Statistical and Algebraic Properties of DES

Stian Fauskanger[1] and Igor Semaev[2]

[1] Norwegian Defence Research Establishment (FFI), PB 25, 2027 Kjeller, Norway
[2] Department of Informatics, University of Bergen, Bergen, Norway

stian.fauskanger@ffi.no
igor@ii.uib.no

Abstract. D. Davies and S. Murphy found that there are at most 660 different probability distributions on the output from any three adjacent S-boxes after 16 rounds of DES [3]. In this paper it is shown that there are at most 72 different distributions for S-boxes 4, 5 and 6. The distributions from S-box triplets are linearly dependent and the dependencies are described. E.g. there are only 13 linearly independent distributions for S-boxes 4, 5 and 6. A coset representation of DES S-boxes which reveals their hidden linearity is studied. That may be used in algebraic attacks. S-box 4 can be represented by significantly fewer cosets than the other S-boxes and therefore has more linearity. Open cryptanalytic problems are stated.

Key words: DES · S-box · output distributions · linear dependencies · coset representation

1 Introduction

The Data Encryption Standard (DES) is a symmetric block cipher from 1977. It has block size of 64 bits and a 56-bit key. DES in its original form is deprecated due to the short key. Triple DES [10] however, is still used in many applications (e.g. in chip-based payment cards). It is therefore still important to analyze its security. DES is probably the most analyzed cipher, and is broken by linear [9] and differential [5] cryptanalysis. Even so, the most effective method in practice is still exhaustive search for the key. There are also some algebraic attacks on DES that can break 6-round DES [11].

Donald Davies and Sean Murphy described in [3] some statistical properties of S-boxes in DES. They found that there are at most 660 different distributions on the output from any three adjacent S-boxes after 16 rounds. These distributions divide the key space into classes where equivalent keys make the output follow the same distributions. The correct class is found by identifying which distribution a set of plaintext/ciphertext pairs follow. They used this to give a known-plaintext attack. The time complexity of the attack is about the same as brute-force attack and requires approximately $2^{56.6}$ plaintext/ciphertext pairs. The attack was improved by Biham and Biryukov [4] where the key can be found with 2^{50} plaintext/ciphertext pairs with 2^{50} operations. Later, Kunz-Jacques and Muller [12] further improved the attack to a chosen-plaintext attack with time complexity 2^{45} using 2^{45} chosen plaintexts.

In this paper we study new statistical and algebraic properties of DES. In Section 2 we show Davies and Murphy's results, using different notations than theirs. We also show a new exceptional property of S_4, and use this to show that there are fewer different distributions on the output from $S_4S_5S_6$ compared to other triplets. The new propertiey are related to the forth S-box in DES, and is used to show that the number of different distributions on the output from S-box 4, 5 and 6 is at most 72 (after 16 rounds). This divides the key space into fewer, but larger, classes compared to Davies and Murphy's results.

The distributions from S-box triplets are linearly dependent. We give a description of the relations between the distributions, and upper bound the number of linearly independent distributions for each triplet. E.g. among the 72 different distributions for S-box 4, 5 and 6 there are only 13 linearly independent.

A coset representation of the DES S-boxes is suggested in Section 3. It is found that S-box 4 is abnormal again. It can be covered by 10 sub-cosets while the other S-boxes require at least 16. Also, the coset representation of S-box 4 contains 6 sub-cosets of size 8, while the other S-boxes contain at most one sub-cosets of such size. The coset representation of S-boxes makes it possible to write the system of equations for DES in a more compact form than in [6,7].

Like the linear approximations discovered by Shamir [2] was later used by Matsui [9] to successfully break DES, these new properties might improve some attacks in the future. Two open problems are stated at the end of the paper. If solved that would improve statistical and algebraic attacks on DES.

1.1 Notations

Let X_{i-1}, X_i denote the input to the i-th round and X_i, X_{i+1} denote the i-th round output. So X_0, X_1 and X_{17}, X_{16} are plaintext and ciphertext blocks respectively, where the initial and final permutations are ignored. Let K_j be 48-bit round key at round j. Then

$$X_{j-1} \oplus X_{j+1} = Y_j \ , \quad Y_j = P(S(\bar{X}_j \oplus K_j)) \ , \tag{1}$$

where $\bar{X}_j$ is a 48-bit expansion of X_j, P denotes a permutation on 32 symbols, and S is a transform implemented by 8 S-boxes. Let S_i be a DES S-box, so

$$S_i(u_5, u_4, u_3, u_2, u_1, u_0) = (v_3, v_2, v_1, v_0) \ , \tag{2}$$

where u_j and v_j are input and output bits respectively.

2 (Ciphertext) $\oplus$ (plaintext) distributions and linear dependencies between the distributions

By (1), the XOR of the plaintext/ciphertext blocks are representable as follows

$$X_{17} \oplus X_1 = Y_2 \oplus Y_4 \oplus ... \oplus Y_{14} \oplus Y_{16} \ , \tag{3}$$

$$X_{16} \oplus X_0 = Y_1 \oplus Y_3 \oplus ... \oplus Y_{13} \oplus Y_{15} \ . \tag{4}$$

When analysing (3) and (4) we assume the round function inputs $X_2, X_4, \dots, X_{16}$ and $X_1, X_3, \dots, X_{15}$ are uniformly random and independent respectively. That common assumption was already in [3]. We study the joint distribution of bits in $X_{17} \oplus X_1$ and in $X_{16} \oplus X_0$ which come from the output of 3 adjacent S-boxes in DES round function, and therefore in Y_j. The output of 3 adjacent S-boxes is called (S_{i-1}, S_i, S_{i+1})-output when i is specified.

When we look at reduced number of rounds in DES (k rounds), we see that $X_{k+1} \oplus X_1$ and $X_k \oplus X_0$ follows the distribution for the XOR of $k/2$ round-outputs (for even k). We will throughout this paper use $2n$ to denote the number of rounds. n is the number if outputs that are XORed, and full DES is represented by $n = 8$.

2.1 Definitions and a basic lemma

We can assume the input to S_i is uniformly random then three distributions are related to each S_i. We use notation (2).

1. The distribution of $u_1, u_0, v_3, v_2, v_1, v_0$ is called **right hand side distribution** and we denote $p^{(i)}_{y,r} = \mathbf{Pr}((u_1, u_0) = y \text{ and } (v_3, v_2, v_1, v_0) = r)$.
2. The distribution of $u_5, u_4, v_3, v_2, v_1, v_0$ is called **left hand side distribution** and we denote $q^{(i)}_{x,r} = \mathbf{Pr}((u_5, u_4) = x \text{ and } (v_3, v_2, v_1, v_0) = r)$.
3. The distribution of $u_5, u_4, u_1, u_0, v_3, v_2, v_1, v_0$ is called **LR distribution** and we denote $Q^{(i)}_{x,y,r} = \mathbf{Pr}((u_5, u_4) = x, \text{ and } (u_1, u_0) = y, \text{ and } (v_3, v_2, v_1, v_0) = r)$.

Obviously, $p^{(i)}_{y,r} = \sum_x Q^{(i)}_{x,y,r}$ and $q^{(i)}_{x,r} = \sum_y Q^{(i)}_{x,y,r}$, the sums are over 2-bit x, y respectively.

Lemma 1. *For any 2-bit x, y and any 4-bit r holds*

$$p^{(i)}_{y\oplus 2,r} + p^{(i)}_{y,r} = \frac{1}{32} \ , \tag{5}$$

$$q^{(i)}_{x\oplus 1,r} + q^{(i)}_{x,r} = \frac{1}{32} \ , \tag{6}$$

$$Q^{(i)}_{x,y,r} + Q^{(i)}_{x,y\oplus 2,r} + Q^{(i)}_{x\oplus 1,y,r} + Q^{(i)}_{x\oplus 1,y\oplus 2,r} = \frac{1}{64} \ . \tag{7}$$

Proof. The equalities (5) and (6) were found directly from the values of $p^{(i)}_{y,r}$, $q^{(i)}_{x,r}$, for instance, see those distributions listed for S_4 in Appendix 1. Alternatively, by DES S-box definition, for any fixed (u_5, u_0) the distribution of (v_3, v_2, v_1, v_0) is uniform. So $(u_0, v_3, v_2, v_1, v_0)$ and $(u_5, v_3, v_2, v_1, v_0)$ are uniformly distributed and that implies (5) and (6) as A. Kholosha [1] later observed. The former implies (7) as well.

2.2 (S_{i-1}, S_i, S_{i+1})-output distributions

We study the distribution of the output from three adjacent S-boxes S_{i-1}, S_i, S_{i+1} in DES round function. An expression for that distribution slightly different from that in [3] is here presented.

Let $(a_5, a_4, a_3, a_2, a_1, a_0)$, $(b_5, b_4, b_3, b_2, b_1, b_0)$ and $(c_5, c_4, c_3, c_2, c_1, c_0)$ be the input to three adjacent S-boxes in one DES round. Then

$$(a_1, a_0) \oplus (b_5, b_4) = k \qquad \text{and} \qquad (b_1, b_0) \oplus (c_5, c_4) = k' \ ,$$

where k and k' are called the **common key bits**, they are both 2-bit linear combinations of round-key-bits. By $k_j = (k_{j1}, k_{j0})$ and $k'_j = (k'_{j1}, k'_{j0})$ we denote the common key bits in round j.

Let (r, s, t) be a 12-bit output from S_{i-1}, S_i, S_{i+1} in one DES round. Then

$$\mathbf{Pr}(r, s, t \mid k, k') = 2^4 \times \sum_{x,y} p^{(i-1)}_{x\oplus k,r} \, Q^{(i)}_{x,y,s} \, q^{(i+1)}_{y\oplus k',t} \ . \tag{8}$$

The distribution of (r, s, t) after $2n$ rounds is the n-fold convolution of $\mathbf{Pr}(r, s, t \mid k, k')$:

$$\mathbf{Pr}(r, s, t \mid k_1, k'_1, ..., k_n, k'_n) = \sum \prod_{i=1}^{n} \mathbf{Pr}(r_i, s_i, t_i \mid k_i, k'_i) \ ,$$

where the sum is over (r_i, s_i, t_i) such that $\bigoplus_i (r_i, s_i, t_i) = (r, s, t)$. By changing the order of summation and using (8) we get

$$\begin{aligned} &\mathbf{Pr}(r, s, t \mid k_1, k_1', ..., k_n, k_n') \\ &= 2^{4n} \times \sum p^{(i-1)}_{x_1 \oplus k_1, ..., x_n \oplus k_n, r} \times Q^{(i)}_{x_1, y_1, ..., x_n, y_n, s} \times q^{(i+1)}_{y_1 \oplus k_1', ..., y_n \oplus k_n', t} \ , \end{aligned} \tag{9}$$

the sum is over 2-bit $x_1, y_1, ..., x_n, y_n$, and where

$$\begin{aligned} p^{(i)}_{x_1, ..., x_n, r} &= \sum_{\bigoplus_j r_j = r} p^{(i)}_{x_1, r_1} \times \cdots \times p^{(i)}_{x_n, r_n} \ , \\ q^{(i)}_{y_1, ..., y_n, t} &= \sum_{\bigoplus_j t_j = t} q^{(i)}_{y_1, t_1} \times \cdots \times q^{(i)}_{y_n, t_n} \ , \\ Q^{(i)}_{x_1, y_1, ..., x_n, y_n, s} &= \sum_{\bigoplus_j s_j = s} Q^{(i)}_{x_1, y_1, s_1} \times \cdots \times Q^{(i)}_{x_n, y_n, s_n} \ . \end{aligned}$$

Davies-Murphy's results Lemma 1 implies

Lemma 2. *For any 2-bit $x_1, y_1 ..., x_n, y_n$ and 4-bit r, t*

$$\begin{aligned} p^{(i)}_{x_1 \oplus k_1, ..., x_n \oplus k_n, r} &= p^{(i)}_{x_1 \oplus k_{10}, ..., x_{n-1} \oplus k_{(n-1)0},\, x_n \oplus 2\overline{k},\, r} \ , \\ q^{(i)}_{y_1 \oplus k_1', ..., y_n \oplus k_n', t} &= q^{(i)}_{y_1 \oplus 2k_{11}', ..., y_{n-1} \oplus 2k_{(n-1)1}',\, y_n \oplus \overline{k'},\, t} \ , \end{aligned} \tag{10}$$

where $\overline{k}$ and $\overline{k'}$ are the parity of $(k_{11}, ..., k_{n1})$ and $(k_{10}', ..., k_{n0}')$.

Each value for the vector $(k_1, k_1', ..., k_n, k_n')$ can be mapped to a distribution on (r, s, t). Many of these distributions are equal to each other. Lemma 2 is now used to give an upper bound on the number of different distributions.

First, one can permute any (k_j, k_j') and (k_i, k_i') and get the same distribution. Also the distribution is defined by the parity of $(k_{11}, ..., k_{n1})$ and $(k_{10}', ..., k_{n0}')$. There are 4 values for the two parity-bits, and there are $\binom{3+n}{n}$ combinations for the remaining $2n$ bits $(k_{10}, ..., k_{n0})$ and $(k_{11}', ..., k_{n1}')$. Therefore there are at most $4 \times \binom{3+n}{n}$ different distributions on the output from three adjacent S-boxes. Table 1 lists the maximum number of different distributions after multiple rounds. Again, 16-round DES is specified by n=8.

Table 1. Upper bound on number of different distributions for $2n$ rounds

n	1	2	3	4	5	6	7	8
Upper bound	16	40	80	140	224	336	480	660

Exceptional property of S_4 In this section we find an exceptional property of S_4. In particular, we prove

Lemma 3. *For any* 2*-bit* x, y, a *and* 4*-bit* r *holds*

$$\sum_h p^{(4)}_{x\oplus a,h}\, p^{(4)}_{y\oplus a,h\oplus r} = \sum_h p^{(4)}_{x,h}\, p^{(4)}_{y,h\oplus r}\ .$$

Proof. By Lemma 1, $p^{(4)}_{x\oplus 2,h} + p^{(4)}_{x,h} = \frac{1}{32}$ for any 2-bit x and 4-bit h. It is easy to see the lemma is true for $a = 2$. All other cases are reduced to $a = 1$ and $x = y = 0$. Let

$$f(h) = \begin{cases} 0, & \text{if } h \notin \{0, 6, 9, 15\}; \\ 1, & \text{if } h \in \{0, 9\}; \\ -1, & \text{if } h \in \{6, 15\}\ . \end{cases}$$

From S_4 right hand side distribution values, see Table 5 in Appendix 1, we find

$$p^{(4)}_{x\oplus 1,\,h} + p^{(4)}_{x,\,h} = \frac{1}{32} + \frac{(-1)^{x_1} f(h)}{64} \tag{11}$$

and then

$$\sum_h f(h) f(h\oplus r) = 4\, f(r)\ , \tag{12}$$

$$\sum_h p^{(4)}_{x,h}\ f(h\oplus r) = \frac{(-1)^{x_1}\, 2\, f(r)}{64}\ , \tag{13}$$

for any 4-bit r and any 2-bit $x = (x_1, x_0)$. Hence

$$\sum_h p^{(4)}_{1,h}\ p^{(4)}_{1,h\oplus r} = \sum_h \left(\frac{1}{32} + \frac{f(h)}{64} - p^{(4)}_{0,h}\right) \left(\frac{1}{32} + \frac{f(h\oplus r)}{64} - p^{(4)}_{0,h\oplus r}\right) =$$

$$\sum_h \frac{f(h) f(h\oplus r)}{64^2} - 2\sum_h p^{(4)}_{0,h}\ \frac{f(h\oplus r)}{64} + \sum_h p^{(4)}_{0,h}\ p^{(4)}_{0,h\oplus r} = \sum_h p^{(4)}_{0,h}\ p^{(4)}_{0,h\oplus r}\ .$$

The lemma is proved.

This surprising property holds because (11), (12),(13) are true simultaneously for the right hand side distribution $p^{(4)}_{x,h}$.

Corollary 1. *For any* 2*-bit* $x_1, ..., x_n$ *and* 4*-bit* r *holds*

$$p^{(4)}_{x_1\oplus k_1,\ldots,x_n\oplus k_n,r} = p^{(4)}_{x_1,\ldots,x_{n-1},x_n\oplus\bar{k},r}\ ,$$

where $\bar{k} = k_1 \oplus \cdots \oplus k_n$.

Proof. By Lemma 3,

$$\sum_{h_1\oplus h_2=r} p^{(4)}_{x_1\oplus k_1,h_1}\, p^{(4)}_{x_2\oplus k_2,h_2} = \sum_{h_1\oplus h_2=r} p^{(4)}_{x_1,h_1}\, p^{(4)}_{x_2\oplus(k_1\oplus k_2),h_2}$$

for any x_1, x_2, k_1, k_2 and r. Therefore the corollary is true for $n = 2$. The general case follows recursively.

The number of different (S_4, S_5, S_6)-output distributions after $2n$ rounds Davies and Murphy found that there are at most $4 \times \binom{3+n}{n}$ different distributions of the output from 3 adjacent S-boxes after $2n$ rounds. In this section we show (S_4, S_5, S_6)-output has at most $(8n + 8)$ different distributions.

Lemma 4. *Let (r, s, t) be (S_4, S_5, S_6)-output after $2n$ rounds. There are at most $8n + 8$ different distributions (r, s, t) can follow.*

Proof. By Corollary 1 and Lemma 2 the distribution of (r, s, t) only depends on $\bigoplus_{j=1}^{n} k_j$, $\bigoplus_{j=1}^{n} k'_{j0}$ and common key bits $(k'_{11}, ..., k'_{n1})$, where the order of the last n bits is irrelevant. There are $n + 1$ combinations for $(k'_{11}, ..., k'_{n1})$ and 8 possible values for the three parity bits. The maximum number of different distributions is therefore at most $8n + 8$ as the lemma states.

We made a computer program that computed the actual number of different distributions for all 8 triplets. Table 2 lists the results for $n = 1, ..., 8$ together with the bound from Lemma 4 and Davies-Murphy's bound. Remark that 16-round DES is specified by $n = 8$.

Table 2. Number of different distributions for output of 3 adjacent S-boxes

n	1	2	3	4	5	6	7	8
$8n + 8$	16	24	32	40	48	56	64	72
(S_4, S_5, S_6)	16	24	32	40	48	56	64	72
$4 \times \binom{3+n}{n}$	16	40	80	140	224	336	480	660
$(S_{i-1}, S_i, S_{i+1}), i \neq 5$	16	40	80	140	224	336	480	660

It is not clear wether or not fewer different distribution can improve Davies-Murphy's attack. Intuitively, distinguishing between few distributions should be easier than distinguishing between many distributions (if the distance between them are approximately the same). At the same time, the number of keys in the class representing a given distribution is larger, so more work is required to identify the correct key in the class. Also, the triplet attack described by Davies and Murphy does not perform better than the attack based on the two S-box pairs in the triplet. An attack on S-box pairs S_4S_5 and S_5S_6 behave worse than the triplet attack on $S_6S_7S_8$ and $S_7S_8S_1$ (which is the best attack on triplets) [3]. We do not know if it is possible to alter Davies-Murphy's attack so that fewer distribution would give an advantage.

2.3 Linear dependencies between the distributions after $2n$ rounds

In this section we describe linear relations between distributions on the output from three adjacent S-boxes. We will see how (S_4, S_5, S_6) compares to the other triplets. A distribution can be represented by a row-vector $(v_0, ..., v_{2^{12}-1})$, where v_j is the probability of the output $j = (r, s, t)$.

Let M be a matrix whose rows are (S_{i-1}, S_i, S_{i+1})-output distributions. M is then called a **distribution matrix**. A non-zero vector r such that $rM = 0$ is called a linear relation for M. Let

R be a matrix whose rows are linear relations for M, then R is called a **relation matrix** for M. Then

$$\mathrm{rank}(M) \leq k - \mathrm{rank}(R)\ , \tag{14}$$

where k is the number of rows in M. There are five independent linear relations inside the right, LR and left distribution that can be used to find linear relation between the rows of M. By Lemma 1,

$$p^{(i)}_{x,r} - p^{(i)}_{x\oplus 1,r} + p^{(i)}_{x\oplus 2,r} - p^{(i)}_{x\oplus 3,r} = 0\ , \qquad q^{(i)}_{x,r} + q^{(i)}_{x\oplus 1,r} - q^{(i)}_{x\oplus 2,r} - q^{(i)}_{x\oplus 3,r} = 0\ .$$

In other words,

$$\sum_a C^1_a \times p^{(i)}_{x\oplus a,r} = 0 \qquad \text{and} \qquad \sum_a C^2_a \times q^{(i)}_{x\oplus a,r} = 0\ , \tag{15}$$

where $C^1 = (1, -1, 1, -1)$ and $C^2 = (1, 1, -1, -1)$. By Lemma 1, for any 2-bit x, y and 4-bit r

$$\sum_a Q^{(i)}_{x\oplus a,y,r} + Q^{(i)}_{x\oplus a,y\oplus 2,r} = \frac{1}{32}\ , \tag{16}$$

$$\sum_b Q^{(i)}_{x,y\oplus b,r} + Q^{(i)}_{x\oplus 1,y\oplus b,r} = \frac{1}{32}\ , \tag{17}$$

$$Q^{(i)}_{x,y,r} + Q^{(i)}_{x,y\oplus 2,r} + Q^{(i)}_{x\oplus 1,y,r} + Q^{(i)}_{x\oplus 1,y\oplus 2,r} = \frac{1}{64}\ . \tag{18}$$

One now subtracts (16) and (16) after changing $y \leftarrow y \oplus 1$, (17) and (17) after changing $x \leftarrow x \oplus 2$, then (18) and (18) after changing $y \leftarrow y \oplus 1$. So

$$\sum_{k,k'} C_{k,k'} \times Q_{x\oplus k,y\oplus k',r} = 0\ , \tag{19}$$

for any x, y and r, where C is any of

$$\begin{aligned} C^3 &= (1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1)\ , \\ C^4 &= (1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ -1,\ -1,\ -1,\ -1,\ -1,\ -1,\ -1,\ -1)\ , \\ C^5 &= (1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 1,\ -1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0)\ . \end{aligned}$$

For instance, C^3 comes from

$$\sum_a Q^{(i)}_{x\oplus a,y,r} + Q^{(i)}_{x\oplus a,y\oplus 2,r} - \sum_a Q^{(i)}_{x\oplus a,y\oplus 1,r} + Q^{(i)}_{x\oplus a,y\oplus 3,r} = 0\ .$$

Both (15) and (19) are used to build linear relations between the distributions of (r, s, t), the output from three adjacent S-boxes after one round.

Lemma 5.

For any k'
$$\sum_k C^1_k \times \boldsymbol{Pr}(r, s, t \mid k, k') = 0\ , \tag{20}$$

for any k
$$\sum_{k'} C^2_{k'} \times \boldsymbol{Pr}(r, s, t \mid k, k') = 0\ , \tag{21}$$

for $C \in \{C^3, C^4, C^5\}$
$$\sum_{k,k'} C_{k,k'} \times \boldsymbol{Pr}(r, s, t \mid k, k') = 0\ . \tag{22}$$

Proof. We will prove (20):

$$\begin{aligned}\sum_k C^1_k \times \mathbf{Pr}(r,s,t \mid k,k') &= 2^4 \times \sum_k C^1_k \times \left(\sum_{x,y} p^{(i-1)}_{x\oplus k,r}\, Q^{(i)}_{x,y,s}\, q^{(i+1)}_{y\oplus k',t} \right) \\ &= 2^4 \times \sum_{x,y} \sum_k C^1_k \times \left(p^{(i-1)}_{x\oplus k,r}\, Q^{(i)}_{x,y,s}\, q^{(i+1)}_{y\oplus k',t} \right) \\ &= 2^4 \times \sum_{x,y} Q^{(i)}_{x,y,s}\, q^{(i+1)}_{y\oplus k',t} \times \left(\sum_k C^1_k \times p^{(i-1)}_{x\oplus k,r} \right) = 0 \ .\end{aligned}$$

Similarly (21) is proved. We will prove (22).

$$\begin{aligned}\sum_{k,k'} C_{k,k'} \times \mathbf{Pr}(r,s,t \mid k,k') &= 2^4 \times \sum_{k,k'} C_{k,k'} \times \left(\sum_{x,y} p^{(i-1)}_{x,r}\, Q^{(i)}_{x\oplus k,y\oplus k',s}\, q^{(i+1)}_{y,t} \right) \\ &= 2^4 \times \sum_{x,y} \sum_{k,k'} C_{k,k'} \times \left(p^{(i-1)}_{x,r}\, Q^{(i)}_{x\oplus k,y\oplus k',s}\, q^{(i+1)}_{y,t} \right) \\ &= 2^4 \times \sum_{x,y} p^{(i-1)}_{x,r}\, q^{(i+1)}_{y,t} \times \left(\sum_{k,k'} C_{k,k'} \times Q^{(i)}_{x\oplus k,y\oplus k',s} \right) = 0 \ .\end{aligned}$$

Lemma 5 implies there are 11 linear dependencies between rows of the distribution matrix after one round. The rank of the relation matrix is 10. We have also computed the rank of the distribution matrix which is 6. Since there are 16 distributions in total, we have found all 10 independent linear relations between the distributions. Lemma 5 is now used to build linear relations between the distributions after $2n$ rounds.

Lemma 6. *For any* $(k_1, ..., k_n)$, $(k'_1, ..., k'_n)$, *and* i

$$\sum_{k_i} C^1_{k_i} \times \boldsymbol{Pr}(r,s,t \mid k_1,k'_1,...,k_n,k'_n) = 0 \ , \tag{23}$$

$$\sum_{k'_i} C^2_{k'_i} \times \boldsymbol{Pr}(r,s,t \mid k_1,k'_1,...,k_n,k'_n) = 0 \ , \tag{24}$$

$$\sum_{k_i,k'_i} C_{k_i,k'_i} \times \boldsymbol{Pr}(r,s,t \mid k_1,k'_1,...,k_n,k'_n) = 0 \ , \tag{25}$$

where $C \in \{C^3, C^4, C^5\}$.

Proof. It is enough to prove (23) for $i = 1$.

$$\begin{aligned}
&\sum_{k_1} C^1_{k_1} \times \mathbf{Pr}(r, s, t \mid k_1, k'_1, ..., k_n, k'_n) \\
&= \sum_{k_1} C^1_{k_1} \times \sum_{\bigoplus_j (r_j, s_j, t_j) = (r,s,t)} \prod_{j=1}^{n} \mathbf{Pr}(r_j, s_j, t_j \mid k_j, k'_j) \\
&= \sum_{\bigoplus_j (r_j, s_j, t_j) = (r,s,t)} \prod_{j=2}^{n} \mathbf{Pr}(r_j, s_j, t_j \mid k_j, k'_j) \sum_{k_1} C^1_{k_1} \times \mathbf{Pr}(r_1, s_1, t_1 \mid k_1, k'_1) = 0\ .
\end{aligned}$$

The proofs of (24) and (25) are similar.

Generating all relations from (23), (24) and (25) for all values of $(k_1, ..., k_n)$, $(k'_1, ..., k'_n)$, and i will make a relation matrix too large to calculate the rank when $n \geq 4$. We will instead consider a distribution matrix M, where each distribution occurs only once. We then generate a relation matrix for M. This way, by using (14), we find an upper bound on the rank of M for all triplets and $n \leq 8$, see columns 2 and 3 in Table 3. Triplet $S_4S_5S_6$ have an an upper bound on the rank which is lower than the other triplets. Full DES is specified by $n = 8$. We also computed the actual rank of M for each triplet, see columns 4-11.

Table 3. Rank of the distribution matrix for each triplet

n	$S_4S_5S_6$ *	$S_iS_{i+1}S_{i+2}$ * $i \neq 4$	$S_1S_2S_3$	$S_2S_3S_4$	$S_3S_4S_5$	$S_4S_5S_6$	$S_5S_6S_7$	$S_6S_7S_8$	$S_7S_8S_1$	$S_8S_1S_2$
1	6	6	6	6	6	6	6	6	6	6
2	7	9	9	9	9	**7**	9	9	9	9
3	8	13	13	13	13	**8**	13	13	13	13
4	9	18	18	18	18	**9**	18	18	18	18
5	10	24	24	24	24	**10**	24	24	24	24
6	11	31	30	31	29	**11**	31	31	31	31
7	12	39	36	39	34	**12**	39	39	39	39
8	13	48	42	48	39	**13**	48	48	48	48

* Upper bound.

Each distribution is determined by a class of DES keys. Table 3 data suggests a strong statistical dependence between ciphertexts generated with representatives of such classes. An open problem is stated in the end of this paper, which if solved, could make use these statistical dependencies to improve probability of success on Davies-Murphy's attack.

3 S-box coset representation and DES equations

For each S_i by (2) a set T_i of 10-bit strings

$$(u_5, u_4, u_3, u_2, u_1, u_0, v_3, v_2, v_1, v_0) \tag{26}$$

is defined. They are vectors in a vector space of dimension 10 over field with two elements F_2 denoted F_2^{10}. Let V be any subspace of F_2^{10}. For any vector a the set $a \oplus V$ is called a coset in F_2^{10}. Let $\dim V = s$, then there are 2^{10-s} cosets associated with V. Also we say $a \oplus V$ has dimension s as well. Any coset of dimension s is a set of the solutions for a linear equation system

$$a \oplus V = \{x \,|\, xA = b\},$$

where A is a matrix of size $10 \times (10-s)$, and rank $A = 10 - s$, and b is a row vector of length $10 - s$.

Any set $T \subseteq F_2^{10}$ may be partitioned into a union of its sub-cosets. We try to partition into sub-cosets of largest possible dimension, in other words of largest size. Denote the set of such cosets by M, it is constructed by the following algorithm. One first constructs a list of all sub-cosets in T maximal by inclusion. Let C be a maximal in dimension coset from the list, then C is added to M and the Algorithm recursively applies to $T \setminus C$. Let

$$M = \{C_1, \dots, C_r\}.$$

Therefore $x \in T$ if and only if x is a solution to the system $xA_k = b_k$ associated with $C_k \in M$.

The algorithm was applied to the vector sets T_i defined by DES S-boxes S_i. Let the sets of cosets M_i be produced. The results are summarised in Table 4, where $2^a\,4^b\,8^c$ means M_i contains a cosets of size 2, b cosets of size 4 and c cosets of size 8. The distribution is uneven. For instance, S_4 admits exceptionally many cosets of size 8. Disjoint sub-cosets which cover T_i for each $i = 1, \dots, 8$ are listed in Appendix 2, where strings (26) have integer number representation

Table 4. Coset distribution for S-boxes

i	1	2	3	4	5	6	7	8
coset dist.	$2^6\,4^{13}$	$2^4\,4^{14}$	$2^6\,4^{11}\,8$	$4^4\,8^6$	4^{16}	$2^6\,4^{13}$	$2^6\,4^{11}\,8$	$2^4\,4^{12}\,8$
# of cosets	19	18	18	10	16	19	18	17

$$u_5 2^9 + u_4 2^8 + u_3 2^7 + u_2 2^6 + u_1 2^5 + u_0 2^4 + v_3 2^3 + v_2 2^2 + v_1 2 + v_0\ .$$

3.1 More compact DES equations

Given one plaintext/ciphertext pair one constructs a system of equations in the key bits by introducing new variables after each S-box application, 128 equations for 16-round DES. By specifying S_i,

$$\begin{array}{c} \bar{X}_{ji} \oplus K_{ji} \\ P^{-1}(X_{j-1\,i} \oplus X_{j-2\,i}) \end{array} = \begin{bmatrix} 0 & \dots & 63 \\ S_i(0) & \dots & S_i(63) \end{bmatrix}\ , \tag{27}$$

with 64 right hand sides, 10-bit vectors T_i written column-wise. Here $\bar{X}_{ji}$ and K_{ji} are 6-bit sub-blocks of $\bar{X}_j$ and K_j respectively. To find the key such equations are solved. That may be done with methods introduced in [6], see also [7]. The complexity heavily depends on the number of right hand sides.

We get a more compact representation, that is with lower number of sides. We use the previous section notation. Let M_i contain r cosets. So $x \in T_i$ if and only if x is a solution to exactly one of the linear equation systems $xA_k = b_k, \quad k = 1, \ldots, r$. We cover the set of right hand side columns in (27) with sub-cosets from M_i and get (27) is equivalent to

$$\begin{bmatrix} \bar{X}_{ji} \oplus K_{ji} \\ P^{-1}(X_{j-1\,i} \oplus X_{j-2\,i}) \end{bmatrix} A_k = b_k, \quad k = 1, \ldots, r \tag{28}$$

in sense that an assignment to the variables is a solution to (27) if and only if it is a solution to one of (28). The number of subsystems(also called sides) in (28), denoted by r, is between 10 and 19 depending on the S-box. For instance, in case of S_4 the equation (28) has only 10 subsystems, while (27) has 64. Such reduction generally allows a faster solution, see [8].

4 Conclusion and open problems

In the present paper new statistical and algebraic properties of the DES encryption were found. They may have cryptanalytic implications upon resolving the following theoretical questions.

The first problem is within the statistical cryptanalysis. Let the cipher key space be split into n classes $K_1, \ldots, K_n$. Each class defines a multinomial distribution on some ≥ 2 outcomes, defined by plaintext and ciphertext bits. Let $P_1, \ldots, P_n$ be all such distributions computed a priori. Let $\nu(k)$ denote a vector of observations on above outcomes for an unknown cipher key k. It is well known that the problem "decide $k \in K_i$" may be solved with maximum likelihood method as in [3]. For the classification of several observation vectors $\nu(k_1), \ldots, \nu(k_s)$ the same method is applied.

Open problem is to improve the method (reduce error probabilities) given the vectors $P_1, \ldots, P_n$ are linearly dependent. That would improve Davies-Murphy type attacks against 16-round DES as for 660 different distributions on (S_{i-1}, S_i, S_{i+1}) outputs (72 for (S_4, S_5, S_6)) only ≤ 48 (13 for (S_4, S_5, S_6)) are linearly independent.

The second problem is related to algebraic attacks against ciphers. A new type time-memory trade-off for AES and DES was observed in [6,7]. Let m be the cipher key size. Let $\leq 2^l$ right hand sides be allowed in the combinations by Gluing of the MRHS equations [6,7] during solution. Gluing means writing several equations as one equation of the same type as (27). Then guessing $\leq m - l$ key-bits is enough before the system of equations is solved by finding and removing contradictory right-hand sides in pairwise agreeing of the current equations. The overall time complexity is at least $2^{m-l} \times 2^l = 2^m$ operations as for each guess one needs to run over the right hand sides of at least one of the equations. However coset representation allows reducing the number of sides by writing them as (28). In case of DES the equation (27) for $i = 4$ is written with only 10 sides instead of 64. For AES instead of 256 right hand sides one can do 64 for each of the equations, see [8]. The combination of two equations (27) with Gluing has $\leq 2^{12}$ right hand sides. With coset representation the number of sides is at most 19^2 (at most 100 for the combination of two equations from S_4). Open problem is to reduce the time complexity of the above trade-off by using coset representation.

References

1. A. Kholosha, Personal conversation with I. Semaev, September 2014
2. A. Shamir, *On the security of DES*, Advances in Cryptology-CRYPTO'85 Proceedings. Springer Berlin Heidelberg, 1986. pp. 280–281
3. D. Davies and S. Murphy, *Pairs and Triplets of DES S-boxes*, Journal of Crypt. vol. 8(1995), pp. 1–25
4. E. Biham and A. Biryukov, *An Improvement of Davies' Attack on DES*, Journal of Crypt. vol. 8(1997), pp. 195–205
5. E. Biham and A. Shamir, *Differential cryptanalysis of the full 16-round DES*, Advances in cryptology, proceedings of CRYPTO'92 (1992), pp. 487-496
6. H. Raddum and I. Semaev, *Solving Multiple Right Hand Sides linear equations*, Des., Codes and Crypt., vol. 49, pp. 147–160 (2008), extended abstract in Proceedings of WCC'07, 16-20 April 2007, Versailles, France, INRIA (2007)
7. H. Raddum, *MRHS Equation Systems*, In Selected Areas in Crypt. 2007, 14th Int. Workshop, LNCS 4876, pp. 232–245, 2007
8. I. Semaev and M. Mikuš, *Methods to solve algebraic equations in cryptanalysis*, Tatra Mt. Math. Publ. vol. 45(2010), pp. 107–136
9. M. Matsui, *Linear cryptanalysis method for DES cipher*, Advances in Cryptology-EUROCRYPT'93, Springer Berlin Heidelberg 1994, pp. 386–397
10. NIST, SP. *800-67, Revision 1: Recommendation for the Triple Data Encryption Algorithm (TDEA) Block Cipher*, National Institute of Standards and Technology (2012)
11. N.T. Courtois and G.V. Bard, *Algebraic cryptanalysis of the data encryption standard*, Cryptography and Coding. Springer Berlin Heidelberg, 2007. pp. 152–169.
12. S. Kunz-Jacques and F. Muller, *New Improvements of Davies-Murphy Cryptanalysis*, Advances in Cryptology-ASIACRYPT 2005, pp. 425–442

5 Appendix 1 - S_4 right, left and LR distribution

Section 2.1 define the right, left and LR distribution. The tables below show these distributions for S-box 4.

Table 5. Right hand side distribution of S-box 4 (each entry = $2^6 \times p^{(4)}_{x,r}$)

x\r	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	1	1	0	1	1	1	0	2	2	2	1	1	1	0	1	1
1	2	1	2	1	1	1	1	0	0	1	1	1	1	2	1	0
2	1	1	2	1	1	1	2	0	0	0	1	1	1	2	1	1
3	0	1	0	1	1	1	1	2	2	1	1	1	1	0	1	2

Table 6. Left hand side distribution of S-box 4 (each entry = $2^6 \times q^{(4)}_{x,r}$)

x\r	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	2	0	0	2	0	1	2	1	1	1	1	1	0	2	1	1
1	0	2	2	0	2	1	0	1	1	1	1	1	2	0	1	1
2	2	1	0	1	0	0	2	1	1	1	2	1	1	2	0	1
3	0	1	2	1	2	2	0	1	1	1	0	1	1	0	2	1

Table 7. LR distribution of S-box 4 (each entry = $2^6 \times Q^{(4)}_{x,y,r}$)

x y\r	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0 0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0
0 1	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
0 2	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0
0 3	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1
1 0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0
1 1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0
1 2	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1
1 3	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0
2 0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0
2 1	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0
2 2	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
2 3	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	1
3 0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1
3 1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0
3 2	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0
3 3	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0

6 Appendix 2 - Disjoint sub-cosets for DES S-boxes

$$\begin{aligned} M_1 = \{ & \{516,626\},\{678,697\},\{812,827\},\{841,894\},\{899,922\},\{944,992\},\\ & \{14,36,326,364\},\{16,87,175,232\},\{63,77,572,590\},\{97,130,545,706\},\\ & \{116,158,298,448\},\{178,221,938,965\},\{203,241,721,747\},\{259,282,653,660\},\\ & \{310,379,437,504\},\{348,389,783,982\},\{409,425,600,616\},\{467,487,851,871\},\\ & \{543,759,789,1021\}\}, \end{aligned}$$

$$\begin{aligned} M_2 = \{ & \{365,490\},\{855,870\},\{892,912\},\{949,1007\},\{15,19,33,61\},\{72,84,962,990\},\\ & \{110,119,134,159\},\{171,178,416,441\},\{195,216,676,703\},\{228,254,396,406\},\\ & \{265,295,475,501\},\{284,304,583,619\},\{322,337,737,754\},\{378,453,822,905\},\\ & \{512,602,931,1017\},\{541,558,795,808\},\{568,625,773,844\},\{650,659,717,724\}\}, \end{aligned}$$

$$\begin{aligned} M_3 = \{ & \{341,497\},\{605,624\},\{648,697\},\{707,759\},\{876,974\},\{978,1020\},\\ & \{10,29,110,121\},\{32,134,301,395\},\{73,80,207,214\},\{163,229,312,382\},\\ & \{180,250,662,728\},\{257,359,420,450\},\{274,412,779,901\},\{443,479,525,617\},\\ & \{529,687,788,938\},\{550,570,834,862\},\{801,883,949,999\},\\ & \{55,147,332,488,580,736,831,923\}\}, \end{aligned}$$

$$\begin{aligned} M_4 = \{ & \{45,56,290,311\},\{395,401,452,478\},\{711,733,968,978\},\{801,820,878,891\},\\ & \{7,29,328,338,683,689,996,1022\},\{78,91,257,276,749,760,930,951\},\\ & \{99,117,428,442,652,666,835,853\},\{128,150,495,505,608,630,783,793\},\\ & \{166,191,201,208,550,575,585,592\},\{234,243,357,380,522,531,901,924\}\}, \end{aligned}$$

$$\begin{aligned} M_5 = \{ & \{2,30,323,351\},\{44,59,230,241\},\{68,82,203,221\},\{97,124,170,183\},\\ & \{135,148,685,702\},\{264,277,577,604\},\{293,304,367,378\},\{397,462,657,722\},\\ & \{403,416,960,1011\},\{441,472,948,981\},\{489,502,516,539\},\{546,744,844,902\},\\ & \{568,711,869,922\},\{619,650,783,1006\},\{631,765,809,931\},\{790,831,848,889\}\}, \end{aligned}$$

$$\begin{aligned} M_6 = \{ & \{467,504\},\{591,693\},\{735,762\},\{795,836\},\{887,897\},\{918,971\},\\ & \{12,26,256,278\},\{33,63,74,84\},\{111,114,232,245\},\{137,151,162,188\},\\ & \{198,301,563,984\},\{217,305,642,874\},\{323,349,398,400\},\{356,423,830,1021\},\\ & \{382,443,521,716\},\{453,491,594,636\},\{532,613,800,849\},\{558,665,775,944\},\\ & \{680,739,941,998\}\}, \end{aligned}$$

$$\begin{aligned} M_7 = \{&\{402, 481\}, \{534, 587\}, \{621, 632\}, \{848, 872\}, \{926, 946\}, \{979, 1020\},\\ &\{29, 43, 143, 185\}, \{48, 66, 426, 472\}, \{91, 110, 329, 380\}, \{148, 160, 730, 750\},\\ &\{200, 237, 513, 548\}, \{209, 250, 969, 994\}, \{259, 286, 652, 657\}, \{300, 341, 447, 454\},\\ &\{307, 359, 675, 759\}, \{571, 605, 793, 895\}, \{778, 815, 896, 933\},\\ &\{4, 119, 389, 502, 692, 711, 821, 838\}\}, \end{aligned}$$

$$\begin{aligned} M_8 = \{&\{446, 498\}, \{519, 684\}, \{806, 911\}, \{949, 1019\}, \{13, 17, 100, 120\}, \{34, 63, 649, 660\},\\ &\{72, 134, 297, 487\}, \{154, 179, 857, 880\}, \{175, 203, 266, 366\}, \{215, 244, 530, 561\},\\ &\{309, 323, 828, 842\}, \{342, 379, 965, 1000\}, \{389, 400, 460, 473\}, \{555, 765, 768, 982\},\\ &\{580, 698, 877, 915\}, \{609, 631, 718, 728\}, \{93, 225, 284, 416, 606, 738, 799, 931\}\}. \end{aligned}$$