbf{E}\left\{\frac{1}{N}\mathcal{L}[(\mathbf{x},\mathbf{y});k]\right\} = \mathbf{E}\left\{\log(P_k[(x_1,y_1)])\right\} = \theta(k).$$

Thus, from Theorem 1, the expected value of the joint log–likelihood function is maximised by the true key k^*. If we define the key partition function $\sigma : K \to Z$ as above, then we can define the joint likelihood function $L[(\mathbf{x}, \mathbf{y}); z]$ of the key class z corresponding to the data $(\mathbf{x}, \mathbf{y})$. The expected value of the joint log–likelihood function at key class z is given by

$$\mathbf{E}\left\{\frac{1}{N}\mathcal{L}[(\mathbf{x}, \mathbf{y}); z]\right\} = \mathbf{E}\left\{\log(P_z[(x_1, y_1)])\right\} = \theta(z),$$

and so is uniquely maximised by the true key class z^*.

We now give a brief description of the properties of the maximum likelihood estimate that make it the optimal estimate of the key.

Definition: Suppose $\{\hat{z}_n\}$ is a sequence of estimates for z^*. Then $\hat{z}_n$ is *consistent* if $\hat{z}_n \to z^*$ (in the appropriate stochastic sense).

Theorem 2 [23]: The maximum likelihood estimate of z^* is consistent.

Sketch Proof: We are essentially estimating $\theta(z)$ with $\frac{1}{N}\mathcal{L}[(\mathbf{x}, \mathbf{y}); z]$. The law of large numbers ensures that for large N and most $(\mathbf{x}, \mathbf{y})$, $\frac{1}{N}\mathcal{L}[(\mathbf{x}, \mathbf{y}); z]$ is near $\theta(z)$. If $\tilde{z}_N$ is the maximum likelihood estimate of z^* based on N plaintext–ciphertext classes, then this shows $\tilde{z}_N \to z^*$ (in the appropriate stochastic sense).□

Definition: A statistic t is *sufficient* for z^* if the distribution of a sample $(\mathbf{x}, \mathbf{y})$ given the value of $t((\mathbf{x}, \mathbf{y}))$ does not depend on z^*.

Equivalently, t is a sufficient statistic for z^* if the distribution within an equivalence class of the partition induced by t is independent of z^*. Thus the distribution of t contains all the information relevant to estimating z^*. A necessary and sufficient condition for t to be sufficient is given by the following factorisation theorem.

Theorem 3 [23]: t is a sufficient statistic for the family $\{P_z[(.,.)]|z \in Z\}$ if and only if $P_z[(x, y)]$ can be expressed as $P_z[(x, y)] = g_z[t((x, y))]h((x, y))$, where h does not depend on z.

Definition: A statistic t is *minimal–sufficient* for z^* if the partition induced by t contains every other sufficient partition.

Equivalently, t is a minimal–sufficient statistic for z^* if it is a function of every other sufficient statistic for z^*. Therefore a minimal–sufficient statistic contains the minimum information relevant to estimating z^*. The following theorem concerning the maximum likelihood estimator is a corollary of the above theorem.

Theorem 4 [23]: The maximum likelihood estimate is a function of a minimal–sufficient statistic.

Thus the maximum likelihood estimate depends only on the minimal relevant information in the sample. Hence the maximum likelihood estimate of the key is the optimal estimate of the key since it is both consistent and minimal–sufficient.

It is often convenient to express the likelihood in a different form. Suppose $P(x)$ denotes the probability that plaintext class x occurs, then we can write the likelihood function in the form

$$L[(x,y);k] = P_k[(x,y)] = P_k(y|x)P(x),$$

where $P_k(y|x)$ denotes the probability that a plaintext in class x is encrypted to a ciphertext in class y under key k. In any particular attack, the distribution of the plaintexts induces a distribution on the plaintext classes. Thus in a ciphertext–only attack where we had some information about the plaintexts, for example they were in ASCII or a natural language, we could use this information to calculate the values of $P(x)$. For many known plaintext attack, the distribution of plaintexts can be assumed to be uniform, and so if the plaintext classes all have equal size, then $P(x)$ is a constant and $L[(x,y);k] \propto P_k(y|x)$. For a chosen plaintext attack, $P(x)$ is essentially chosen by the cryptanalyst. In many attacks, the plaintexts are all chosen to be in one class, in which case we have a likelihood function $L[y;k] = P_k[y]$.

We saw above how the likelihood function defines an equivalence relation on the key space K in which two keys are equivalent if their likelihood functions are identical. Therefore to find the key by maximum likelihood estimation, we perform a "divide and conquer" cryptanalytic attack. By using

the plaintext and ciphertext classes, we first find the equivalence class of the key space which maximises the likelihood function. We then search this key class using the plaintexts and ciphertexts for the true key. The complexity of calculating the likelihood for a general pair of plaintext–ciphertext classes determines the computational complexity of finding the true key class, as the joint likelihood can easily be calculated from this. The number of plaintext–ciphertext pairs determines the error probability of estimating the true key class. The complexity of the cryptanalysis is then determined by the above factors and the sizes of the key classes.

For a simple example of a cryptanalytic attack consider a known plaintext exhaustive key search. We define ϕ and ψ both to be identity functions, so $\phi(m) = m$ and $\psi(c) = c$. Thus

$$L[(m,c);k] = P_k[(c|m)]P(m) = \begin{cases} P(m) & \text{if } f_k(m) = c \\ 0 & \text{otherwise,} \end{cases}$$

so the maximum value of the likelihood function distinguishes precisely those keys which transform the known plaintext m to the known ciphertext c.

In this example, it is necessary to perform an encryption or decryption for each key in order to evaluate the likelihood function, so that evaluating the likelihood requires $|K|$ encryptions. However, in various recent successful cryptanalytic attacks, defects in certain block ciphers have been discovered which allow ϕ and ψ to be chosen in such a way that it is possible to find the key class that maximises the likelihood function with an acceptable error probability and then search this key class in order to determine the key more quickly than would be expected in an exhaustive key search.

For certain block ciphers it is also possible to maximise or calculate the likelihood function very efficiently. Andelman and Reeds [1] [2] considered such a situation for both rotor machines (stream ciphers) and SP networks (block ciphers). For an SP network with a small number of rounds and an n-bit key $k = (k_1, \cdots, k_n)$, the likelihood function for k is well–approximated by a product of functions each involving only one key bit as

$$L[(m,c);k] \approx \prod_{i=1}^{n} l_i[(m,c);k_i].$$

Andelman and Reeds showed how to find the true key bits k_i by the fast maximisation of this product. For a general block cipher (including SP networks

with many rounds) such an approximation of the likelihood does not hold. Biham's *related key* chosen plaintext cryptanalysis [4] exploits key scheduling defects in certain ciphers. In the likelihood framework, this corresponds to sets of related keys having related likelihood functions given related plaintexts. The calculation of the likelihood function can thus be carried out more efficiently.

Maximum likelihood estimation of the true key class can also be thought of as a key partition inducing a partition on the plaintext–ciphertext classes, and this is the usual statistical formulation of such problems. Let $\sigma : K \to Z$ be a function from the key set K onto a set Z, so σ partitions K into equivalence classes indexed by elements of Z. Suppose we now wish to estimate the value of $z^* = \sigma(k^*)$, the equivalence class of k^* the true key, given some plaintext–ciphertext data $(\mathbf{m}, \mathbf{c})$. The maximum likelihood estimate of z^*, which is optimal in the sense described above, is necessarily a function of a minimal–sufficient statistic of the data $(\mathbf{m}, \mathbf{c})$. This minimal–sufficient statistic gives the minimal partition of the plaintext–ciphertext space$(M \times C)$ that allows the estimation of $\sigma(k^*)$. The minimal partition enables us to define optimal partition functions $\phi : M \to X$ and $\psi : C \to Y$.

We conclude this Section by discussing other ways of viewing likelihood estimation for block cipher keys. The problem of estimating the true key k^* has a Bayesian formulation. If the cryptanalyst has a prior distribution on the set of keys, then Bayes' Theorem provides a method for estimating the key. If we have a plaintext class x and a ciphertext class y, then we can use an inference form of Bayes' Theorem [23]:

Bayes' Theorem: Suppose $p(.)$ is the prior mass function on the elements of the key space K, $p(.|(x, y))$ is the posterior mass function on the elements of the key space K given the data (x, y), and L[(x,y);k] is the likelihood of the key k corresponding to the data (x, y), then

$$p(k|(x, y)) \propto L[(x, y); k]p(k).$$

Usually each key is equally likely, so the prior mass function $p(.)$ is constant. In this case Theorem 1 shows that the posterior probability $p(k^*|(x, y))$ is maximal, or equivalently k^* is a mode of the posterior distribution. This

is a standard result of Bayesian statistics, but whatever the prior distribution, the mode of the posterior is the Bayes' estimate with respect to the zero-one loss function. Note that if we have many corresponding pairs of plaintext and ciphertext classes, the posterior distribution obtained by using a joint likelihood for many pairs is the same as that obtained by iteratively calculating posterior distributions one pair at a time and using them as prior distributions for the next iteration. The Bayesian approach has often been used in the analysis of block and stream ciphers both explicitly, for example [18] [3] [8], and implicitly in "key ranking" techniques, for example [13] [16].

The problem of testing whether a key class is the true key class can also be expressed in terms of hypothesis testing, another routine statistical technique [23]. Hypothesis testing has been used by Brynielsson [6] to find stream cipher keys. Suppose we wish to test the null hypothesis $H_0 : z = z^*$ that z is the true key class z^* against the alternative hypothesis $H_A : z \neq z^*$ that z is not the true key class. The *likelihood ratio test* is based on the ratio

$$\lambda\left((\mathbf{x},\mathbf{y})\right) = \sup_{z \neq z^*} \left\{ \frac{L[(\mathbf{x},\mathbf{y});z]}{L[(\mathbf{x},\mathbf{y});z^*]} \right\}.$$

Intuitively, we would reject the null hypothesis H_0 if the ratio $\lambda\left((\mathbf{x},\mathbf{y})\right)$ is too large. Indeed the likelihood ratio test has a critical region for rejecting H_0 of the form $\{(\mathbf{x},\mathbf{y}) | \lambda((\mathbf{x},\mathbf{y})) > c\}$, where c is fixed by the error probability $\alpha = \mathbf{P}\left[\lambda((\mathbf{x},\mathbf{y})) > c | z = z^*\right]$ of rejecting the true key class z^*. Such a test is called a test of *size* α. A full analysis of this likelihood test would involve the evaluation of the *power function* $\pi : H_A \rightarrow [0,1]$, which is defined by

$$\pi(z) = 1 - \mathbf{P}\left[\text{ Accept } z \text{ is true key class} | z \in H_A\right],$$

so $\pi(z)$ is the probability of rejecting the false key class z. We should like to choose a test of size α that is optimal in some sense over all tests of size α. In the case, when H_A consists of a single key class, that is H_A is a simple hypothesis, the Neyman-Pearson Lemma [23] shows that a test based on a critical region $\{(\mathbf{x},\mathbf{y}) | \lambda(\mathbf{x},\mathbf{y}) > c\}$, where c is determined by $\alpha = \mathbf{P}\left(\lambda(\mathbf{x},\mathbf{y}) > c\right)$, is the *most powerful* test of size α. Even when H_A is a composite hypothesis, for most cryptographic purposes testing H_0 against H_A is equivalent to a one–sided test involving normal distributions with different means. For such tests, the likelihood ratio test is *uniformly most powerful*, that is it has maximal power $\pi(z)$ over all tests of size α for every $z \in H_A$.

Note that we may be able to perform a hypothesis test more efficiently by performing a *sequential likelihood ratio test* [23].

3 Iterated Block Ciphers

Most block ciphers are constructed by iterating a relatively simple cryptographic function a number of times. Consider such an n-round block cipher. For such a cipher $M = C$ and the i^{th} round transformation under a key k is an invertible function from the $(i-1)^{th}$-round message space to the i^{th}-round message space, which we denote by $\alpha_i(k) : M \to M$. Thus for a plaintext m_0 and key k, we can regard the encryption process as producing a sequence of message states $m_0, m_1, \cdots, m_n$, where m_n is the ciphertext, and $m_i = m_{i-1}\alpha_i(k)$. Thus the encryption function $\alpha(k) : M \to M$ from plaintext to ciphertext is given by

$$\alpha(k) = \prod_{i=1}^{n} \alpha_i(k).$$

For such an iterated block cipher, we generalise the approach outlined in the previous Section, by partitioning all the i^{th}-round message spaces ($i = 1, \cdots, n$) as well as the plaintext and ciphertext spaces. For each $i = 0, \cdots n$, let $\phi_i : M \to Y_i$ be a function from the message space M onto a set Y_i. The *partition function* ϕ_i partitions the i^{th}-round message space into equivalence classes indexed by the elements of Y_i ($i = 0, \cdots, n$).

Suppose now that we have a sequence of message states $m_0, m_1, \cdots m_n$ obtained by encrypting under the key k. We can then obtain a sequence of message classes $y_0, y_1, \cdots, y_n$, where $y_i = \phi_i(m_i)$. This sequence can be regarded as a realisation of a random process on $Y_0, Y_1, \cdots, Y_n$. Suppose the matrix of transition probabilities from Y_{i-1} to Y_i under key k is $Q_i(k)$. The entries of the transition matrix $Q_i(k)$ can be easily calculated as

$$(Q_i(k))_{y_{i-1}, y_i} = P_k\left(Y_i = y_i | Y_{i-1} = y_{i-1}\right) = \frac{|\{m \in y_{i-1} | m\alpha_i(k) \in y_i\}|}{|y_{i-1}|}.$$

We wish to calculate the matrix $Q(k)$ of transition probabilities from the plaintext classes Y_0 to the ciphertext classes Y_n. If the random process on

the message classes $Y_0, \cdots, Y_n$ under key k is well–approximated by a first order Markov process, this matrix $Q(k)$ is given by the product

$$Q(k) = \prod_{i=1}^{n} Q_i(k).$$

A first order Markov random process is one in which the distribution of the i^{th} message class y_i conditional on all the previous message classes $y_{i-1}, \cdots, y_0$ is identical to the distribution of y_i conditional on the most recent message class y_{i-1}. As part of the modelling approach, we make the assumption that the random process on the message classes $Y_0, \cdots, Y_n$ under key k is well–approximated by a first order Markov process. For the S-box pairs analysis considered in Section 4, this is shown to be a valid assumption [8], whereas it has been implicitly assumed in the other cryptanalyses we consider. This assumption could be justified empirically in the different cryptanalyses we consider, but we give the following heuristic justification. The round partition functions used in the various cryptanalyses are algebraic homomorphisms, so the set of round message classes form an algebraic quotient space. The round functions contain transformations which are highly non–homomorphic with respect to these quotient spaces. Therefore the algebraic structure of an $(i-1)^{th}$ round message class y_{i-1} is not "inherited" by an i^{th} message class y_i, that is the elements of y_{i-1} are sufficiently "randomised" within y_i. In cases where the round functions are partially homomorphic with respect to these quotient spaces, for example the differential analysis of DES discussed in Section 6, the"randomisation" property can often be restored by using the partial homomorphism to construct a finer partition.

We can thus calculate the likelihood function as above with $\phi = \phi_0$ and $\psi = \phi_n$ since the relevant entry of $Q(k)$ gives the probability $P_k(y|x)$. For such ciphers, the equivalence classes of the key space K induced by the likelihood function are indexed by the set of matrices $\{Q(k)|k \in K\}$, that is k and k' belong to the key class if $Q(k) = Q(k')$.

4 S-Box Pairs Cryptanalysis of DES

For a general example of a cryptanalysis that can be thought of in the above framework, consider the cryptanalysis of DES [20] by Davies and Murphy

[8] using pairs of S-boxes. Consider a pair of adjacent DES S-boxes with respective 6-bit inputs $a_1, \cdots, a_6$ and $b_1, \cdots, b_6$. Over all 2^{12} possible inputs to this pair of DES S-boxes, each possible 8-bit output occurs equally often (16 times). However, the expansion phase of DES (E) means that the values of $a_5 \oplus b_1$ and $a_6 \oplus b_2$ are determined solely by the key. If we fix these bits, then over the remaining possible 2^{10} inputs, we obtain a non-uniform distribution of the 2^8 possible outputs. Furthermore this non-uniform distribution has one of two forms depending on the value of $a_5 \oplus a_6 \oplus b_1 \oplus b_2$, the XOR of the middle four bits. To calculate the distribution of the XOR of the outputs of n pairs of S-boxes given random inputs we have to calculate the n-fold convolution of this distribution with itself, and this distribution is also one of only two non-uniform distributions depending on the total XOR of $a_5 \oplus a_6 \oplus b_1 \oplus b_2$ from the input of each pair of S-boxes. From a plaintext–ciphertext pair it is easy to obtain a realisation of the XOR of the outputs of 8 pairs of S-boxes with approximately random inputs. With many such plaintext–ciphertext pairs, we can construct an estimate of the XOR of all the common key bits. We can do this for all 8 pairs of S-boxes on both the left and right halves of the cipher, so this potentially gives us 16 bits of key information. However only S-box pair 7,8 gives us this information faster than an exhaustive key search, so this attack only yields two bits of key information. We show how this cryptanalysis can be viewed in the likelihood framework. For the purposes of this cryptanalysis of DES, it is convenient to ignore the initial and final permutations of DES, IP and IP^{-1}, as there is no loss of generality. We actually analyse a variant of the DES cryptosystem given by Desmedt [9], which we call EXPDES. This variant is a block cipher which can be regarded as DES with 48-bit registers rather than 32-bit registers. If E, S and P denote the usual expansion, S-box and permutation of the encryption function f of DES, then we can define $T : \mathbb{Z}_2^{48} \to \mathbb{Z}_2^{48}$ by $T = S.P.E$ (with maps on the right). EXPDES has message space $M = \mathbb{Z}_2^{96}$, and we denote the i^{th}-round message by (l_i, r_i) for $l_i, r_i \in \mathbb{Z}_2^{48}$. An encryption round of EXPDES is given by

$$\begin{aligned} (l_{2i+1}, r_{2i+1}) &= (l_{2i} \oplus (r_{2i} \oplus k_{2i+1})T, r_{2i}) \qquad [i = 0, \cdots, 7] \\ (l_{2i}, r_{2i}) &= (l_{2i-1}, r_{2i-1} \oplus (l_{2i-1} \oplus k_{2i})T) \qquad [i = 1, \cdots, 8], \end{aligned}$$

where k_i is the usual 48-bit i^{th}-round DES subkey. EXPDES encrypts plaintext (l_0, r_0) to (l_{16}, r_{16}), and by analogy with DES, the ciphertext is (r_{16}, l_{16}). Note that if we encrypt a 64-bit message by first expanding both 32-bit halves

to 48-bits using the expansion function E, then encrypting with EXPDES and finally "unexpanding" the 96-bits to a 64-bit cryptogram using E, this is equivalent to encrypting the original 64-bit message using DES.

At any particular stage, EXPDES is divided into a left and a right register, and we regard each of them as a 48-dimensional vector space over $\mathbb{Z}_2$. Let π denote the projection of this space onto an 8-dimensional subspace corresponding to the 8-bit output of the adjacent pair of S-boxes after the functions P and E have been applied. Let $I_1, \cdots, I_6$ and $J_1, \cdots, J_6$ denote the unit vectors which have a 1 in the appropriate position corresponding to the 6-bit inputs to a pair of adjacent S-boxes, and let ρ denote the projection of the 48-dimensional space onto the 1-dimensional subspace generated by $I_5 \oplus I_6 \oplus J_1 \oplus J_2$. Note that the linear transformation $E\rho$ from a 32-dimensional to a 1-dimensional space satisfies $E\rho = 0$, and so the non-linear transformation $T\rho$ of the 48-dimensional space satisfies $T\rho = SPE\rho = 0$. We are going to use the projection π to track the 8-bit XOR of the output of the adjacent pair of S-boxes and ρ to track the fixed value $a_5 \oplus a_6 \oplus b_1 \oplus b_2$ given above. Let Y_i be the 9-dimensional vector space which is the direct sum $Im(\rho) \oplus Im(\pi)$ and define partition functions $\phi_i : X_i \to Y_i$ by

$$\phi_i\left((l_i, r_i)\right) = (l_i\rho, r_i\pi), \qquad [i = 0, \cdots, 16].$$

The encryption transformation from the $(2i)^{th}$-round to the $(2i+1)^{th}$-round leaves the message classes unaltered since,

$$\begin{aligned} \phi_{2i+1}\left((l_{2i+1}, r_{2i+1})\right) &= (l_{2i+1}\rho, r_{2i+1}\pi) \\ &= (l_{2i}\rho \oplus (k_{2i+1} \oplus r_{2i})T\rho, r_{2i}\pi) \\ &= (l_{2i}\rho, r_{2i}\pi) = \phi_{2i}\left((l_{2i}, r_{2i})\right) \quad [i = 0, \cdots, 7], \end{aligned}$$

which means that the transition matrix is the identity matrix.

The encryption transformation from the $(2i-1)^{th}$-round to the $(2i)^{th}$-round permutes the message classes according to the output of the two adjacent S-boxes. We have

$$\begin{aligned} \phi_{2i}\left((l_{2i}, r_{2i})\right) &= (l_{2i}\rho, r_{2i}\pi) \\ &= (l_{2i-1}\rho, r_{2i-1}\pi \oplus (k_{2i} \oplus l_{2i-1})T\pi) \\ &= \phi_{2i-1}\left(l_{2i-1}, r_{2i-1}\right) \oplus (0, (k_{2i} \oplus l_{2i-1})T\pi) \quad [i = 1, \cdots, 8] \end{aligned}$$

giving a block diagonal transition matrix of the form

$$\begin{pmatrix} M_0(k_{2i}) & 0 \\ 0 & M_1(k_{2i}) \end{pmatrix},$$

where the subscript denotes the value of $l_{2i-1}\rho$. $T\pi$ is the output of the pair of adjacent S-boxes, and given that the value of $l_{2i-1}\rho$ is known, the distribution of $T\pi$ depends on $k_{2i}\rho$. It can be shown [8] (as each DES S-box consists of four permutations) that the $(2^8 \times 2^8)$ matrices $M_{\cdot}(k_{2i})$ can be expressed as

$$M_0(k_{2i}) = E + (-1)^{k_{2i}\rho}W, \qquad M_1(k_{2i}) = E - (-1)^{k_{2i}\rho}W,$$

where E denotes the $2^8 \times 2^8$ matrix with entries 2^{-8} and $W = (w_{ij})$ has row and column sum zero, "constant diagonals" (ie. $w_{(i,j)\oplus(d,d)} = w_{(i,j)}$ for all $d \in \mathbb{Z}_2^8$), and is easily calculated. Thus we can write

$$M(k_{2i}\rho) = \begin{pmatrix} E + (-1)^{k_{2i}\rho}W & 0 \\ 0 & E - (-1)^{k_{2i}\rho}W \end{pmatrix}$$

as the transition matrix from the $(2i-1)^{th}$-round to the $(2i)^{th}$-round message classes.

The random process on round message classes is well–approximated by a Markov process [8], so the overall transition matrix is well-approximated by the product of the round transition matrices. Now $EW = WE = 0$, so for any $n_1, \cdots, n_j = 0, 1$ we have

$$\prod_{i=1}^{j} M(n_i) = \begin{pmatrix} E + (-1)^{\oplus n_i}W^l & 0 \\ 0 & E + (-1)^{\oplus n_i}(-W)^l \end{pmatrix}.$$

Thus we can write the transition matrix from plaintext classes to ciphertext classes as

$$Q(k) = \prod_{i=1}^{8} M(k_{2i}\rho) = M\left(\oplus_{i=1}^{8}(k_{2i}\rho)\right) = M\left((\oplus_{i=1}^{8}k_{2i})\rho\right)$$

$$\text{so } Q(k) = \begin{pmatrix} E + (-1)^{s}W^8 & 0 \\ 0 & E + (-1)^{s}W^8 \end{pmatrix}, \text{ where } s = (\oplus_{i=1}^{8}k_{2i})\rho$$

Our analysis of DES has used an expanded form of DES, EXPDES. However, the generalisation of DES to EXPDES means that we use plaintext–ciphertext classes corresponding to the "top left" block of $Q(k)$, giving a transition matrix between plaintext and ciphertext classes of $E + (-1)^s W^8$, where $s = (\oplus_{i=1}^{8}k_{2i})\rho$. We can thus define two classes on the key space, depending on the value of $(\oplus_{i=1}^{8}k_{2i})\rho$, and, since we need only to discriminate

between two cases, we can use a likelihood ratio hypothesis test for two simple hypothesis as described in Section 2. The two approaches are equivalent since this hypothesis test "selects" the value of $(\oplus_{i=1}^{8} k_{2i})\rho$ giving the higher value of the likelihood function. We can also define a function

$$\phi : Y_0 \times Y_{16} \to Y = Im(\pi)$$

from the plaintext–ciphertext classes onto a set Y by addition of the plaintext and ciphertext classes, so

$$\phi\left((y_0, y_{16})\right) = y_0 \oplus y_{16} = (r_0 \oplus r_{16})\pi.$$

This defines a partition of the plaintext–ciphertext classes into equivalence classes on which the likelihood function is constant for each key, since $E + (-1)^s W^8$ has constant diagonals. This partition corresponds to the minimal–sufficient statistic for estimating $s = (\oplus_{i=1}^{8} k_{2i})\rho$.

5 Linear Cryptanalysis

In this Section, we show how the technique of linear cryptanalysis applied by Matsui and Yamagishi to FEAL [17], and by Matsui to DES [15] [16], can be viewed in the likelihood framework. Linear cryptanalysis is very similar to the analysis of S-box pairs analysed in the previous Section. Linear cryptanalysis essentially considers projections onto linear (1-dimensional) subspaces, rather than the projection onto the 9-dimensional subspace used in the previous Section. In our analysis, we consider the analysis of DES given by Matsui [15], and for consistency we use the notation of that paper. We regard the left and right registers as a 32-dimensional binary vector space $\mathbb{Z}_2^{32}$, and for $X \in \mathbb{Z}_2^{32}$ we define $X[i]$ to be the projection onto the i^{th} position of X and $X[i_1, \cdots, i_j] = X[i_1] \oplus \cdots \oplus X[i_j]$.

In the annex to Matsui [15] the following equation is given

$$P_H[7, 18, 24] \oplus P_L[12, 16] \oplus C_H[15] \oplus C_L[7, 18, 24, 29, 27, 28, 30, 31] = k$$

where P_H, P_L, C_H, C_L are the plaintext left and right halves and the ciphertext left and right halves respectively, and k is the sum of a certain collection of key bits. This equation holds with probability $\frac{1}{2} - p$, where $p = 1.49 \times 2^{-24}$.

In the notation of Section 2, if we define the partition function ϕ on the plaintext by $\phi((P_H, P_L)) = P_H[7, 18, 24] \oplus P_L[12, 16]$, and the partition function ψ on the ciphertext by $\psi((C_H, C_L)) = C_H[15] \oplus C_L[7, 18, 24, 29, 27, 28, 30, 31]$ to be the ciphertext class, then the transition matrix between plaintext and ciphertext classes is given by

$$\begin{pmatrix} \frac{1}{2} - (-1)^k p & \frac{1}{2} + (-1)^k p \\ \frac{1}{2} + (-1)^k p & \frac{1}{2} - (-1)^k p \end{pmatrix}.$$

This transition matrix can be calculated by using i^{th}-round message classes as explained in Section 3, and this is essentially how the probabilistic equation given above was derived. We need to test whether $k = 0$ or $k = 1$. The log–likelihood ratio statistic, Λ, for this test is given by

$$\Lambda = (N - 2n)4p \qquad \text{for small } p.$$

This is the statistic used by Matsui. Therefore we choose $k = 0$ if $n < \frac{N}{2}$ and $k = 1$ otherwise. This gives, for example, for a one–sided test of size 0.0228, a sample size N of 2^{47} plaintext–ciphertexts pairs, the value given by [15].

Matsui [15] also gives another probabilistic equation, namely

$$P_H[7, 18, 24] \oplus P_L[12, 16] \oplus C_H[15] \oplus C_L[7, 18, 24, 29] \oplus F_{16}(C_L, k_{16})[15] = k,$$

where k is a certain sum of key bits and $F_{16}(C_L, k_{16})[15]$ denotes the output of the DES F function in bit position 15. This equation holds with probability $\frac{1}{2} + p$, where $p = 1.19 \times 2^{-22}$. Since this probabilistic equation depends on k and the 6 bits of K_{16} used as input to S-box 1, we now have the potential to estimate 7 bits of key information. These are k^*, the true value of k, and h^*, the 6 true key input bits of K_{16} to S-box 1 which affect the value of $F_{16}(C_l, K_{16})[15]$. This probabilistic equation can be derived using i^{th}-round message classes as described in Section 3, with plaintext and ciphertext classes defined by $\phi : \mathbb{Z}_2^{64} \to \mathbb{Z}_2$ and $\psi : \mathbb{Z}_2^{64} \to \mathbb{Z}_2^7$ respectively, where

$$\begin{aligned} &\phi\left((P_H, P_L)\right) = (P_H[7, 18, 24] \oplus P_L[12, 16]), \\ &\psi\left((C_H, C_L)\right) = (C_H[15] \oplus C_L[7, 18, 24, 29], C_L^1), \end{aligned}$$

and C_L^1 denotes the 6 input bits from C_L to S-box 1. We can now estimate the true key class (k^*, h^*) by likelihood estimation. Matsui's [15] estimate of

the true key class is the key class (k, h) which has the largest count for the number of times the above equation holds.

In [16] Matsui improves the efficiency of this cryptanalysis by using the technique of *key ranking*. In this technique, the key classes are ranked according to their count for the number of times the equation holds. The key classes are then tested in rank order. As mentioned in Section 2 this technique is the calculation of an empirical Bayesian posterior distribution on the key classes.

Kaliski and Robshaw [12] describe a *multiple approximations* technique which uses n linear approximations simultaneously. This corresponds to defining the partition functions to be projections onto certain n-dimensional subspaces rather than 1-dimensional ones.

6 Differential Cryptanalysis

In this Section, we show how differential cryptanalysis first described by Biham and Shamir [5] and generalised by Lai, Massey and Murphy [13] can be expressed in the likelihood framework. Suppose we have a block cipher with message space M, and i^{th}-round subkey $k_i \in M$ in which m_0 is the plaintext and m_n is the ciphertext, and the i^{th}-round encryption is given by

$$m_i = (m_{i-1} \odot k_i)T \qquad [i = 1 \cdots n],$$

for some invertible function T on the message space M, where $(M, \odot)$ is a group. We note that, whilst many iterated block ciphers can be described in this form, DES cannot because of the expansion phase (E). However the expanded form of DES described in Section 4, EXPDES, does fit this formulation, so results for DES can be obtained by analysing EXPDES.

We can define another (double) block cipher based on the original cipher with message space $A = M \times M$ and key space K by running a pair of encryptions in parallel. For this cipher the i^{th}-round encryption is given by

$$(m_i, m_i') = \left((m_{i-1} \odot k_i)T, (m_{i-1}' \odot k_i)T\right).$$

Thus we can define a group $(A, \circ) = (M, \odot) \times (M, \odot)$, an i^{th}-round subkey $k_i = (k_i, k_i)$ and an invertible function S on A such that for $x_{i-1}, x_i \in A$ an

encryption round can be defined by

$$x_i = (x_{i-1} \circ k_i)S.$$

An i^{th}-round differential can be regarded as a surjective function $\psi_i : A = M \times M \to M$ on the i^{th}-round message classes of the double cipher defined by

$$\psi_i\left((m_i, m_i')\right) = m_i \odot m_i'^{-1}.$$

Suppose now that ψ_i and ψ_{i+1} were used as partition functions in a likelihood cryptanalysis. We first note that for any constant $d \in M$,

$$\psi_i\left((m_i \odot d, m_i' \odot d)\right) = (m_i \odot d) \odot (m_i' \odot d)^{-1} = m_i \odot m_i'^{-1} = \psi_i\left((m_i, m_i')\right).$$

Thus, in particular, the introduction of a subkey $k_i = (k_i, k_i)$ to this double cipher on A leaves the value of ψ_i unaltered. Hence the transition probabilities between the i^{th}-round and the $(i+1)^{th}$-round message classes defined by ψ_i and ψ_{i+1} are key–independent. Biham and Shamir [5] term such differential classes *characteristics*. For many ciphers, though not EXPDES, it is a reasonable assumption (as discussed in Section 3) that the random process on the round differential classes forms a first order Markov processes. Thus we can calculate key–independent transition probabilities between differential characteristics several rounds apart.

If we defined all the partition functions ϕ_i $(i = 0, \cdots, n)$ to be the differential functions ψ_i, we would obtain a key–independent transition matrix between the plaintext classes and the ciphertext classes. Whilst this would give us information about the correlation between plaintext and ciphertext differentials, it would give no information about the key. Suppose we define partition functions ϕ_i by

$$\begin{array}{rll} \phi_i = \psi_i & (Y_i = M) & i = 0, \cdots, (n-l-1), \\ \phi_{n-l} = \text{Indicator for } D & (Y_{n-l} = \{D, M \setminus D\}), & \\ \phi_i = \text{Identity} & (Y_i = A) & i = (n-l+1), \cdots n, \end{array}$$

where $D \subset M$ and $\phi_{n-l}((m, m')) = D$ if and only if $m_i \odot m_i'^{-1} \in D$. With these partition functions, we obtain a key–independent transition matrix from the plaintext classes to the $(n-l)^{th}$-round message classes. This enables us to find a transition matrix depending on a small set of key classes for the

transition from the $(n-l)^{th}$-round message classes to the ciphertext. We can use the likelihood estimation process outlined earlier to find the true key class, and then search this class for the true key. In practice, for a differential attack we would use a chosen plaintext attack by specifying a given plaintext difference. This gives key–independent probabilities p_D and $1-p_D$ for the $(n-l)^{th}$-round message classes $D, M \setminus D$. The plaintext difference and the set D would be chosen so as to maximise the transition probability p_D. This makes the likelihood estimation process more efficient, and may reduce the number of key classes. In this case, if we observe the ciphertext $y \in A$, we can write the likelihood and the log–likelihood function as

$$L[y;k] = P_k(y), \qquad \mathcal{L}[y;k] = \log\left(P_k(y)\right).$$

If we observe N ciphertexts $\mathbf{y} = (y_1, \cdots y_N)$, then the joint log–likelihood is given by

$$\mathcal{L}[\mathbf{y};k] = \sum_{i=1}^{N} \log P_k(y_i).$$

We now show how some of the methods used in the recent differential attacks may be regarded in the likelihood framework.

In order to simplify the following discussion, we assume that $l = 1$. This corresponds to what Biham and Shamir call a 1R-attack. For a given plaintext equivalence class (differential), the $(n-1)^{th}$-round message class D occurs with probability p_D, and the other $(n-1)^{th}$-round message class $M \setminus D$ occurs with probability $1-p_D$. For $y \in A$, we can define

$$u(k_n) = P_{k^*}[\, yS^{-1} \circ k_n^{-1} \in D\,],$$

which can be expressed in terms of the following two probabilities,

$$q(k_n) = \mathbf{P}[\, y \in D \,|\, y \circ k_n \in D\,], \text{ and } r(k_n) = \mathbf{P}[\, y \notin D \,|\, y \circ k_n \notin D\,].$$

If we let $w = yS^{-1} \circ k_n^{*-1}$ denote the true value of the $(n-1)^{th}$-round message, we have

$$\begin{aligned} u(k_n) &= P_{k^*}(yS^{-1} \circ k_n^{-1} \in D) \\ &= P_{k^*}(w \circ (k_n^* \circ k_n^{-1}) \in D) \\ &= \mathbf{P}(w \circ (k_n^* \circ k_n^{-1}) \in D | w \in D) P_{k^*}(w \in D) \\ &\quad + \mathbf{P}(w \circ (k_n^* \circ k_n^{-1}) \in D | w \notin D) P_{k^*}(w \notin D) \\ &= p_D q(k_n^* \circ k_n^{-1}) + (1-p_D)(1 - r(k_n^* \circ k_n^{-1})). \end{aligned}$$

Thus we have

$$\begin{aligned}
\mathbf{E}[P_k(y_i)] &= \mathbf{E}[P_k(y_i \mid y_i S^{-1} \circ k_n^{-1} \in D\,)]P_{k^*}(y_i S^{-1} \circ k_n^{-1} \in D) \\
&\quad +\mathbf{E}[P_k(y_i | y_i S^{-1} \circ k_n^{-1} \notin D)]P_{k^*}(y_i S^{-1} \circ k_n^{-1} \notin D) \\
&= \begin{cases} \frac{u(k_n)}{|D|} & \text{if } y_i S^{-1} \circ k_n^{-1} \in D \\ \frac{1-u(k_n)}{|M|-|D|} & \text{if } y_i S^{-1} \circ k_n^{-1} \notin D \end{cases},
\end{aligned}$$

and so the expected value of the joint log–likelihood for N ciphertexts $\mathbf{y} = (y_1, \cdots, y_N)$ is given by

$$\begin{aligned}
\mathbf{E}[\mathcal{L}[\mathbf{y}; k]] &= \textstyle\sum_{i=1}^{N} \mathbf{E}[\log P_k(y_i)] \\
&= \mathbf{E}[m_D(k_n)] \log \tfrac{u(k_n)}{|D|} + (N - \mathbf{E}[m_D(k_n)]) \log \tfrac{1-u(k_n)}{|M|-|D|} \\
&= N \log \tfrac{1-uk_n}{|M|-|D|} + \mathbf{E}[m_D(k_n)] \log \left(\tfrac{u(k_n)}{(1-u(k_n))} \tfrac{(|M|-|D|)}{|D|} \right),
\end{aligned}$$

where $m_D(k_n)$ is the number of times the $(n-1)^{th}$-round message class occurs by decrypting the ciphertexts under n^{th}-round subkey k_n.

For differential cryptanalyses $|D| << |M|$, $u(k_n^*) = p_D$, and we can usually make the simplifying assumption that for $k_n \neq k_n^*$, $u(k_n) \approx \frac{|D|}{|M|}$. Thus, for $k \neq k^*$,

$$\mathbf{E}[\mathcal{L}[\mathbf{y}; k]] \approx -N \log |M|,$$

whereas for the true key k^*,

$$\mathbf{E}[\mathcal{L}[\mathbf{y}; k^*]] \approx -N \log |M| + \mathbf{E}[m_D(k^*)] \log \left(\frac{p_D}{(1-p_D)} \frac{|M|}{|D|} \right).$$

However $\mathbf{E}[m_D(k)]$ is maximal when $k = k^*$, so maximising $m_D(k)$ gives an approximate maximum likelihood estimate.

Gilbert and Chassé's [11] differential analysis of FEAL-8 is one in which the true key is found by calculating which key gives rise to a distribution that best fits a theoretical distribution. We note that this attack is obtained if we take $\phi_{n-l} = \psi_{n-l}$ in the above model.

Biham and Shamir [5] use $m_D(k_n)$ to estimate the true key class. Their estimate of the sample size is based on a *signal to noise ratio* (S/N). In likelihood terms, this is the ratio of the likelihood under the true key class to the average value of the likelihood. In the theory of statistical hypothesis tests (outlined in Section 2), we consider the ratio of the likelihood under

the true key class to the largest likelihood under all the other key classes. Thus the signal to noise ratio (S/N) does not allow for correlated counts for incorrect key classes and so is an approximation which may underestimate the sample size in situations where counts for different keys are highly correlated.

As we mentioned above, the random process on the differential classes for EXPDES is not approximated by a first order Markov process. This is because the invertible function S on the "double" cipher with message space A is partially homomorphic (due to the DES S-boxes mapping 48 bits to 32) with respect to the differential classes. By using S to construct a finer partition by projecting onto certain bit positions as well as the difference, we obtain an approximate first order process. Biham and Shamir [5] termed such partition functions for DES *enhanced characteristics*. The transition probabilities between enhanced characteristics are therefore key–dependent, but the analysis of enhanced characteristics is similar to that described above. The probabilities used for ordinary characteristics for DES are averages across keys of the probabilities of the enhanced characteristics.

7 Linear Structures

In a series of papers, Reeds and Manferdelli [21], Chaum and Evertse [7], and Evertse [10] consider how to analyse a block cipher if the encryption function has some partial linearity associated with it. The basic idea is to find linear or affine transformations of the plaintext, ciphertext and key spaces respectively such that the mapped ciphertext is a function of the mapped plaintext and mapped key. If these transformations have less than full rank, then it may be possible to cryptanalyse the smaller mapped cryptosystem and determine a subspace of the key space containing the true key.

If a block cipher has this partial linearity, then we can define partition functions based on the various linear or affine transformations and obtain a set of key–dependent permutation matrices describing the transitions between the plaintext and ciphertext classes. In terms of the likelihood framework, these attacks can be regarded as the special case when the transition matrix is a permutation matrix. In this case the likelihood function only takes the values zero or one and the algebraic structure may allow us to evaluate the likelihood function quickly.

The likelihood framework however allows us to relax the condition that the transition matrices must be permutation matrices, and still enables us to find the key. Thus for a general block cipher, when the partition functions have been defined in terms of linear or affine transformations, we are really considering *probabilistic linear structures*. This is certainly the case with the three examples, S-box pairs, linear and differential cryptanalysis, that are considered above. In the analysis of SAFER in the next Section, the partition functions are based on a module homomorphism.

8 Analysis of the SAFER Block Cipher

In this Section, we consider the SAFER block cipher [14], and show that viewing SAFER in the likelihood framework exposes a cryptographic weakness. This cryptographic weakness is described more fully in [19].

SAFER (K-64) is a block cipher that operates on 8-byte (64-bit) blocks under the control of an 8-byte (64-bit) key. It is a byte–oriented cipher in that all of its basic operations are on bytes or pairs of bytes. Thus we consider the message space to be a $\mathbb{Z}_{256}$-module of rank 8, V say. SAFER is an r-round cipher that uses $2r+1$ 8-byte (64-bit) subkeys $K_1 \cdots K_{2r+1}$. The key scheduling is such that the j^{th} byte of any subkey depends only on the j^{th} byte of the key. Thus the subkey bytes are independent. The i^{th} round function splits naturally into two parts:

1. *Keyed Byte–Separated Layer:* Each of the 8 bytes is processed separately from the other bytes. A byte is obtained from the j^{th} byte using the j^{th} bytes from subkeys K_{2i-1} and K_{2i}, 8-bit XOR, addition modulo 256 and a byte function.

2. *Pseudo Hadamard Transform (PHT) Layer:* This is a module homomorphism α of V, so the bytes are mixed linearly. If α has matrix M with respect to the standard basis, then

$$M = \begin{pmatrix} 8 & 4 & 4 & 2 & 4 & 2 & 2 & 1 \\ 4 & 2 & 4 & 2 & 2 & 1 & 2 & 1 \\ 4 & 2 & 2 & 1 & 4 & 2 & 2 & 1 \\ 2 & 1 & 2 & 1 & 2 & 1 & 2 & 1 \\ 4 & 4 & 2 & 2 & 2 & 2 & 1 & 1 \\ 2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 & 2 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

After r rounds there is a final output transformation, for which the j^{th} byte output depends only on the j^{th} byte input and the j^{th} byte from subkey K_{2r+1}.

In the likelihood framework we are interested in defining partition functions ϕ_i on V which determine a partition on the key space such that the likelihood function on the key classes is easily calculated and can be used to find the true key class of the true key. We therefore consider the following two submodules of V,

$$\begin{aligned} R &= \langle\, e_2 - e_5, e_3 - e_5, e_6 - e_4, e_7 - e_4 \,\rangle, \\ S &= \langle\, e_1, e_2 + e_3 + e_5, e_4 + e_6 + e_7, e_8 \,\rangle, \end{aligned}$$

where e_i denotes the element with 1 as its i^{th} coordinate and 0 everywhere else. Now $V = R \oplus S$, and R and S are α-invariant submodules of rank 4. We thus define partion functions to be the module homomorphisms $\phi_i : V \to R \cong V/S$, which are defined by the matrix

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with respect to the standard basis. The class of an element $v \in V$ is its "projection" onto the submodule R.

We now consider the effect of the i^{th} round function on the classes. For an input in a given class to the keyed byte–separated layer, the probability of an output class is independent of the 1^{st} and 8^{th} bytes of subkey K_{2i-1} and K_{2i} as neither e_1 nor e_8 is included in any of the basis elements of R. Thus the output class of the keyed byte–separated layer is independent of the 1^{st} and 8^{th} bytes of the key. The PHT layer permutes the classes as R is an α-invariant subspace. Therefore the transition probability from any input class to the i^{th} round to any output class is independent of the 1^{st} and 8^{th} bytes of the key. By using this argument iteratively (and handling the final output transformation similarly), we have shown that the transition probability from plaintext class to ciphertext class is independent of the 1^{st} and 8^{th} bytes of the key. In the language of Section 7, we have found a probabilistic algebraic structure.

We can therefore define a key partition function $\sigma : K = \mathbb{Z}_{256}^8 \to \mathbb{Z}_{256}^6$ on the key space K by

$$(k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8)\sigma \;=\; (k_2, k_3, k_4, k_5, k_6, k_7).$$

For plaintext class x and ciphertext class y, the log–likelihood function is given by

$$\mathcal{L}[(x, y); k] = \mathcal{L}[(x, y); \sigma(k)] \qquad \text{for all } k.$$

There are 2^{48} key classes and so at most 2^{48} evaluations of the log–likelihood function. Thus if there are any non–negligible correlations between plaintext and ciphertext classes we have a reduced key search. A detailed analysis is given in [19]. In any case, we have given partitions of the plaintext and ciphertext spaces which are independent of a quarter of the key bits.

Shannon's principle of confusion [22], which is "to make the relationship between simple statistics of the ciphertext and simple statistics of the key a very complex and involved one", is one of SAFER's design criteria [14]. Our analysis of SAFER in the likelihood framework has given partitions of the plaintext and ciphertext spaces which are independent of a quarter of the key bits. We have thus exposed a cryptographic weakness of the SAFER algorithm and shown that it does not satisfy one of its own design criteria.

References

[1] D. Andelman. *Maximum Likelihood Estiamtion Applied to Cryptanalysis.* PhD thesis, Stanford University, 1979.

[2] D. Andelman and J. Reeds. On the Cryptanalysis of Rotor Machines and Substitution–Permutation Networks. *IEEE Transactions on Information Theory*, IT-28:578–584, 1982.

[3] B. Preneel, M. Nuttin, V. Rijmen and J. Buelens. Cryptanalysis of the CFB mode of DES with a Reduced Number of Rounds. In *Advances in Cryptology, Proceedings of CRYPTO 93*, pages 212–223. Springer–Verlag LNCS 773, 1994.

[4] E. Biham. New Types of Cryptanalytic Attacks using Related Keys. *Journal of Cryptology*, 7:229–246, 1994.

[5] E. Biham and A. Shamir. Differential Cryptanalysis of DES-like Cryptosystems. *Journal of Cryptology*, 4:3–72, 1991.

[6] L. Brynielsson. Hypothesenprüfung in der Kryptologie. Personal Communication, 1992.

[7] D. Chaum and J-H. Evertse. Cryptanalysis of DES with a Reduced Number of Rounds : Sequences of Linear Factors in Block Ciphers. In *Advances in Cryptology, Proceedings of CRYPTO 85*, pages 192–211. Springer–Verlag LNCS 218, 1986.

[8] D. Davies and S. Murphy. Pairs and Triplets of DES S-Boxes. *Journal of Cryptology*, 8:1–25, 1995.

[9] Y. Desmedt. *Analysis of the Security and New Algorithms for Modern Industrial Cryptography.* PhD thesis, Katholieke Universiteit of Leuven, 1984.

[10] J-H. Evertse. Linear Structures in Block Ciphers. In *Advances in Cryptology, Proceedings of EUROCRYPT 87*, pages 249–266. Springer–Verlag LNCS 304, 1988.

[11] H. Gilbert and G. Chassé. A Statistical Attack of the FEAL-8 Cryptosystem. In *Advances in Cryptology, Proceedings of CRYPTO 90*, pages 22–33. Springer–Verlag LNCS 537, 1991.

[12] B.S. Kaliski and M.J.B. Robshaw. Linear Cryptanalysis using Multiple Approximations. In *Advances in Cryptology, Proceedings of CRYPTO 94*, pages 26–39. Springer–Verlag LNCS 839, 1994.

[13] J.L. Lai, X. Massey and S. Murphy. Markov Ciphers and Differential Crytpanalysis. In *Advances in Cryptology, Proceedings of EUROCRYPT 91*, pages 17–38. Springer–Verlag LNCS 547, 1991.

[14] J.L. Massey. SAFER K-64: A Byte-Oriented Block-Ciphering Algorithm. In *Fast Software Encryption, Proceedings of Cambridge Security Workshop 1993*, pages 1–17. Springer–Verlag LNCS 809, 1994.

[15] M. Matsui. Linear Cryptanalysis Method for DES Cipher. In *Advances in Cryptology, Proceedings of EUROCRYPT 93*, pages 386–397. Springer–Verlag LNCS 765, 1994.

[16] M. Matsui. The First Experimental Cryptanalysis of the Data Encryption Standard. In *Advances in Cryptology, Proceedings of CRYPTO 94*, pages 1–11. Springer–Verlag LNCS 839, 1994.

[17] M. Matsui and A. Yamagishi. A new Method of Known Plaintext Attack of the FEAL cipher. In *Advances in Cryptology, Proceedings of EUROCRYPT 92*, pages 81–91. Springer–Verlag LNCS 658, 1993.

[18] M.J. Mihaljević and J.D. Golić. Convergence of a Bayesian Iterative Error–correction Proceedure on a Noisy Shift Register. In *Advances in Cryptology, Proceedings of EUROCRYPT 92*, pages 124–138. Springer–Verlag LNCS 658, 1993.

[19] S. Murphy. An Analysis of SAFER. *Journal of Cryptology*, submitted, 1995.

[20] National Bureau of Standards. Data Encryption Standard. *U.S. Department of Commerce*, FIPS pub. 46, 1977.

[21] J.A. Reeds and J.L. Manferdelli. DES has no Per Round Linear Factors. In *Advances in Cryptology, Proceedings of CRYPTO 84*, pages 377–389. Springer–Verlag LNCS 196, 1985.

[22] C.E. Shannon. Communication Theory of Secrecy Systems. *Bell. System Technical Journal*, 28:656–715, 1949.

[23] S.D. Silvey. *Statistical Inference.* Chapman and Hall, 1975.