University of Milano-Bicocca
Department of Informatics, Systems and Communications

Construction of Boolean functions and S-boxes with evolutionary algorithms

Luca Mariot

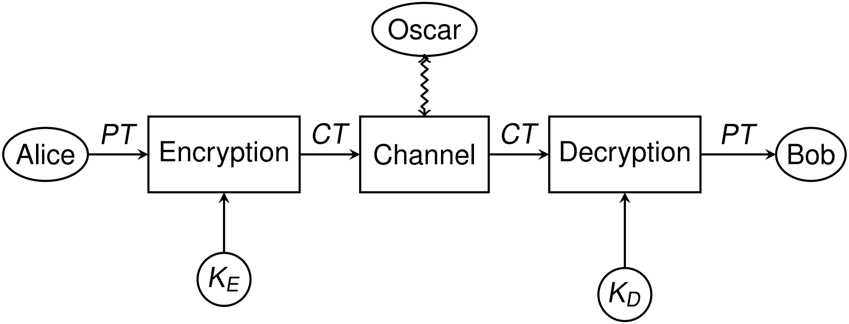
luca.mariot@unimib.it

Laboratory of population-based optimisation methods

Milan – June 17, 2019

Cryptography

Basic Goal of Cryptography: Enable two parties (Alice and Bob, A and B) to securely communicate over an insecure channel, even in presence of an opponent (Oscar, O)



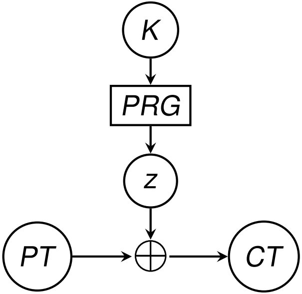
- PT: plaintext
- CT: ciphertext
- K_E: encryption key
- K_D: decryption key

Symmetric cryptosystems

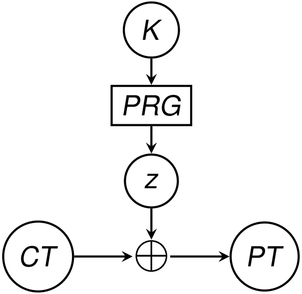
Symmetric cryptosystems ($K_E = K_D = K$) can be classified as:

- *Stream ciphers*: each symbol of *PT* is combined with a symbol of a *keystream*, computed from K
  - Grain
  - Trivium
  - ...
- *Block ciphers*: *PT* is divided in *blocks* combined with *round keys* derived from K through a *round function*
  - DES
  - Rijndael (AES)
  - ...

Vernam Stream Cipher



(a) Encryption



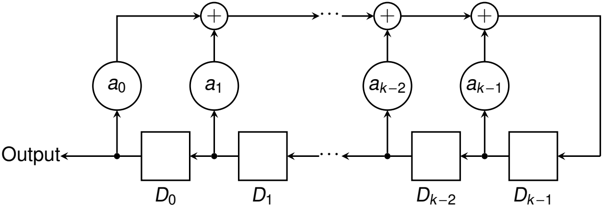
(b) Decryption

- K: secret key
- PRG: Pseudorandom Generator
- z: keystream
- $\bigoplus$: bitwise XOR
- PT: Plaintext
- CT: Ciphertext

Linear Feedback Shift Registers (LFSR)

- Device computing the binary linear recurring sequence

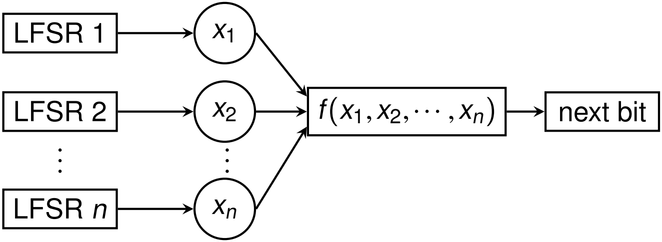
$$s_{n+k} = a + a_0 s_n + a_1 s_{n+1} + \cdots + a_{k-1} s_{n+k-1}$$



- Too weak as a PRG: $2k$ consecutive bits of keystream are enough to recover the LFSR initialization via the Berlekamp-Massey algorithm

An Example of PRG: The Combiner Model

- a Boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$ combines the outputs of n LFSR [1]



- Security of the combiner $\Leftrightarrow$ cryptographic properties of f

Boolean Functions - Basic Definitions

Boolean function: a mapping $f : \mathbb{F}_2^n \to \mathbb{F}_2$, where $\mathbb{F}_2 = \{0, 1\}$

- Truth table: vector Ω_f specifying $f(x)$ for all $x \in \mathbb{F}_2$

(x_1, x_2, x_3)	000	100	010	110	001	101	011	111
Ω_f	0	1	1	1	1	0	0	0

- Algebraic Normal Form (ANF): Sum (XOR) of products (AND) over the finite field $\mathbb{F}_2$

$$f(x_1, x_2, x_3) = x_1 \cdot x_2 \oplus x_1 \oplus x_2 \oplus x_3$$

- Walsh Transform: correlation with the *linear* functions defined as $\omega \cdot x = \omega_1 x_1 \oplus \cdots \oplus \omega_n x_n$

$$\hat{F}(\omega) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) \oplus \omega \cdot x}$$

Cryptographic Properties: Balancedness

- Hamming weight $w_H(f)$: number of 1s in Ω_f
- A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is balanced if $w_H(f) = 2^{n-1}$
- Walsh characterization: f balanced $\Leftrightarrow \hat{F}(0) = 0$

(x_1, x_2, x_3)	000	100	010	110	001	101	011	111
Ω_f	0	1	1	1	1	0	0	0

$$\Downarrow$$

f is balanced

- Unbalanced functions present a statistical bias that can be exploited in attacks

Cryptographic Properties: Algebraic Degree

- Algebraic degree d: the degree of the multivariate polynomial representing the ANF of f

$$f(x_1, x_2, x_3) = x_1 \cdot x_2 \oplus x_1 \oplus x_2 \oplus x_3$$

$$\Downarrow$$

$$f \text{ has degree } d = 2$$

- *Linear* functions $\omega \cdot x = \omega_1 x_1 \oplus \cdots \oplus \omega_n x_n$ have degree $d = 1$
- Boolean functions of high degree make the attack based on Berlekamp-Massey algorithm less effective

Cryptographic Properties: Nonlinearity

- Nonlinearity $nl(f)$: Hamming distance of f from linear functions
- Walsh characterization:

$$nl(f) = 2^{n-1} - \frac{1}{2} \max_{\omega \in \mathbb{F}_2^n} \left\{ \left| \hat{F}(\omega) \right| \right\}$$

(x_1, x_2, x_3)	000	100	010	110	001	101	011	111
Ω_f	0	1	1	1	1	0	0	0
$\hat{F}(\omega)$	0	0	0	0	-4	4	4	4

$$\Downarrow$$

$$nl(f) = 2^{3-1} - \frac{1}{2} \cdot 4 = 2$$

- Functions with high nonlinearity resist fast-correlation attacks

Cryptographic Properties: Resiliency

- *t*-Resiliency: when fixing any t variables, the restriction of f stays balanced
- Walsh characterization:

$$\hat{F}(\omega) = 0 \ \forall \omega : w_H(\omega) \leq t$$

(x_1, x_2, x_3)	000	100	010	110	001	101	011	111
Ω_f	0	1	1	1	1	0	0	0
$\hat{F}(\omega)$	0	0	0	0	-4	4	4	4

$$\Downarrow$$

$$F(001) = -4 \Rightarrow f \text{ is NOT 1-resilient}$$

- Resilient functions of high order t resist to correlation attacks

Bounds and Trade-offs

In summary, $f : \mathbb{F}_2^n \to \mathbb{F}_2$ should:

- be balanced
- be resilient of high order m
- have high algebraic degree d
- have high nonlinearity nl

But most of these properties cannot be satisfied simultaneously!

- *Covering Radius bound*: $nl \leq 2^{n-1} - 2^{\frac{n}{2}-1}$
- *Siegenthaler's bound*: $d \leq n - t - 1$
- *Tarannikov's bound*: $nl \leq 2^{n-1} - 2^{t+1}$

Constructions of good Boolean Functions

- Number of Boolean functions of n variables: 2^{2^n}
- $\Rightarrow$ too huge for exhaustive search when $n > 5$!
- Functions used in the combiner model have $n \geq 13$ variables

In practice, one usually resorts to:

- Algebraic constructions [1]
  - *Maiorana-McFarland construction*
  - *Rothaus' construction*
  - ...
- Heuristic techniques
  - *Simulated Annealing* [3]
  - *Evolutionary Algorithms* [5]
  - ...

Evolutionary Search of Boolean Functions

- Evolutionary search offers a promising way to optimize cryptographic boolean functions
- Usual approach: directly search the space of truth tables, represented as 2^n-bit strings [6]
- Fitness function measuring nonlinearity, algebraic degree, and deviation from correlation-immunity
- Specialized variation operators for preserving balancedness

Spectral Inversion [2]

- Applying the Inverse Walsh Transform to a generic spectrum yields a pseudoboolean function $f : \mathbb{F}_2^n \to \mathbb{R}$

$$\mathcal{S}_f = (0, -4, -2, 2, 2, 4, 4, -2)$$

$$\Downarrow \hat{F}^{-1}$$

$$\Omega_{\hat{f}} = (0, 0, 0, -1, 0, -1, 2)$$

- New objective: minimize the deviation of Walsh spectra which satisfy the desired cryptographic constraints
- Heuristic techniques proposed for this optimization problem:
  - Clark et al. [2]: Simulated Annealing (SA)
  - Mariot and Leporati [5]: Genetic Algorithms (GA)

Plateaued Functions [8]

- Our GA evolves spectra of plateaued functions
- A (pseudo)boolean function f is plateaued if its Walsh spectrum takes only three values: $-W_M(f)$, 0 and $+W_M(f)$

$$\mathcal{S}_f = (0, 0, 0, 0, -4, 4, 4, 4) \Rightarrow \text{plateaued}$$

- Motivations:
  - Simple combinatorial representation of candidate solutions, determined by a single parameter $r \geq n/2$
  - Plateaued functions reach both Siegenthaler's and Tarannikov's bounds

Chromosome Encoding

- Resiliency Constraint: ignore positions with at most m ones

x	<u>000</u>	<u>100</u>	<u>010</u>	110	<u>001</u>	101	011	111
$\mathcal{S}_f$	0	0	0	−4	0	4	4	4

- The chromosome c is the permutation of the spectrum in the positions with more than m ones:

x	110	101	011	111
c	−4	4	4	4

- The multiplicities of 0, $-W_M(f)$ and $+W_M(f)$ in the permutation depend on plateau index r

Fitness Function

- Given $\hat{f} : \mathbb{F}_2^n \to \mathbb{R}$, the nearest boolean function $\hat{b} : \mathbb{F}_2^n \to \mathbb{F}_2$ is defined for all $x \in \mathbb{F}_2^n$ as:

$$\hat{b}(x) = \begin{cases} +1 & \text{, if } \hat{f}(x) > 0 \\ -1 & \text{, if } \hat{f}(x) < 0 \\ +1 \text{ or } -1 \text{ (chosen randomly)} & \text{, if } \hat{f}(x) = 0 \end{cases}$$

- Objective function proposed in [2]:

$$obj(f) = \sum_{x \in \mathbb{F}_2^n} (\hat{f}(x) - \hat{b}(x))^2$$

- Fitness function maximised by our GA: $fit(f) = -obj(f)$

Genetic Operators

- Crossover between two Walsh spectra p_1, p_2 must preserve the multiplicities of $-W_M(f)$, 0 and $+W_M(f)$
- Idea: use counters to keep track of the multiplicities [6]
- Mutation: swap two random positions in the chromosome with different values
- Selection operators adopted:
  - Roulette-Wheel (*RWS*)
  - Deterministic Tournament (*DTS*)

Experimental Settings

Common parameters:

- Number of variables $n = 6,7$ and plateau index $r = 4$

(n,m,d,nl)	$\|0_{res}\|$	$\|0_{add}\|$	$\|-W_M(f)\|$	$\|+W_M(f)\|$
$(6,2,3,24)$	22	26	6	10
$(7,2,4,56)$	29	35	28	36

GA-related parameters:

- Population size $N = 30$
- max generations $G = 500000$
- GA runs $R = 500$
- Crossover probability $p_\chi = 0.95$
- Mutation probability $p_\mu = 0.05$
- Tournament size $k = 3$

SA-related parameters:

- Inner loops $MaxIL = 3000$
- Moves in loop $MIL = 5000$
- SA runs $R = 500$
- Initial temperatures $T = 100, 1000$
- Cooling parameter: $\alpha = 0.95, 0.99$

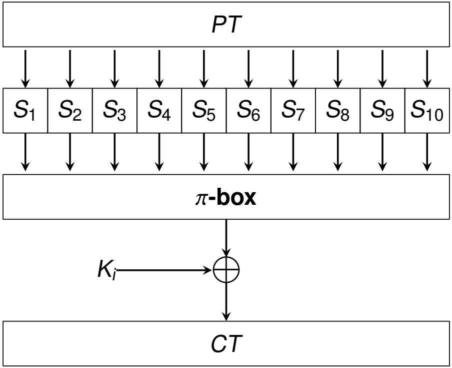
Results

Statistics of the best solutions found by our GA and SA over $R = 500$ runs.

n	Stat	$GA(RWS)$	$GA(DTS)$	$SA(T_1, \alpha_1)$	$SA(T_2, \alpha_2)$
6	avg_o	14.08	13.02	19.01	19.03
	min_o	0	0	0	0
	max_o	16	16	28	28
	std_o	5.21	6.23	4.89	4.81
	$\#opt$	$\boxed{60}$	$\boxed{93}$	$\boxed{11}$	$\boxed{10}$
	avg_t	83.3	79.2	79.1	79.4
7	avg_o	53.44	52.6	45.09	44.85
	min_o	47	44	32	27
	max_o	58	59	63	57
	std_o	2.40	2.77	4.39	4.18
	$\#opt$	0	0	0	0
	avg_t	204.2	204.5	180.3	180.2

Block Ciphers: Substitution-Permutation Network

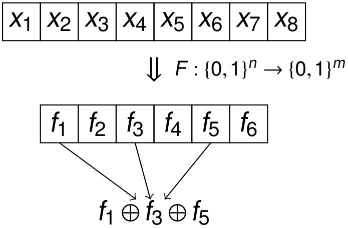
Round function of a SPN cipher:



- $S_i : \mathbb{F}_2^n \to \mathbb{F}_2^n$ are S-boxes providing confusion [7]
- Security of confusion layer $\Leftrightarrow$ cryptographic properties of S_i

Background on S-boxes (1/2)

- A Substitution Box (S-box) is a mapping $F : \{0,1\}^n \rightarrow \{0,1\}^m$ defined by m coordinate functions $f_i : \{0,1\}^n \rightarrow \{0,1\}$
- The component functions $v \cdot F : \{0,1\}^n \rightarrow \{0,1\}$ for $v \in \{0,1\}^m$ of F are the linear combinations of the f_i



- The nonlinearity of a S-box F is defined as the minimum nonlinearity among all its component functions
- S-boxes with high nonlinearity allow to resist to linear cryptanalysis attacks

Background on S-Boxes (2/2)

- delta difference table of F wrt a, b:

$$D_F(a,b) = \left\{x \in \mathbb{F}_2^n : F(x) \oplus F(x \oplus a) = b\right\}.$$

- Given $\delta_F(a,b) = |D_F(a,b)|$, the differential uniformity of F is:

$$\delta_F = \max_{\substack{a \in \{0,1\}^{n*} \\ b \in \{0,1\}^m}} \delta_F(a,b).$$

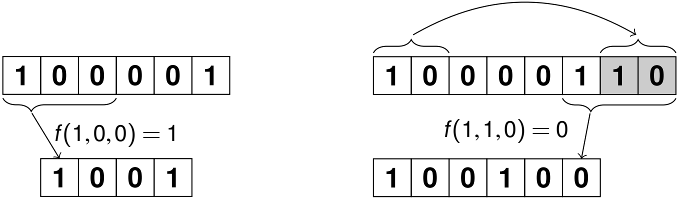
- S-boxes with low differential uniformity are able to resist differential cryptanalysis attacks

Cellular Automata S-boxes

- One-dimensional Cellular Automaton (CA): a discrete parallel computation model composed of a finite array of n cells
- Each cell updates its state $s \in \{0,1\}$ by applying a local rule $f : \{0,1\}^d \to \{0,1\}$ to itself and the $d-1$ cells to its right

Example: $n = 6$, $d = 3$, $f(s_i, s_{i+1}, s_{i+2}) = s_i \oplus s_{i+1} \oplus s_{i+2}$,

Truth table: $\Omega(f) = 01101001 \to$ Rule 150



No Boundary CA – NBCA

Periodic Boundary CA – PBCA

Problem Statement

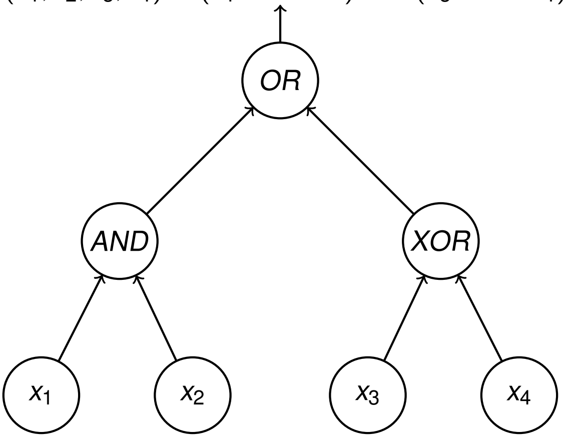
- Goal: Find PBCA of length n and diameter $d = n$ having cryptographic properties equal to or better than those of other real-world S-boxes
- Considered S-boxes sizes: from $n = 4$ to $n = 8$
- Using tree encoding, exhaustive search is already unfeasible for $n = 4$
- We adopted Genetic Programming to address this problem

Genetic Programming (GP)

- Optimization method inspired by evolutionary principles, introduced by Koza [4]
- Each candidate solution (individual) is represented by a tree
  - Terminal nodes: input variables
  - Internal nodes: Boolean operators (AND, OR, NOT, XOR, ...)
- New solutions are created through genetic operators like tree crossover and subtree mutation applied to a population of candidate solutions
- Optimization is performed by evaluating the new candidate solutions wrt a fitness function

GP Tree Encoding – Example

$$f(x_1, x_2, x_3, x_4) = (x_1\ AND\ x2)\ OR\ (x_3\ XOR\ x_4)$$



Fitness Function

- Considered cryptographic properties:
  - balancedness/invertibility ($BAL = 0$ if F is balanced, -1 otherwise)
  - nonlinearity N_F
  - differential uniformity δ_F
- Fitness function maximized:

$$fitness = BAL + \Delta_{BAL,0}\left(N_F + \left(1 - \frac{nMinN_F}{2^n}\right) + (2^n - \delta_F)\right).$$

where $\Delta_{BAL,0} = 1$ if F is balanced and 0 otherwise, and $nMinN_F$ is the number of occurrences of the current value of nonlinearity

Experimental Setup

- Problem instance / CA size: $n = 4$ up to $n = 8$
- Maximum tree depth: equal to n
- Genetic operators: simple tree crossover, subtree mutation
- Population size: 2000
- Stopping criterion: 2000000 fitness evaluations
- Parameters determined by initial tuning phase on $n = 6$ case

Results

Table: Statistical results and comparison.

S-box size	T_max	GP			N_F	δ_F
		Max	Avg	Std dev		
4×4	16	**16**	16	0	4	4
5×5	42	**42**	41.73	1.01	12	2
6×6	86	**84**	80.47	4.72	24	4
7×7	182	**182**	155.07	8.86	56	2
8×8	364	318	281.87	13.86	82	20

- From $n = 4$ to $n = 7$, we obtained CA rules inducing S-boxes with optimal crypto properties
- Only for $n = 8$ the performances of GP are consistently worse wrt to the theoretical optimum

Conclusions

- Boolean functions and S-boxes play a fundamental role in the design of symmetric ciphers
- The design of Boolean functions and S-boxes with good properties is a hard optimization problem
- For Boolean functions, GA are more efficient than SA under the spectral inversion approach
- For S-boxes, GP is able to find optimal solutions up to size 7×7

References I

C. Carlet.
Boolean Functions for Cryptography and Error Correcting Codes.
In Y. Crama and P. L. Hammer, editors, *Boolean Models and Methods in Mathematics, Computer Science, and Engineering*, pages 257–397. Cambridge University Press, 2010.

J. A. Clark, J. L. Jacob, S. Maitra, and P. Stanica.
Almost boolean functions: The design of boolean functions by spectral inversion.
Computational Intelligence, 20(3):450–462, 2004.

J. A. Clark, J. L. Jacob, S. Stepney, S. Maitra, and W. Millan.
Evolving boolean functions satisfying multiple criteria.
In *Progress in Cryptology - INDOCRYPT 2002, Third International Conference on Cryptology in India, Hyderabad, India, December 16-18, 2002*, pages 246–259, 2002.

J. R. Koza.
Genetic programming - on the programming of computers by means of natural selection.
Complex adaptive systems. MIT Press, 1993.

L. Mariot and A. Leporati.
A genetic algorithm for evolving plateaued cryptographic boolean functions.
In *Theory and Practice of Natural Computing - Fourth International Conference, TPNC 2015, Mieres, Spain, December 15-16, 2015. Proceedings*, pages 33–45, 2015.

References II

W. Millan, A. J. Clark, and E. Dawson.
Heuristic design of cryptographically strong balanced boolean functions.
In *Advances in Cryptology - EUROCRYPT '98, International Conference on the Theory and Application of Cryptographic Techniques, Espoo, Finland, May 31 - June 4, 1998, Proceeding*, pages 489–499, 1998.

C. E. Shannon.
Communication theory of secrecy systems.
Bell system technical journal, 28(4):656–715, 1949.

Y. Zheng and X. Zhang.
Plateaued functions.
In *Information and Communication Security, Second International Conference, ICICS'99, Sydney, Australia, November 9-11, 1999, Proceedings*, pages 284–300, 1999.