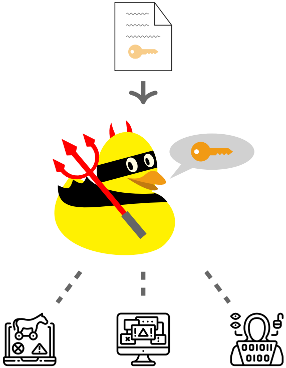
Analysis and Improvement of Differential Computation Attacks against Internally-Encoded White-Box Implementations

Matthieu Rivain [1] **Junwei Wang** [1,2,3]

[1]CryptoExperts [2]University of Luxembourg [3]University Paris 8

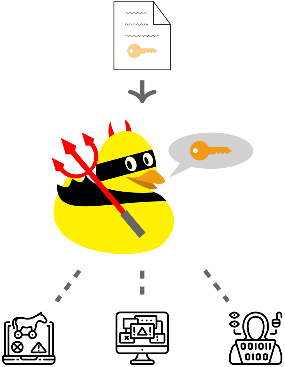
CHES 2019, Atalanta

White-Box Threat Model



- **Goal:** to extract a cryptographic key, $\cdots$
- **Where:** from a software impl. of cipher
- **Who:** malwares, co-hosted applications, user themselves, $\cdots$
- **How:** (*by all kinds of means*)
  - analyze the code
  - spy on the memory
  - interfere the execution
  - $\cdots$

White-Box Threat Model



- **Goal:** to extract a cryptographic key, $\cdots$
- **Where:** from a software impl. of cipher
- **Who:** malwares, co-hosted applications, user themselves, $\cdots$
- **How:** (*by all kinds of means*)
  - analyze the code
  - spy on the memory
  - interfere the execution
  - $\cdots$

In theory: no provably secure white-box scheme for standard block ciphers.

Typical Applications

Digital Content Distribution

videos, music, games, e-books, ⋯

Host Card Emulation

mobile payment without a *secure element*

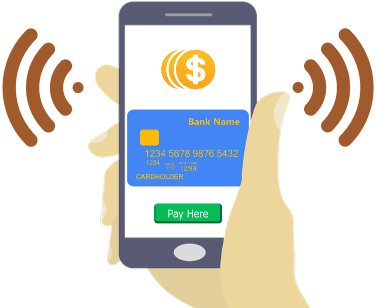
Typical Applications

Digital Content Distribution

videos, music, games, e-books, $\cdots$

Host Card Emulation

mobile payment without a *secure element*



In practice: heuristic solutions / security through obscurity

Internal Encoding Countermeasure [SAC02]

$$\boxed{X} \rightarrow \boxed{\mathcal{R}_1} \rightarrow \boxed{\mathcal{R}_2} \rightarrow \cdots \rightarrow \boxed{\mathcal{R}_r} \rightarrow \boxed{Y}$$

1. Represent the cipher into a *network* of transformations

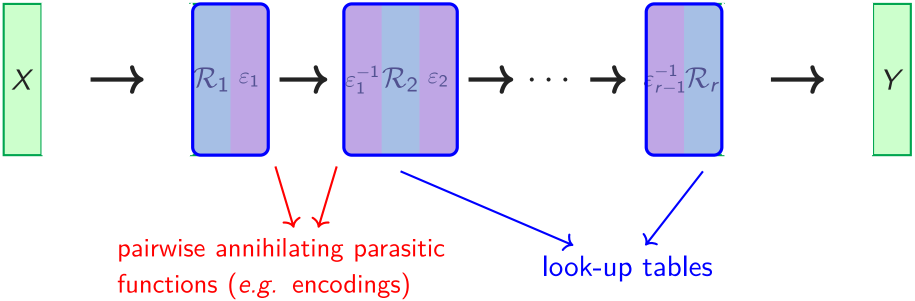
Internal Encoding Countermeasure [SAC02]

$$X \longrightarrow \mathcal{R}_1 \ \varepsilon_1 \longrightarrow \varepsilon_1^{-1}\mathcal{R}_2 \ \varepsilon_2 \longrightarrow \cdots \longrightarrow \varepsilon_{r-1}^{-1}\mathcal{R}_r \longrightarrow Y$$

pairwise annihilating parasitic functions (*e.g.* encodings)

1. Represent the cipher into a *network* of transformations
2. Obfuscate the network by encoding *adjacent* transformations

Internal Encoding Countermeasure [SAC02]

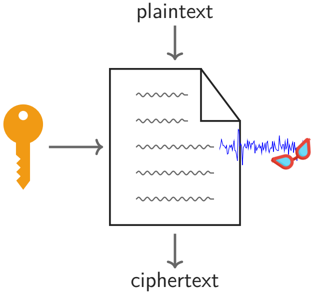


1. Represent the cipher into a *network* of transformations
2. Obfuscate the network by encoding *adjacent* transformations
3. Store the encoded transformations into look-up tables

Attacks in This Talk

1. Differential Computation Analysis
2. Collision Attack

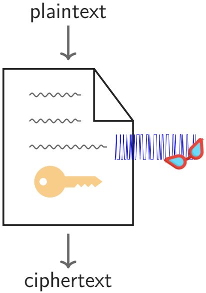
Differential Computation Analysis [CHES16]



gray-box model

side-channel leakages (*noisy*)

e.g. power/EM/time/···



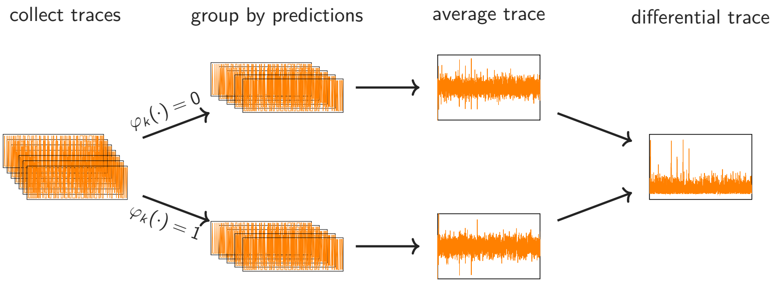
white-box model

computational leakage (*perfect*)

e.g. registers/accessed memory/···

Differential Computation Analysis [CHES16]

Differential power analysis techniques on computational leakages

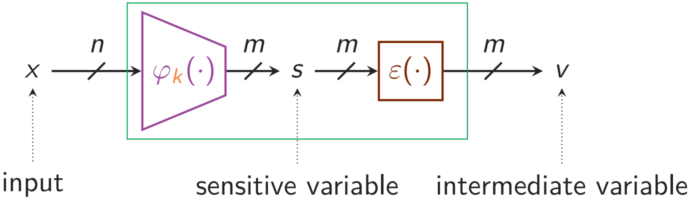


Implying strong *linear correlation* between the sensitive variables and the leaked samples in the computational traces.

DCA Attack Limitations

1. The seminal work [CHES16] lacks in-depth understanding of DCA
2. The follow-up analysis [ACNS18] is
   - *partly* experimental (in particular for wrong key guesses)
   - *Only* known to work on **nibble** encodings
   - *Only* known to work on the **first** and **last** rounds
   - Success probability is unknown
3. The computational traces are only sub-optimally exploited

Internal Encoding Leakage



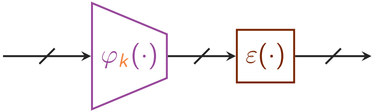
- A *key-dependent* (n, m) selection function φ_k in a block cipher
- A *random* selected m-bit bijection ε
- $\varepsilon \circ \varphi_k$, as a result of some **table look-ups**, is **leaked in the memory**

- To exploit the leakage of $\varepsilon \circ \varphi_k$, it is necessary that $n > m$

DCA Analysis

Based on well-established theory – *Boolean correlation*, instead of *difference of means*: for any key guess k

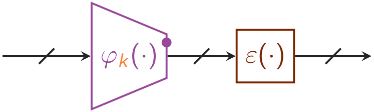
$$\rho_k = \mathrm{Cor}\Big(\qquad\qquad , \qquad\qquad\qquad \Big)$$



DCA Analysis

Based on well-established theory – *Boolean correlation*, instead of *difference of means*: for any key guess k

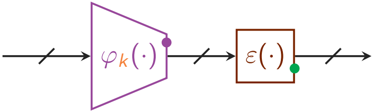
$$\rho_k = \mathrm{Cor}\Big(\varphi_k(\cdot)[i] \quad , \qquad\qquad\qquad \Big)$$



DCA Analysis

Based on well-established theory – *Boolean correlation*, instead of *difference of means*: for any key guess k

$$\rho_k = \mathrm{Cor}\Big(\varphi_k(\cdot)[i] \quad , \quad \varepsilon \circ \varphi_{k^*}(\cdot)[j] \Big)$$

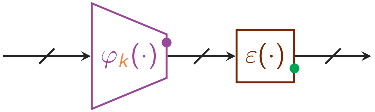


CryptoExperts

DCA Analysis

Based on well-established theory – *Boolean correlation*, instead of *difference of means*: for any key guess k

$$\rho_k = \mathrm{Cor}\Big(\ \varphi_k(\cdot)[i]\ \ ,\ \ \varepsilon \circ \varphi_{k^*}(\cdot)[j]\ \Big)$$



DCA success (roughly) requires:

$$\left|\rho_{k^*}\right| > \max_{k^\times} \left|\rho_{k^\times}\right|$$

ρ_{k^*} and $\rho_{k^\times}$: Distributions

- **Ideal** assumption: $\left(\varphi_k\right)_k$ are mutually independent random (n, m) functions

ρ_{k^*} and $\rho_{k^\times}$: Distributions

- **Ideal** assumption: $\left(\varphi_k\right)_k$ are mutually independent random (n, m) functions

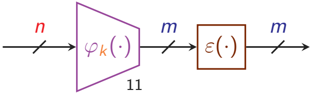
Correct key guess k^*,

$$\rho_{k^*} = 2^{2-m} N^* - 1$$

where

$$N^* \sim \mathcal{HG}(2^m, 2^{m-1}, 2^{m-1}) \; .$$

Only depends on m.



ρ_{k^*} and $\rho_{k^\times}$: Distributions

- **Ideal** assumption: $\left(\varphi_k\right)_k$ are mutually independent random (n, m) functions

Correct key guess k^*,

$$\rho_{k^*} = 2^{2-m} N^* - 1$$

where

$$N^* \sim \mathcal{HG}(2^m, 2^{m-1}, 2^{m-1}) \ .$$

Only depends on m.

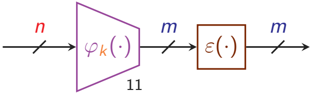
Incorrect key guess $k^\times$,

$$\rho_{k^\times} = 2^{2-n} N^\times - 1$$

where

$$N^\times \sim \mathcal{HG}(2^n, 2^{n-1}, 2^{n-1}) \ .$$

Only depends on n.



Lemma

Lemma

Let $\mathcal{B}(n)$ be the set of balanced n-bit Boolean function. If $f \in \mathcal{B}(n)$ and $g \xleftarrow{\$} \mathcal{B}(n)$ independent of f, then the balanceness of $f + g$ is $\mathrm{B}(f + g) = 4 \cdot N - 2^n$ where $N \sim \mathcal{HG}(2^n, 2^{n-1}, 2^{n-1})$ denotes the size of $\{x : f(x) = g(x) = 0\}$.

With

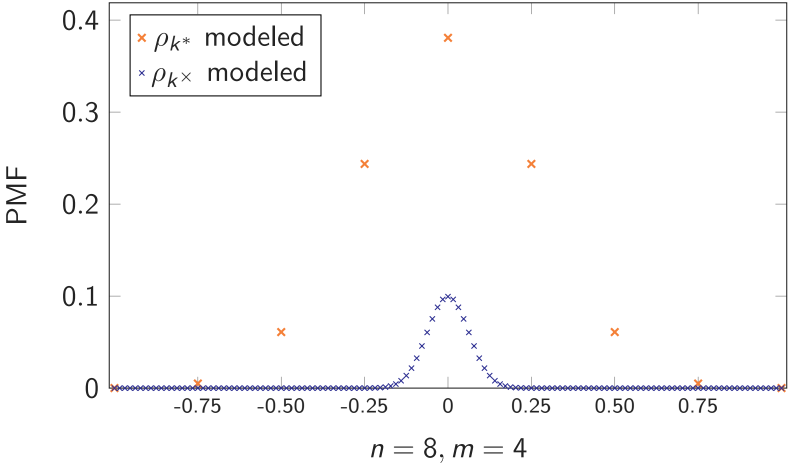
$$\mathrm{Cor}(f + g) = \frac{1}{2^n}\mathrm{B}(f + g)$$

$\Rightarrow$

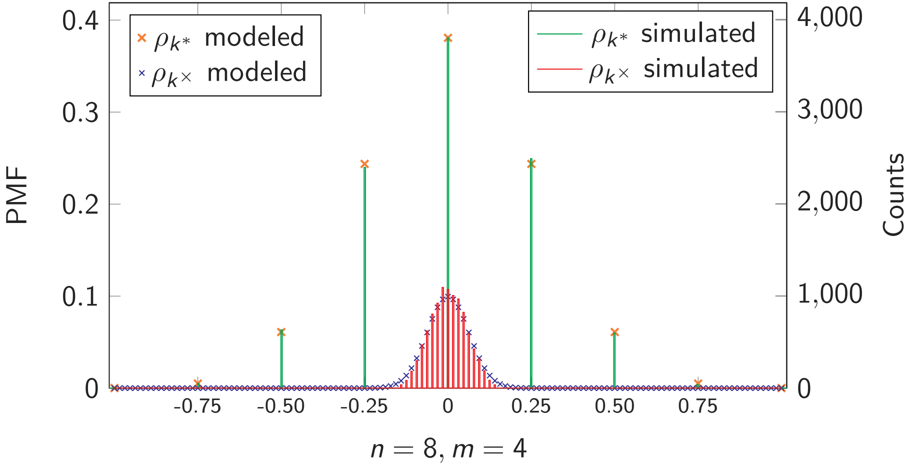
$$\rho_{k^*} = 2^{2-m}N^* - 1 \quad \text{and} \quad \rho_{k^\times} = 2^{2-n}N^\times - 1$$

where $N^* \sim \mathcal{HG}(2^m, 2^{m-1}, 2^{m-1})$ and $N^\times \sim \mathcal{HG}(2^n, 2^{n-1}, 2^{n-1})$.

ρ_{k^*} and $\rho_{k^\times}$: Distributions

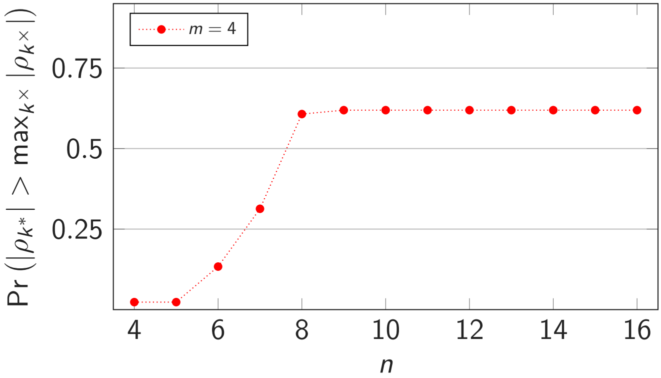


ρ_{k^*} and $\rho_{k^\times}$: Distributions



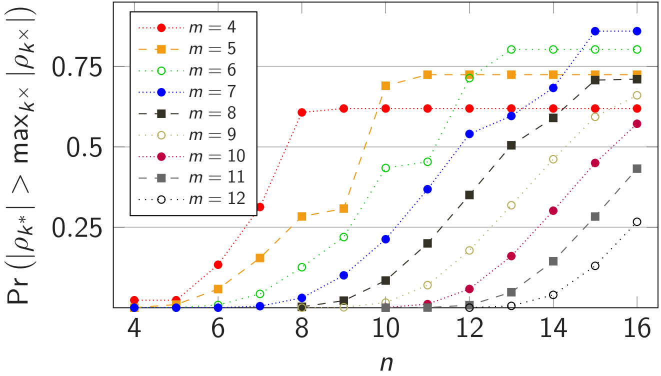
CryptoExperts

DCA Success Rate: $|\rho_{k^*}| > \max_{k^\times} |\rho_{k^\times}|$



DCA success probability converges towards $\approx 1 - \Pr_{N^*}\left(2^{m-2}\right)$ for $n \geq 2m + 2$.

DCA Success Rate: $|\rho_{k^*}| > \max_{k^\times} |\rho_{k^\times}|$



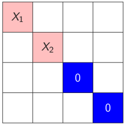
DCA success probability converges towards $\approx 1 - \Pr_{N^*}\left(2^{m-2}\right)$ for $n \geq 2m + 2$.

Attack a NSC Variant: a White-Box AES

- *Byte* encoding protected
- DCA has failed to break it *before this work*

Attack a NSC Variant: a White-Box AES

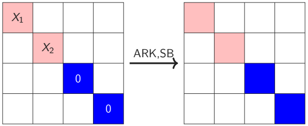
- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



x_1 x_2

Attack a NSC Variant: a White-Box AES

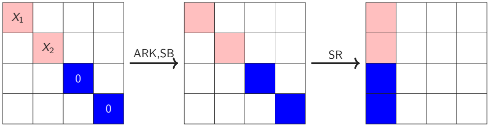
- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



$\mathbf{Sbox}(x_1 \oplus k_1)$ $\mathbf{Sbox}(x_2 \oplus k_2)$ $\mathbf{Sbox}(k_3)$ $\mathbf{Sbox}(k_4)$

Attack a NSC Variant: a White-Box AES

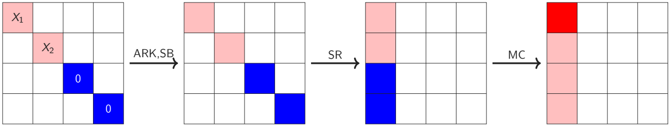
- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



$\mathbf{Sbox}(x_1 \oplus k_1)$ $\mathbf{Sbox}(x_2 \oplus k_2)$ $\mathbf{Sbox}(k_3)$ $\mathbf{Sbox}(k_4)$

Attack a NSC Variant: a White-Box AES

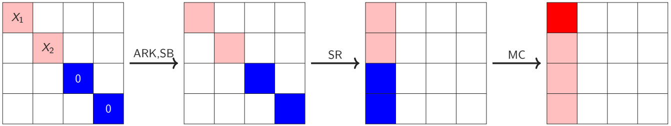
- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



$$2 \cdot \textbf{Sbox}(x_1 \oplus k_1) \ \oplus\ 3 \cdot \textbf{Sbox}(x_2 \oplus k_2) \ \oplus\ \textbf{Sbox}(k_3) \ \oplus\ \textbf{Sbox}(k_4)$$

Attack a NSC Variant: a White-Box AES

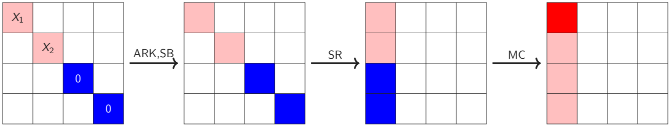
- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



$$2 \cdot \textbf{Sbox}(x_1 \oplus k_1) \oplus 3 \cdot \textbf{Sbox}(x_2 \oplus k_2) \oplus \qquad c$$

Attack a NSC Variant: a White-Box AES

- *Byte* encoding protected
- DCA has failed to break it *before this work*
- Our approach: target a output byte of MixColumn in the first round



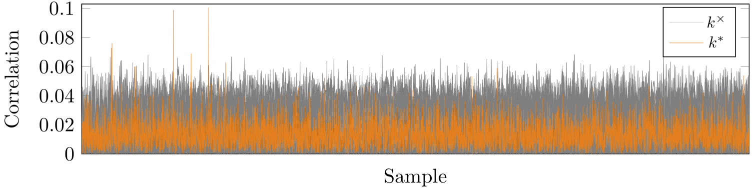
$$\varphi_{k_1||k_2}(x_1||x_2) = 2 \cdot \mathbf{Sbox}(x_1 \oplus k_1) \oplus 3 \cdot \mathbf{Sbox}(x_2 \oplus k_2)$$

$$\varepsilon' = \varepsilon \circ \oplus_c \ ,$$

$$n = 16, m = 8 \ , |\mathcal{K}| = 2^{16}.$$

Attack a NSC Variant: a White-Box AES

- Attack results: $\sim$ 1800 traces



- Similar attack can be applied to a "masked" white-box implementation, which intends to resist DCA.

Attacks in This Talk

1. Differential Computation Analysis
2. Collision Attack

Collision Attack

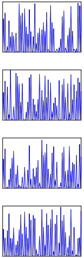
N inputs & raw traces

x_1

x_2

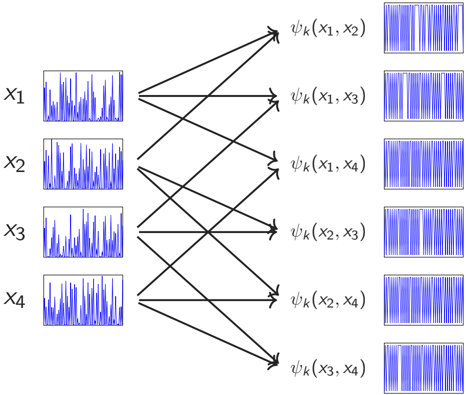
x_3

x_4



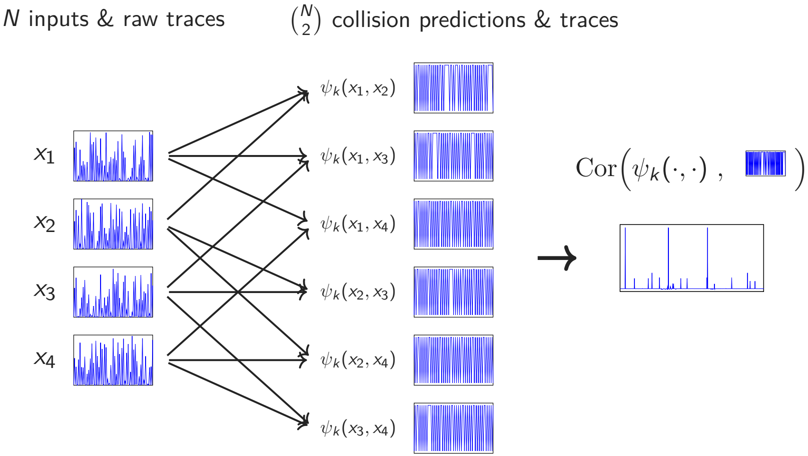
Collision Attack

N inputs & raw traces $\binom{N}{2}$ collision predictions & traces



$$\psi_k(x_1, x_2) := \Big(\varphi_k(x_1) = \varphi_k(x_2)\Big)$$

Collision Attack



$$\psi_k(x_1, x_2) := \big(\varphi_k(x_1) = \varphi_k(x_2)\big)$$

Collision Attack: Explanation

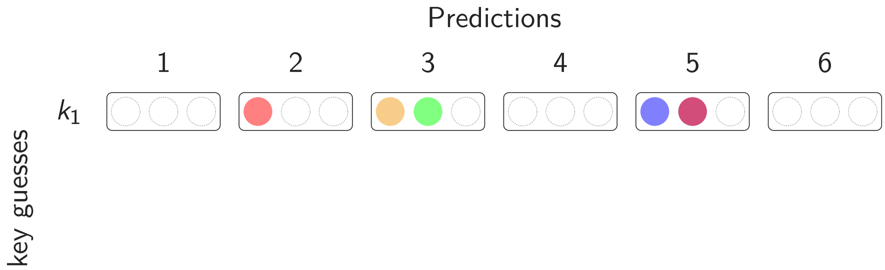
Based on the principle:

$$\varphi_k(x_1) = \varphi_k(x_2) \quad \Leftrightarrow \quad \varepsilon \circ \varphi_k(x_1) = \varepsilon \circ \varphi_k(x_2)$$

Trace Complexity:

$$N = O\left(2^{\frac{m}{2}}\right)$$

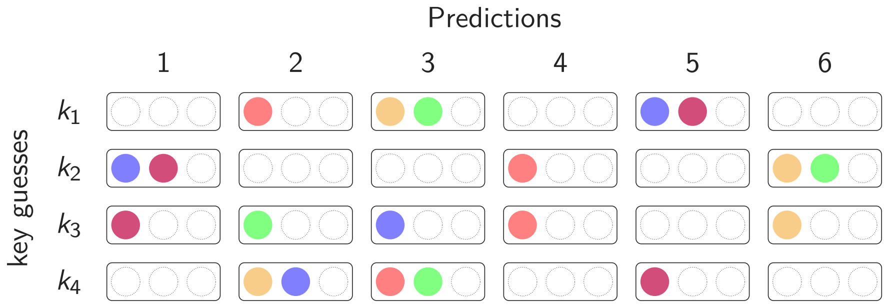
Collision Attack: Explanation



k^* "**collides**" $\bigwedge$ $\forall k^\times$, k^* and $k^\times$ are not "**isomorphic**"

$$\Rightarrow N = O\left(2^{\frac{m}{2}}\right)$$

Collision Attack: Explanation

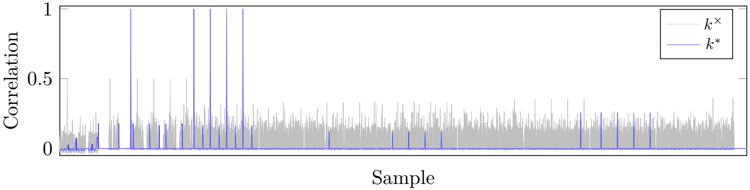


k^* "**collides**" $\bigwedge$ $\forall k^\times$, k^* and $k^\times$ are not "**isomorphic**"

$$\Rightarrow N = O\left(2^{\frac{m}{2}}\right)$$

Attack the NSC Variant

- Same to DCA: targeting at one 1-st round MixColumn output byte
- Attack results: 60 traces



Conclusion

- DCA against internal encodings has been analysed in depth
  - Allows to attack wider encodings
- Computation traces have been further exploited
  - Showcase to attack variables beyond the first round of the cipher
  - New class of collision attack with very low trace complexity
- Hence, protecting AES with internal encodings in the beginning rounds is insufficient

Thank You !

ia.cr/2019/076