Multidimensional Extension of Matsui's Algorithm 2

Miia Hermelin and Joo Yeon Cho and Kaisa Nyberg

Department of Information and Computer Science
Helsinki University of Technology

FSE 2009, 24th February, Leuven Belgium

Abstract

In the paper we studied different methods to extend Matsui's Alg. 2 to multiple dimensions. The efficiency of the methods were compared by the "advantage" (Selçuk). This presentation will focus on the method based on the log-likelihood ratio.

Outline

History - Multiple linear approximations

- Matsui EUROCRYPT'93: Uses one biased approximate linear equation to recover one bit information of the inner key (Alg. 1) or several bits of the last round key (Alg. 2)
- Robshaw and Kaliski CRYPTO'94: Alg. 1 and Alg. 2 several linear approximations, obtain one bit of information of the inner key (assumes statistical independence)
- Biryukov, et al., CRYPTO'04: Alg. 1 and Alg. 2 with multiple approximate linear equations (assumes statistical independence), recovers multiple bits of information of the key, success measured using *gain*
- Collard, et al., FSE 2008: Experiments of Biryukov's algorithms on Serpent

History - Probability distributions of multidimensional linear approximations

- Baignères, et al., ASIACRYPT'04: Distinguishing probability distributions based on log-likelihood ratio LLR
- Maximov, 2006: Algorithms for computing large probability distributions of multidimensional approximate linear equations
- Baignères and Vaudenay, ICITS'08: Different scenarios in hypothesis testing
- Hermelin, et al., ACISP 2008: Multidimensional Alg. 1, using G-test and comparison with the algorithm of Biryukov, et al.
- Hermelin, et al., Dagstuhl 2009 (to appear): Multidimensional Alg. 1 with LLR and χ^2

Assumption about statistical independence

Problem

Customised *special-purpose* statistical test under the assumption about *statistical independence* of simultaneous 1D linear approximations

Assumption about statistical independence

Problem

Customised *special-purpose* statistical test under the assumption about *statistical independence* of simultaneous 1D linear approximations

Our contribution

Use of LLR (optimal distinguisher) and other known tools and no assumption about statistical independence

Computing the multidimensional probability distribution

Problem

Computing *large probability distributions* needed in the multidimensional attack

Computing the multidimensional probability distribution

Problem

Computing *large probability distributions* needed in the multidimensional attack

Our contribution

Use Cramér-Wold Theorem (1936) for computing efficiently the probability distribution
$\Rightarrow$ Only information essential to the attack is taken into account and probability distribution computed with smaller dimension

Adding linearly or statistically dependent approximations

Problem

Is it correct to use linearly or statistically dependent 1D approximations?

Adding linearly or statistically dependent approximations

Problem

Is it correct to use linearly or statistically dependent 1D approximations?

Solution

Theoretical justification for this enhancement

Outline

Boolean functions

- Correlation between Boolean function $f : V_n \to V$ and zero

$$c(f) = c(f, 0) = 2^{-n} \left(\#\{\xi \in V_n \,|\, f(\xi) = 0\} - \#\{\xi \in V_n \,|\, f(\xi) \neq 0\}\right)$$

- $f = (f_1, \ldots, f_m) : V_n \to V_m$ an m-dimensional vector Boolean function
- $W = (w_1, \ldots, w_m) : V_n \to V_m$ a linear Boolean function
  $Wx = (w_1 \cdot x, \ldots, w_m \cdot x)$

Probability distribution

- Probability distribution (p.d.) $p = (p_0, \ldots, p_M)$ of random variable Y taking values in the set $\{0, 1, \ldots, M\}$:

$$\Pr(Y = y) = p_y,\ y = 0, \ldots, M,$$

- If random variable Y has p.d. p, denote $Y \sim p$
- θ uniform distribution
- Let $f : V_n \to V_m$ and $X \sim \theta$. If $f(X) \sim p$ we call p the p.d. of f

Kullback-Leibler distance

Definition

Let $p = (p_0, \ldots, p_M)$ and $q = (q_0, \ldots, q_M)$ be two p.d.'s. Their *relative entropy* or *Kullback-Leibler distance* is

$$D(p||q) = \sum_{\eta=0}^{M} p_\eta \log \frac{p_\eta}{q_\eta},$$

where we use the convention $0 \log 0/b = 0$, $b \neq 0$ and $b \log b/0 = \infty$.

Capacity

Close p.d.'s

We say that p.d p *is close to* p.d. q if $|p_\eta - q_\eta| \ll q_\eta, \forall \eta = 0, 1, \ldots, M$

Capacity

Close p.d.'s

We say that p.d p *is close to* p.d. q if $|p_\eta - q_\eta| \ll q_\eta$, $\forall \eta = 0, 1, \ldots, M$

Definition

The *capacity* between two p.d.'s p and q is defined by

$$C(p, q) = \sum_{\eta=0}^{M} \frac{(p_\eta - q_\eta)^2}{q_\eta}$$

We denote $C(p, \theta)$ by $C(p)$ and call $C(p)$ the capacity of p (cf. Biryukov, et al.). It is identical to the notion of squared Euclidean imbalance of p used by Baignères, et al.

Log-likelihood ratio (LLR)

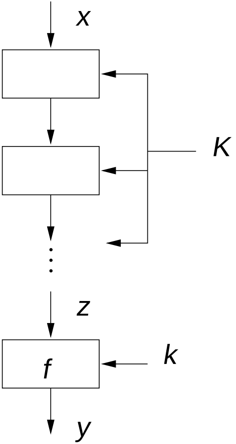
- Independent and identically distributed data $\hat{d}_1, \ldots, \hat{d}_N, \hat{d}_i \in V_m$, is drawn from p or q , $p \neq q$
- LLR is the optimal distinguisher between the two p.d.'s (hypotheses)
- Empirical p.d. $\hat{q} = (\hat{q}_0, \ldots, \hat{q}_M)$, $M = 2^m - 1$, where $\hat{q}_\eta = \frac{1}{N}\#\{i = 1, \ldots, N \,|\, \hat{d}_i = \eta\}$ are the relative observed frequencies
- We decide p if

$$\mathrm{LLR}(\hat{q}, p, q) = \sum_{\eta=0}^{M} N\hat{q}_\eta \log \frac{p_\eta}{q_\eta} \geq \gamma$$

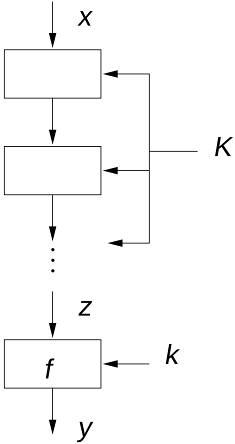
and otherwise we decide q, where γ is a threshold, usually taken equal to zero

Outline

Linear approximation of a block cipher (Alg. 2)

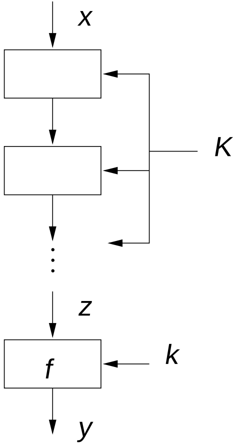


Linear approximation of a block cipher (Alg. 2)



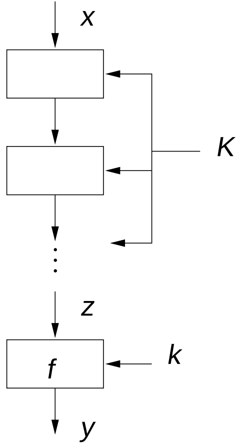
- Plaintext x, ciphertext y, last round key $k \in V_l$, all but last round key data K, last round function f, $z = f^{-1}(y, k)$

Linear approximation of a block cipher (Alg. 2)



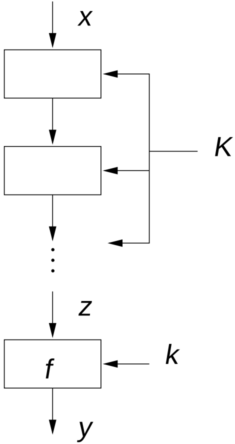
- Plaintext x, ciphertext y, last round key $k \in V_l$, all but last round key data K, last round function f, $z = f^{-1}(y, k)$
- Alg. 2 exploits 1D approximation $u \cdot x + w \cdot z + v \cdot K$ with non-negligible correlation c

Linear approximation of a block cipher (Alg. 2)



- Plaintext x, ciphertext y, last round key $k \in V_l$, all but last round key data K, last round function f, $z = f^{-1}(y, k)$
- Usually one has multiple 1D approximations with large correlations: $u_i \cdot x + w_i \cdot z + v_i \cdot K$, $i = 1, \ldots, m$ linearly independent *base approximations*

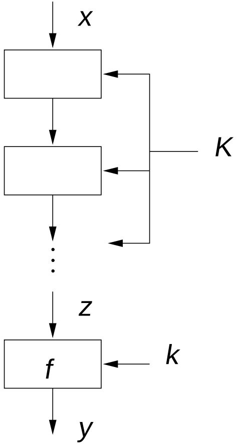
Linear approximation of a block cipher (Alg. 2)



- Plaintext x, ciphertext y, last round key $k \in V_l$, all but last round key data K, last round function f, $z = f^{-1}(y, k)$
- Usually one has multiple 1D approximations with large correlations: $u_i \cdot x + w_i \cdot z + v_i \cdot K$, $i = 1, \ldots, m$ linearly independent *base approximations*

Q: How to efficiently exploit them all?

Linear approximation of a block cipher (Alg. 2)



- Plaintext x, ciphertext y, last round key $k \in V_l$, all but last round key data K, last round function f, $z = f^{-1}(y, k)$
- Usually one has multiple 1D approximations with large correlations: $u_i \cdot x + w_i \cdot z + v_i \cdot K$, $i = 1, \ldots, m$ linearly independent *base approximations*

Q: How to efficiently exploit them all?

A: Determine the p.d. p of $Ux + Wz + VK$, $U = (u_1, \ldots, u_m)$, $W = (w_1, \ldots, w_M)$, $V = (v_1, \ldots, v_m)$

From one to many

- The p.d. p of $Ux + Wz + VK$ and 1D correlations $\rho(a) = c(a \cdot (Ux + Wz + VK))$, $a \in V_m$ are related as follows:

$$p_\eta = 2^{-m} \sum_{a \in V_m} (-1)^{a \cdot \eta} \rho(a),\ \eta \in V_m.$$

  That is, p.d. is determined using 1D projections, a well-known statistical method due to Cramér and Wold (1936)

From one to many

- The p.d. p of $Ux + Wz + VK$ and 1D correlations $\rho(a) = c(a \cdot (Ux + Wz + VK))$, $a \in V_m$ are related as follows:

$$p_\eta = 2^{-m} \sum_{a \in V_m} (-1)^{a \cdot \eta} \rho(a),\ \eta \in V_m.$$

  That is, p.d. is determined using 1D projections, a well-known statistical method due to Cramér and Wold (1936)
- One can and *must* add all non-negligible 1D approximations when calculating p

From one to many

- The p.d. p of $Ux + Wz + VK$ and 1D correlations $\rho(a) = c(a \cdot (Ux + Wz + VK))$, $a \in V_m$ are related as follows:

$$p_\eta = 2^{-m} \sum_{a \in V_m} (-1)^{a \cdot \eta} \rho(a), \ \eta \in V_m.$$

  That is, p.d. is determined using 1D projections, a well-known statistical method due to Cramér and Wold (1936)
- One can and *must* add all non-negligible 1D approximations when calculating p
- To strengthen the attack one should choose the m base approximations such that there are as many as possible non-neglible correlations $\rho(a)$, $a \in V_m$.

From one to many

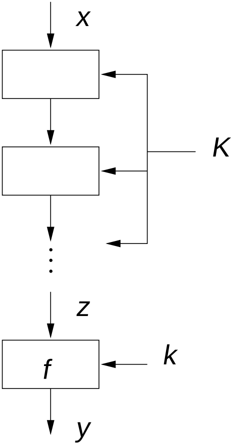
- The p.d. p of $Ux + Wz + VK$ and 1D correlations $\rho(a) = c(a \cdot (Ux + Wz + VK))$, $a \in V_m$ are related as follows:

$$p_\eta = 2^{-m} \sum_{a \in V_m} (-1)^{a \cdot \eta} \rho(a), \; \eta \in V_m.$$

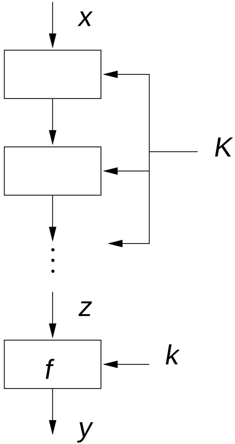
  That is, p.d. is determined using 1D projections, a well-known statistical method due to Cramér and Wold (1936)

- One can and *must* add all non-negligible 1D approximations when calculating p
- To strengthen the attack one should choose the m base approximations such that there are as many as possible non-neglible correlations $\rho(a)$, $a \in V_m$.
- We do not assume statistical independence of base approximations!

Multidimensional linear approximation of a block cipher

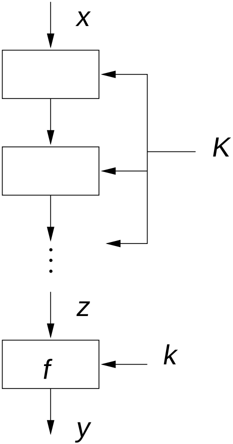


Multidimensional linear approximation of a block cipher



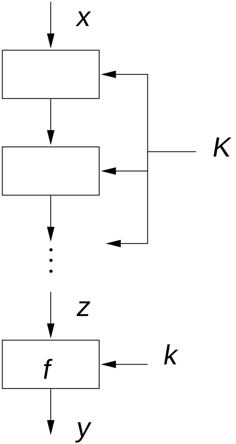
- Multidimensional linear approximation $Ux + Wz + VK \in V_m$ with p.d. p

Multidimensional linear approximation of a block cipher



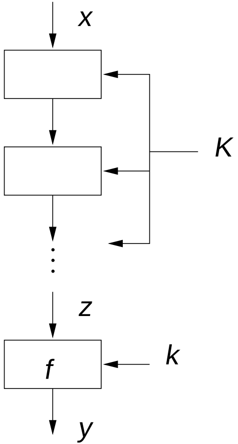
- Multidimensional linear approximation $Ux + Wz + VK \in V_m$ with p.d. p
- Parity bits $g = VK$ called the *inner key class*

Multidimensional linear approximation of a block cipher



- Multidimensional linear approximation $Ux + Wz + VK \in V_m$ with p.d. p
- Parity bits $g = VK$ called the *inner key class*
- For $g = VK$, the data $U\hat{x} + W\hat{z} \sim p^g$, a permutation of p dependent on g

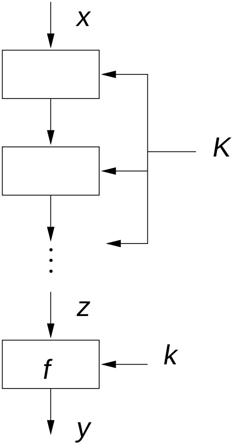
Multidimensional linear approximation of a block cipher



- Multidimensional linear approximation $Ux + Wz + VK \in V_m$ with p.d. p
- Parity bits $g = VK$ called the *inner key class*
- For $g = VK$, the data $U\hat{x} + W\hat{z} \sim p^g$, a permutation of p dependent on g

$\Rightarrow$ We have nice symmetry properties, e.g., $C(p^g) = C(p)$ for all $g \in V_m$

Multidimensional linear approximation of a block cipher



- Multidimensional linear approximation $Ux + Wz + VK \in V_m$ with p.d. p
- Parity bits $g = VK$ called the *inner key class*
- For $g = VK$, the data $U\hat{x} + W\hat{z} \sim p^g$, a permutation of p dependent on g

$\Rightarrow$ We have nice symmetry properties, e.g., $C(p^g) = C(p)$ for all $g \in V_m$

- g_0 is the right inner key class
  k_0 is the right last round key
  both unknown!

Outline

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:
  - **On-line phase**: collect data

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:
  - **On-line phase**: collect data
  - **Off-line phase**: using some statistic T, calculate mark $T(k)$ for each key $k \in V_l$ using the empirical data

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:
  - **On-line phase**: collect data
  - **Off-line phase**: using some statistic T, calculate mark $T(k)$ for each key $k \in V_l$ using the empirical data
  - **Sort phase**: rank keys according to order statistic $T(k)$

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:
  - **On-line phase**: collect data
  - **Off-line phase**: using some statistic T, calculate mark $T(k)$ for each key $k \in V_l$ using the empirical data
  - **Sort phase**: rank keys according to order statistic $T(k)$
  - **Search phase**: run through the list to find k_0 which should be at the top of the list

Key ranking as a statistical problem

- Given key candidates $k \in V_l$ we wish to find the right key k_0
- Four phases:
  - **On-line phase**: collect data
  - **Off-line phase**: using some statistic T, calculate mark $T(k)$ for each key $k \in V_l$ using the empirical data
  - **Sort phase**: rank keys according to order statistic $T(k)$
  - **Search phase**: run through the list to find k_0 which should be at the top of the list
- How well T ranks? Measure using *advantage*

Advantage

Definition (Selçuk's a-bit advantage, JoC'08)

We say that a key recovery attack for an l-bit key achieves an advantage of a bits over exhaustive search, if the correct key is ranked among the top 2^{l-a} out of all 2^l key candidates.

Advantage

Definition (Selçuk's a-bit advantage, JoC'08)

We say that a key recovery attack for an l-bit key achieves an advantage of a bits over exhaustive search, if the correct key is ranked among the top 2^{l-a} out of all 2^l key candidates.

We derived the relationship between the data complexity of the on-line phase and the advantage to describe the trade-off between search phase and data complexity.

Outline

On-line phase of Alg. 2

- Draw data $(\hat{x}_i, \hat{y}_i), i = 1, \ldots, N$

On-line phase of Alg. 2

- Draw data $(\hat{x}_i, \hat{y}_i), i = 1, \ldots, N$
- For each $k \in V_l$, calculate empirical p.d. $\hat{q}^k = (\hat{q}^k_0, \ldots, \hat{q}^k_M)$:

On-line phase of Alg. 2

- Draw data $(\hat{x}_i, \hat{y}_i), i = 1, \ldots, N$
- For each $k \in V_l$, calculate empirical p.d. $\hat{q}^k = (\hat{q}^k_0, \ldots, \hat{q}^k_M)$:

$$\hat{q}^k_\eta = \frac{1}{N}\#\{i = 1, \ldots, N \mid U\hat{x}_i + W\hat{z}^k_i = \eta\}, \text{ where } \hat{z}^k_i = f^{-1}(\hat{y}_i, k)$$

On-line phase of Alg. 2

- Draw data $(\hat{x}_i, \hat{y}_i), i = 1, \dots, N$
- For each $k \in V_l$, calculate empirical p.d. $\hat{q}^k = (\hat{q}_0^k, \dots, \hat{q}_M^k)$:

$$\hat{q}_\eta^k = \frac{1}{N}\#\{i = 1, \dots, N \,|\, U\hat{x}_i + W\hat{z}_i^k = \eta\}, \text{ where } \hat{z}_i^k = f^{-1}(\hat{y}_i, k)$$

- Decrypting with the *right* round key, we have empirical p.d. $\hat{q}^{k_0} \sim p^{g_0}$, where $g_0 \in V_m$ is unknown

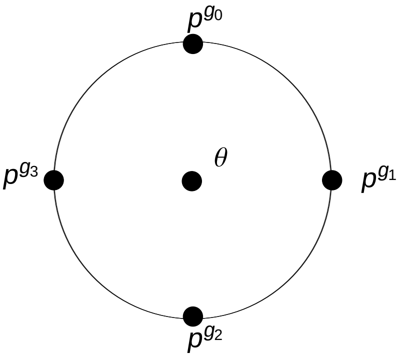
On-line phase of Alg. 2

- Draw data $(\hat{x}_i, \hat{y}_i), i = 1, \ldots, N$
- For each $k \in V_l$, calculate empirical p.d. $\hat{q}^k = (\hat{q}_0^k, \ldots, \hat{q}_M^k)$:

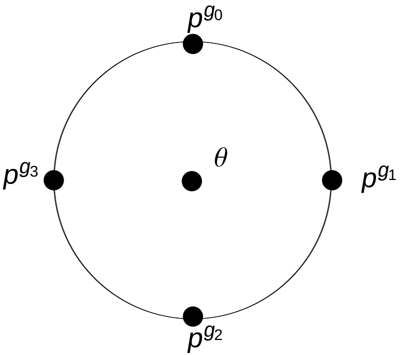
$$\hat{q}_\eta^k = \frac{1}{N}\#\{i = 1, \ldots, N \,|\, U\hat{x}_i + W\hat{z}_i^k = \eta\}, \text{ where } \hat{z}_i^k = f^{-1}(\hat{y}_i, k)$$

- Decrypting with the *right* round key, we have empirical p.d. $\hat{q}^{k_0} \sim p^{g_0}$, where $g_0 \in V_m$ is unknown
- Decryption with *wrong* key $k \neq k_0$ means additional encryption such that $\hat{q}^k \sim \theta$, $k \neq k_0$ (*Wrong-key Randomisation Hypothesis*)

Wrong-key Randomisation Hypothesis

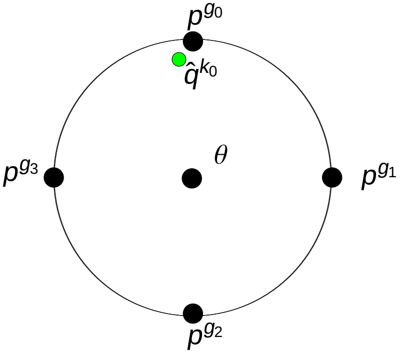


Wrong-key Randomisation Hypothesis



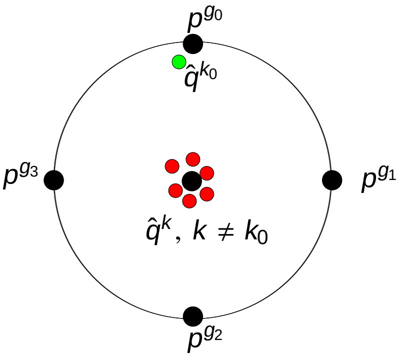
- Ranking statistic (off-line):
  $L(k) = \max_{g \in V_m} L(k, g)$,
  where
  $L(k, g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$

Wrong-key Randomisation Hypothesis



- Ranking statistic (off-line): $L(k) = \max_{g \in V_m} L(k, g)$, where $L(k, g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$
- $\hat{q}^{k_0}$ follows p^{g_0} for some $g_0 \in V_m$ (an unknown permutation of p) and not any other p^g, $g \neq g_0$ or θ, then $L(k_0, g_0) > 0$

Wrong-key Randomisation Hypothesis



- Ranking statistic (off-line): $L(k) = \max_{g \in V_m} L(k,g)$, where $L(k,g) = \mathrm{LLR}(\hat{q}^k, p^g, \theta)$
- $\hat{q}^{k_0}$ follows p^{g_0} for some $g_0 \in V_m$ (an unknown permutation of p) and not any other p^g, $g \neq g_0$ or θ, then $L(k_0, g_0) > 0$
- $\hat{q}^k, k \neq k_0$ follows θ rather than p^g, for any $g \in V_m$, then $L(k,g) < 0, k \neq k_0$

Other approaches

- We could also interpret the problem as a goodness-of-fit test

Other approaches

- We could also interpret the problem as a goodness-of-fit test

$\Rightarrow$ χ^2-based ranking statistic

Other approaches

- We could also interpret the problem as a goodness-of-fit test

$\Rightarrow$ χ^2-based ranking statistic

- Similar calculations

Other approaches

- We could also interpret the problem as a goodness-of-fit test

$\Rightarrow$ χ^2-based ranking statistic

- Similar calculations
- A weaker method both in theory and in experiments

Other approaches

- We could also interpret the problem as a goodness-of-fit test

$\Rightarrow$ χ^2-based ranking statistic

- Similar calculations
- A weaker method both in theory and in experiments
- Unlike LLR, does not benefit from using (many) multiple approximations

Other approaches

- We could also interpret the problem as a goodness-of-fit test

$\Rightarrow$ χ^2-based ranking statistic

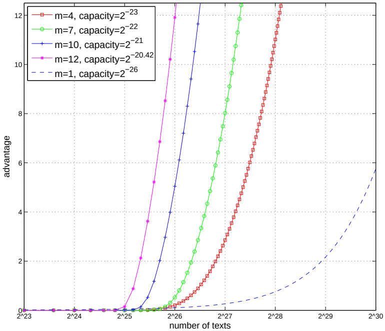
- Similar calculations
- A weaker method both in theory and in experiments
- Unlike LLR, does not benefit from using (many) multiple approximations
- Different ranking statistics are also possible, but the LLR is optimal and its statistical behaviour is well-known

Outline

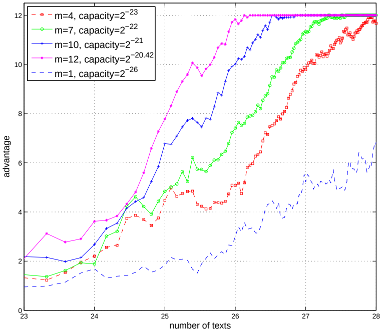
Experimental results

- Experiments on 5-round Serpent, 16 keys, k has 12 bits
- Comparison between LLR and χ^2, in theory and practice
- LLR is more powerful
- Theoretical and experimental advantage behave similarly with dimension m of linear approximation
- For this cipher the optimal value for LLR is $m = 12$ and for χ^2 it is $m = 4$

Advantage of LLR-method as a function of data complexity for different *m*



Theoretical prediction



Empirical results

Outline

Results

- Multidimensional extensions of Matsui's Algorithm 2

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations
- No need to compute huge distributions, just smaller distributions from 1D correlations

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations
- No need to compute huge distributions, just smaller distributions from 1D correlations
- Theoretical justification for the enhancement of Biryukov's method (adding linearly dependent strong 1D approximations)

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations
- No need to compute huge distributions, just smaller distributions from 1D correlations
- Theoretical justification for the enhancement of Biryukov's method (adding linearly dependent strong 1D approximations)
- Order statistics for measuring success of key ranking and to find trade-off between search phase and data complexity

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations
- No need to compute huge distributions, just smaller distributions from 1D correlations
- Theoretical justification for the enhancement of Biryukov's method (adding linearly dependent strong 1D approximations)
- Order statistics for measuring success of key ranking and to find trade-off between search phase and data complexity
- Estimates for data complexities calculated

Results

- Multidimensional extensions of Matsui's Algorithm 2
- No assumption about statistical independence of the 1D linear approximations
- No need to compute huge distributions, just smaller distributions from 1D correlations
- Theoretical justification for the enhancement of Biryukov's method (adding linearly dependent strong 1D approximations)
- Order statistics for measuring success of key ranking and to find trade-off between search phase and data complexity
- Estimates for data complexities calculated
- Different methods and dimensions can be compared

Conclusions

- For fixed dimension m of linear approximation, LLR has higher advantage than χ^2
- Advantage of LLR increases with m further than advantage of χ^2

Conclusions

- For fixed dimension m of linear approximation, LLR has higher advantage than χ^2
- Advantage of LLR increases with m further than advantage of χ^2

$\Rightarrow$ If no reason to suspect a significant error in p, we recommend using LLR rather than χ^2

Open questions and future work

- Measure advantage for finding both last round key and inner key class
- Extensions to nonlinear cryptanalysis (cf. Baignères, et al.2004)?
- Other ciphers? Stream ciphers?

Thank you!