On Distinct Known Plaintext Attacks

Céline Blondeau and Kaisa Nyberg

Department of Computer Science, Aalto University School of Science, Finland
celine.blondeau@aalto.fi, kaisa.nyberg@aalto.fi

Abstract. Among statistical attacks, we often make a distinction between attacks in the linear context, for which the knowledge of plaintexts and corresponding ciphertexts is enough to perform the attack, and attacks in the differential context, for which the plaintexts are chosen. Such attacks are usually referred as known- or chosen-plaintext attacks. It is commonly believed that attacks in the known-plaintext model are more practical than attacks in the chosen-plaintext model. Nevertheless, it is usual in the literature, to only compare the data, time and memory complexity of these attacks without considering the model. In this paper, we reconsider some known-plaintext attacks, by considering them in the distinct-known-plaintext model. We explain and develop the statistical model for the multiple zero-correlation linear cryptanalysis, multidimensional linear cryptanalysis, as well as for the key-difference-invariant-bias related-key attack introduced at ASIACRYPT 2013. Based on these models validated by experiments, we improve attacks on some ciphers.

Keywords: multidimensional linear attack, zero-correlation linear attack, key-difference-invariant-bias attack, known plaintext, distinct known plaintext, statistical model.

1 Introduction

Among statistical attacks, we often make a distinction between attacks in the linear context, for which the knowledge of plaintexts and corresponding ciphertexts is enough to perform the attack, and attacks in the differential context, for which the plaintexts are chosen. Such attacks are usually referred as known- or chosen-plaintext (KP, CP) attacks. It is commonly believed that attacks in the known-plaintext model are more practical than attacks in the chosen-plaintext model. Nevertheless, it is usual in the literature to only compare the data, time and memory complexity of these attacks without considering the model.

In [4, 6, 7] zero-correlation linear attacks were introduced. For these attacks, depending on the number of used approximations and on the relation between the involved linear masks, different statistical models are presented. For instance in [4] the two models to compute respectively the data complexity of multiple zero-correlation linear attacks and multidimensional zero-correlation linear attacks are recalled. While the statistical model for multidimensional zero-correlation linear attack assumes that the plaintexts involved in the attacks are

distinct, the one for multiple zero-correlation linear attack assumes a normal distribution of the expected capacity for the wrong keys.

In [3] key-difference-invariant-bias attacks are introduced. In these related-key attacks, the attacker takes advantage of linear approximations that have the same bias for keys with a specific difference. The statistical model [3] for this related-key known-plaintext attack is similar to the one presented in [7] in the context of multiple zero-correlation linear attack.

In this paper we develop on distinct-known-plaintext (DKP) attacks. In particular, we show the importance of avoiding repetition when the data complexity of the attack is close to the full codebook. More importantly, while it is commonly believed that there is a difference in the evaluation of the data complexity of a multiple zero-correlation linear attack and of a multidimensional zero-correlation linear attack, we show that the difference is only in the way the plaintexts are handled. Based on theoretical and experimental observations, we provide new formulas to compute the data complexity of multiple zero-correlation linear attacks and key-difference-invariant bias attacks. As an illustration, we have instantiated our formulas on existing attacks reducing their data complexity. A model for distinct-known-plaintext multiple/multidimensional linear attacks is also introduced.

The outline of this paper is as follows. The notation is introduced in Sect. 2. In Sect. 3 we present the statistical model which is validated by experiments in Sect. 4. In Sect. 5 we apply our model on some existing attacks and improve their complexity. Sect. 6 concludes this paper.

2 Preliminaries

2.1 Linear Attacks

While the idea of using distinct-known plaintexts can be extended to any statistical attack, we focus on the most common known-plaintext statistical attacks which are generalizations of linear cryptanalysis [11].

Given an n-bit permutation F, we denote by $(u, v) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$, a pair of input and output masks. In linear attacks, we take advantage of linear approximations of the form $u \cdot x \oplus v \cdot F(x) = 0$. The strength of a linear relation is measured by its correlation. The correlation of a Boolean function $f_{u,v}(x) = u \cdot x \oplus v \cdot F(x)$ is defined as

$$\mathrm{cor}(u, v) = 2^{-n}\Big[\#\left\{x \in \mathbb{F}_2^n | f_{u,v}(x) = 0\right\} - \#\left\{x \in \mathbb{F}_2^n | f_{u,v}(x) = 1\right\}\Big].$$

In [2] the idea of taking advantage of multiple independent linear approximations is introduced. In [10] multidimensional linear attacks are presented. In these more recent attacks, the attacker takes advantage of all linear approximations with linear masks (u, v) $u \neq 0$ in a linear space.

The capacity C is used to collect the information of a set of linear approximations. Given sets of input and output linear masks U and V, it is defined as the sum of the squared correlations:

$$C = \sum_{u \in U, v \in V, u \neq 0} \mathrm{cor}^2(u, v).$$

While multiple/multidimensional linear attacks take advantage of a set of linear approximations with large capacity, multiple and multidimensional zero-correlation linear attacks [4, 6, 7] take advantage of linear approximations with correlation equal to zero. These attacks have been proven efficient on word-oriented structures such as Feistel-type ciphers. When multiple approximations with zero-correlation are used, the capacity C of the set of linear approximations is equal to zero.

In the remainder of this paper, we denote by ℓ the number of linear approximations involved in our attacks. Given s the dimension of the linear space $U \times V$, in (zero-correlation) multidimensional linear attacks we have $\ell = 2^s - 1$. The block cipher size is denoted by n.

2.2 Statistics

The data complexity N of a statistical attack corresponds to the number of plaintexts necessary to perform the attack. In general, we want to find the encryption key (right key) by differentiating the score of the right key from the one of the wrong keys. In (zero-correlation) multiple/multidimensional linear attacks the scoring function corresponds to the estimated capacity of the multiple/multidimensional linear approximations:

$$T = N \cdot \sum_{1 \leq i \leq \ell} (\hat{\mathrm{cor}}_i)^2,$$

where $\hat{\mathrm{cor}}_i$ is the empirical correlation of the i-th linear approximation. In (zero-correlation) multidimensional linear attacks the computation of this score T can be simplified and is equivalent to:

$$T = \sum_{j=0}^{\ell} \frac{(V[j] - N/(\ell+1))^2}{N/(\ell+1)},$$

where $V[j]$ corresponds to the number of occurrences of the j-th element of the multidimensional distribution.

Following the notation of [12], we denote by P_s the success probability and by a the advantage of the attack where 2^{-a} is the proportion of discarded keys.

Throughout this paper, we denote by Φ the cumulative distribution function of the central normal distribution. To simplify the notation, we also introduce: $\varphi_a = \Phi^{-1}(1 - 2^{-a})$ and $\varphi_{P_S} = \Phi^{-1}(P_S)$. Given μ_R and σ_R^2 (resp. μ_R and σ_w^2),

the mean and variance of the normal random variable T_R for the right key (resp. T_W for the wrong keys), we have (see i.e. [12]):

$$P_S \approx \Phi\left(\frac{|\mu_R - \mu_W| - \sigma_W \varphi_a}{\sigma_R}\right). \tag{1}$$

3 Statistical Models

3.1 Multiple and Multidimensional Zero-Correlation Linear Attacks

In [4] we have the following two estimates of the data complexity of a multiple and multidimensional zero-correlation linear attacks derived from [6, 7].

Lemma 1. *[7] The number N of known plaintexts required in a multiple zero-correlation linear cryptanalysis is:*

$$N \approx \frac{2^n(\varphi_{PS} + \varphi_a)}{\sqrt{\ell/2} - \varphi_a}. \tag{2}$$

The proof [7] follows from (1) using the fact that the distribution of T_R/N (resp T_W/N) can be estimated by a normal distribution with parameters $\mu_R = \frac{\ell}{N}$ and $\sigma_R = \frac{\sqrt{2\ell}}{N}$ (resp. $\mu_W = \frac{\ell}{N} + \frac{\ell}{2^n}$ and $\sigma_W = \frac{\sqrt{2\ell}}{N} + \frac{\sqrt{2\ell}}{2^n}$).

Lemma 2. *[4, 6][1] The number N of distinct-known plaintexts required in a multidimensional zero-correlation linear cryptanalysis is:*

$$N \approx \frac{2^n(\varphi_{PS} + \varphi_a)}{\sqrt{\ell/2} + \varphi_{PS}}. \tag{3}$$

The proof [4,6] follows from the use of the hypergeometric distribution as given in a more general case in Th. 1.

Assuming as in the proof of Lemma 1 that the correlation of the involved linear approximations are independent, we can adapt this result to the context of multiple zero-correlation linear cryptanalysis. In practice we observed, see Sect. 4.1, that the data complexity of a multiple zero-correlation linear attack can be estimated by (3) when assuming distinct-known plaintexts.

Corollary 1. *Given ℓ the number of used linear approximations. The data complexity of a known-plaintext multiple/multidimensional zero-correlation linear attack is given by (2). If we consider distinct-known plaintexts, the data complexity is given by (3).*

[1] The distribution of the random variables has been derived in [6], the estimation of the data complexity appears in [4].

Since for most attacks $0.5 \leq P_S \leq 0.99$, meaning that $0 \leq \varphi_{PS} \leq 2.4$, the difference between (3) and (2) is particularly noticeable when $\sqrt{\ell/2}$ and φ_a are in the same order of magnitude. From (3) and (2) we deduce that the success probability of a known-plaintext zero-correlation linear attack is:

$$P_S \approx \Phi\left(\frac{N}{2^n}\sqrt{\ell/2} - \varphi_a \cdot (\frac{N}{2^n} + 1)\right), \tag{4}$$

and the one of a distinct-known-plaintext zero-correlation linear attack is:

$$P_S \approx \Phi\left(\frac{N\sqrt{\ell/2}}{2^n - N} - \varphi_a \frac{2^n}{2^n - N}\right). \tag{5}$$

3.2 Multidimensional Linear Attacks

In the previous section we show that the data complexity of zero-correlation linear attacks is reduced once we consider distinct-known plaintexts. In this section we focus on the classical multidimensional linear attack.

Lemma 3. *[10] In the known-plaintext model, the random variable T_R involved in a multiple/multidimensional linear attack follows a normal distribution with parameters:*

$$\begin{aligned} &\mu_R \approx \ell + N \cdot C, \text{ and} \\ &\sigma_R^2 \approx 2(\ell + 2 \cdot N \cdot C). \end{aligned} \tag{6}$$

The random variable T_W follows a normal distribution with parameters $\mu_W = \ell$, $\sigma_W^2 = 2\ell$.

From these results, the data complexity of a known-plaintext multidimensional linear attack is computed as [10]:

$$N \approx \frac{\sqrt{4a\ell} + 4\Phi^{-1}(2P_S - 1)^2}{C}. \tag{7}$$

In the following we study the distribution of the variable T_R in the context of a distinct-known-plaintext attack.

Theorem 1. *Assuming distinct-known plaintexts the random variable T_R involved in a multiple/multidimensional linear attack follows a normal distribution with parameters:*

$$\begin{aligned} &\mu_R \approx \ell(1 - \frac{N}{2^n}) + N \cdot C, \text{and} \\ &\sigma_R^2 \approx 2\ell(1 - N/2^n)^2 + 4(1 - N/2^n)N \cdot C. \end{aligned} \tag{8}$$

Proof. For the purpose of this proof, we denote by $E(X)$ the mean of a random variable X and by $Var(X)$ its variance. We denote by Z_i the random variable corresponding to the number of solutions of the i-th equation of the form $u \cdot x \oplus v \cdot F(x) = 0$ and by $N \cdot p_i$ the expected number of solutions of this equation. Assuming that the plaintexts are distinct, from the hypergeometric distribution, we have $E(Z_i) = N \cdot p_i = \frac{N}{2}(1 + \mathrm{cor}_i)$ and $Var(Z_i) = Np_i\,(1 - p_i)\,\frac{2^n - N}{2^n - 1} = \frac{N}{4}(1 - cor_i^2)\frac{2^n - N}{2^n - 1}$ since $p_i = 1/2(1 + cor_i)$. Given $X_i = 2 \cdot Z_i/N - 1$ we deduce that $E(X_i) = \mathrm{cor}_i$ and $Var(X_i) = \frac{4}{N^2}Var(Z_i) \approx \frac{1}{N} \cdot \frac{2^n - N}{2^n - 1}$.

We have $T_R = N \sum_i X_i^2$. We denote $V = \sum_{i \le \ell} \frac{X_i^2}{Var(X_i)}$. The random variable V follows a non-central χ^2 distribution with parameters $E(V) = \ell + \lambda$ and $Var(V) = 2(\ell + 2\lambda)$ where $\lambda = \sum_i \frac{E(X_i)^2}{Var(X_i)} = N \cdot C\frac{2^n - 1}{2^n - N}$. From $V = T_R\frac{2^n - 1}{2^n - N}$, we deduce that $E(T_R) \approx \ell(1 - N/2^n) + N \cdot C$ and $Var(T_R) \approx 2\ell(1 - N/2^n)^2 + 4(1 - N/2^n) \cdot N \cdot C$. □

Experiments which confirm the value μ_R are presented in Sect. 4.2. From this result we can extract using $C = 0$ the distribution of the random variable T_R (see [6]) in the case of a distinct-known-plaintext zero-correlation linear attack.

4 Experimental Results

As usually zero-correlation linear attacks and multidimensional linear attacks are not targeting the same cipher construction, we have implemented experiments on a Feistel-type cipher and on a SPN-type cipher. The used ciphers are depicted in Fig. 1 and could correspond to scaled versions of CLEFIA [13] (a 16-bit type-II GFN with 4 branches) and PRESENT [5] (SMALLPRESENT-[32], a 32-bit SPN-cipher).

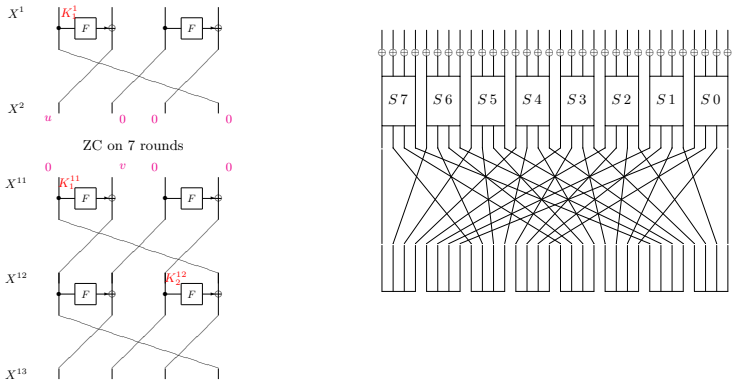

Fig. 1: Left: Description of the key-recovery attack done on a Type-II GFN. Right: One round of SMALLPRESENT-[8].

While in [14] experiments showing the distribution of μ_R and μ_W have been presented, there is, to the best of our knowledge, no mentioning of experimental zero-correlation linear attacks, in the literature.

4.1 Zero-Correlation Attacks on a Type-II GFN with 4 Branches

The results of our experiments averaged over 1000 keys are provided in Fig. 2. In these graphics we compare the success probability of multidimensional and multiple zero-correlation linear attacks with the theoretical ones given by (5) for distinct plaintexts and by (4) for non-distinct plaintexts. This experiments confirm the theory given in Sect. 3.1 showing that the same formula can be used to compute the complexity of multiple zero-correlation and multidimensional zero-correlation linear attacks.

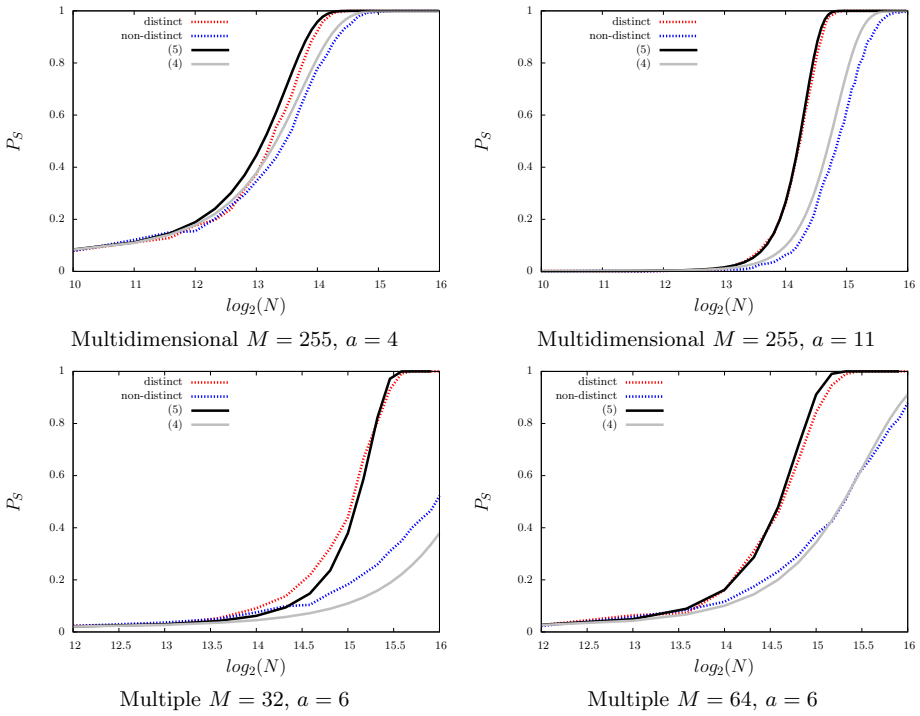

Fig. 2: Attacks on a type-II-GFN cipher. Top: multidimensional zero-correlation linear attacks, bottom: multiple zero-correlation linear attacks.

4.2 Experiments on SMALLPRESENT-[8]

In Fig. 3 we compare the experimental and theoretical mean μ_R of the variable T_R in the cases of distinct-known-plaintext and of known-plaintext distinguishing attacks. For this cipher, we observe that the theoretical value of μ_R given in Th.1 is accurate.

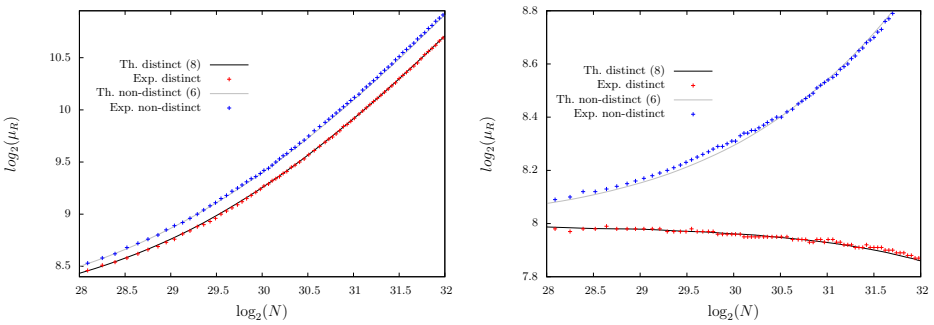

Fig. 3: The mean μ_R of T_R for a 8-bit multidimensional distribution ($\ell = 2^8 - 1$) over 9 rounds of SMALLPRESENT-[8] with capacity $C = 2^{-21.29}$ (left), over 12 rounds with capacity $C = 2^{-24.13} \leq \ell/2^n$ (right).

Remark 1. Distinct-known-plaintext attacks should require less data than known-plaintext attack when the data complexity of the attack is close to the full codebook. In practice, as illustrated by (8), this occurs when the capacity of the multiple/multidimensional linear approximations is such that $C \leq \ell/2^n$. In the distinct-known-plaintext model this would correspond to the case where μ_R decreases as the data complexity increases (see i.e. the right side of Fig. 3). An attack in this model will only be possible if we have a good estimate of the capacity. However, in practice, see for instance [9], we only have an underestimate of the capacity which can be problematic when it comes to estimate the data complexity of a distinct-known-plaintext multidimensional linear attack.

5 Applications

5.1 Multiple Zero-Correlation Linear Attacks

As explained in this paper, by considering distinct-known plaintexts we can use (3) to compute the data complexity of a multiple zero-correlation linear attack. As the data complexity of multidimensional linear attacks has already been computed under this setting, and because other comparable (in number of attacked rounds) attacks have been performed in the chosen-plaintext model, this should give us a better comparison factor. The result of our computation and a comparison with the best attacks on the block cipher Camellia [1] are provided in Table 1. The attack is from [4]. The data complexity has been computed using (3) instead of using (2) with the parameters of the attack chosen as $P_S = 0.85$ and $a = 96$ or $a = 160$. The time complexity has been computed according to the description given in [4].

Similarly we can improve the data complexity of the multiple zero-correlation linear attack on CAST-128 [15]. The parameters of the attack being $n = 128$, $\ell = 64770$, $a = 50$ and $P_S = 0.85$, the data complexity of the attack using known plaintexts[2] is $N = 2^{123.73}$ and the data complexity of the attack using distinct-known plaintexts is $N = 2^{123.67}$.

[2] With these parameters, the data complexity can not be equal to $2^{123.2}$ as given in [15].

Version	#R	Attack	ℓ	a	P_S	N	Time	Mem.	Ref.
Camellia-128	11	Impossible	-	-	1	$2^{118.4}$ CP	$2^{118.43}$	$2^{96.4}$	[8]
Camellia-128	11	Zero-Correlation	2^{14}	96	85%	$2^{125.3}$ KP	$2^{125.8}$	2^{112}	[4]
Camellia-128	11	Zero-Correlation	2^{14}	96	85%	$\mathbf{2^{125.1}}$ **DKP**	$2^{125.8}$	2^{112}	This paper
Camellia-192	12	Impossible	-	-	1	$2^{119.7}$ CP	$2^{161.06}$	$2^{147.7}$	[8]
Camellia-192	12	Zero-Correlation	2^{14}	160	85%	$2^{125.7}$ KP	$2^{125.8}$	2^{112}	[4]
Camellia-192	12	Zero-Correlation	2^{14}	160	85%	$\mathbf{2^{125.46}}$ **DKP**	$2^{125.8}$	2^{112}	This paper

Table 1: Best key-recovery attacks on Camellia (attacks starting from the first round). The memory is expressed in number of bytes. #R denotes the number of attacked rounds.

5.2 Key-Difference-Invariant-Bias Attacks

To the best of our knowledge, the only paper presenting attacks in this context is the similar paper [3]. In Table 2 we resume the complexity of the best related key-attacks on LBlock [17] and show that by assuming distinct-known plaintexts the data and time complexity of the attack can be improved. Similar improvement can be obtained for the related-key attack on TWINE presented in [3].

#R	Type	#Keys	ℓ	a	P_S	N	Time	Mem.	Ref.
23	Imp. Diff	4	-		100%	$2^{61.4}$ RKCP	$2^{78.3}$	$2^{61.4}$	[16]
24	Key Inv Bias	32	$2^{7.81}$	4.5	85%	$2^{62.29}$ RKKP	$2^{74.59}$	2^{61}	[3]
24	Key Inv Bias	32	$2^{7.81}$	8.5	85%	$2^{62.95}$ RKKP	$2^{70.67}$	2^{61}	[3]
24	Key Inv Bias	32	$2^{7.81}$	8.5	85%	$\mathbf{2^{62.38}}$**RKDKP**	$2^{70.67}$	2^{61}	This paper
24	Key Inv Bias	32	$2^{7.81}$	**16**	85%	$\mathbf{2^{62.84}}$ **RKDKP**	$\mathbf{2^{66.57}}$	2^{61}	This paper*

Table 2: Best related-key attacks on LBlock. *: Computation of the time complexity according to the description given in Sect. 5.3 of [3].

6 Conclusion

In this paper, we reconsider the statistical model for multiple zero-correlation linear and key-difference-invariant-bias attacks. We show that when using distinct plaintexts, the attacks can be performed using less plaintexts. We also consider for the first time a statistical model for distinct-known-plaintext multiple/multidimensional linear attacks. Nevertheless questions remain regarding how this model will be useful when we only have an underestimate of the capacity.

References

1. Kazumaro Aoki, Tetsuya Ichikawa, Masayuki Kanda, Mitsuru Matsui, Shiho Moriai, Junko Nakajima, and Toshio Tokita. Camellia: A 128-Bit Block Cipher Suitable for Multiple Platforms - Design and Analysis. In Douglas R. Stinson and Stafford E. Tavares, editors, *SAC 2000*, volume 2012 of *LNCS*. Springer, 2001.

2. Alex Biryukov, Christophe De Cannière, and Michaël Quisquater. On Multiple Linear Approximations. In *CRYPTO 2004*, pages 1–22, 2004.
3. Andrey Bogdanov, Christina Boura, Vincent Rijmen, Meiqin Wang, Long Wen, and Jingyuan Zhao. Key Difference Invariant Bias in Block Ciphers. In Kazue Sako and Palash Sarkar, editors, *ASIACRYPT 2013*, volume 8269 of *LNCS*, pages 357–376. Springer, 2013.
4. Andrey Bogdanov, Huizheng Geng, Meiqin Wang, Long Wen, and Baudoin Collard. Zero-Correlation Linear Cryptanalysis with FFT and Improved Attacks on ISO Standards Camellia and CLEFIA. In *SAC'13*, LNCS. Springer, 2014.
5. Andrey Bogdanov, Lars R. Knudsen, Gregor Leander, Christof Paar, Axel Poschmann, Matthew J. B. Robshaw, Yannick Seurin, and C. Vikkelsoe. PRESENT: An Ultra-Lightweight Block Cipher. In Pascal Paillier and Ingrid Verbauwhede, editors, *CHES*, volume 4727 of *LNCS*, pages 450–466. Springer, 2007.
6. Andrey Bogdanov, Gregor Leander, Kaisa Nyberg, and Meiqin Wang. Integral and Multidimensional Linear Distinguishers with Correlation Zero. In Xiaoyun Wang and Kazue Sako, editors, *ASIACRYPT*, volume 7658 of *LNCS*, pages 244–261. Springer, 2012.
7. Andrey Bogdanov and Meiqin Wang. Zero Correlation Linear Cryptanalysis with Reduced Data Complexity. In Anne Canteaut, editor, *FSE*, volume 7549 of *LNCS*, pages 29–48. Springer, 2012.
8. Christina Boura, María Naya-Plasencia, and Valentin Suder. Scrutinizing and Improving Impossible Differential Attacks: Applications to CLEFIA, Camellia, LBlock and Simon. In Palash Sarkar and Tetsu Iwata, editors, *ASIACRYPT 2014*, volume 8873 of *LNCS*, pages 179–199. Springer, 2014.
9. Joo Yeon Cho. Linear Cryptanalysis of Reduced-Round PRESENT. In Josef Pieprzyk, editor, *CT-RSA*, volume 5985 of *LNCS*, pages 302–317. Springer, 2010.
10. Miia Hermelin, Joo Yeon Cho, and Kaisa Nyberg. Multidimensional Extension of Matsui's Algorithm 2. In *FSE*, volume 5665 of *LNCS*, pages 209–227. Springer, 2009.
11. Mitsuru Matsui. Linear Cryptanalysis Method for DES Cipher. In Tor Helleseth, editor, *EUROCRYPT*, volume 765 of *LNCS*, pages 386–397. Springer, 1993.
12. Ali Aydin Selçuk. On Probability of Success in Linear and Differential Cryptanalysis. *J. Cryptology*, 21(1):131–147, 2008.
13. Taizo Shirai, Kyoji Shibutani, Toru Akishita, Shiho Moriai, and Tetsu Iwata. The 128-Bit Block cipher CLEFIA (Extended Abstract). In Alex Biryukov, editor, *FSE*, volume 4593 of *LNCS*, pages 181–195. Springer, 2007.
14. Hadi Soleimany and Kaisa Nyberg. Zero-correlation linear cryptanalysis of reduced-round LBlock. *Des. Codes Cryptography*, 73(2):683–698, 2014.
15. Long Wen, Meiqin Wang, Andrey Bogdanov, and Huaifeng Chen. General Application of FFT in Cryptanalysis and Improved Attack on CAST-256. In Willi Meier and Debdeep Mukhopadhyay, editors, *INDOCRYPT 2014*, volume 8885 of *LNCS*, pages 161–176. Springer, 2014.
16. Long Wen, Meiqin Wang, and Jingyuan Zhao. Related-Key Impossible Differential Attack on Reduced-Round LBlock. *J. Comput. Sci. Technol.*, 29(1):165–176, 2014.
17. Wenling Wu and Lei Zhang. LBlock: A Lightweight Block Cipher. In Javier Lopez and Gene Tsudik, editors, *ACNS*, volume 6715 of *LNCS*, pages 327–344, 2011.