TELECOM ParisTech

Institut Mines-Télécom

Good Is Not Good Enough

Deriving Optimal Distinguishers from Communication Theory

Annelie Heuser, Olivier Rioul, Sylvain Guilley

CHES 2014

Motivation

Given a side-channel context

simulations (SNR/leakage model)

measurements

knowledge of the attacker

- Questions raised by the community

What is the best distinguisher among all known ones?

Motivation

Given a side-channel context

simulations (SNR/leakage model)

measurements

knowledge of the attacker

- Questions raised by the community

What is the best distinguisher among all known ones?

- Question we would like to answer

What is the best distinguisher among all possible ones?

Outlook

- Side-channel ⟺ communication channel
- Optimal distinguisher
  - Known model
  - Known model on a proportional scale
  - Partially known model
- Empirical results
- What comes next!

SCA as a communication channel

$$\underbrace{\mathbf{X}}_{\text{leakage}} = \underbrace{\varphi}_{\text{device-specific function}}(\underbrace{f}_{\text{algorithmic-specific function}}(\underbrace{\mathbf{T}}_{\text{input/output}}, \underbrace{K^\star}_{\text{secret key}})) + \underbrace{\mathbf{N}}_{\text{noise}}$$

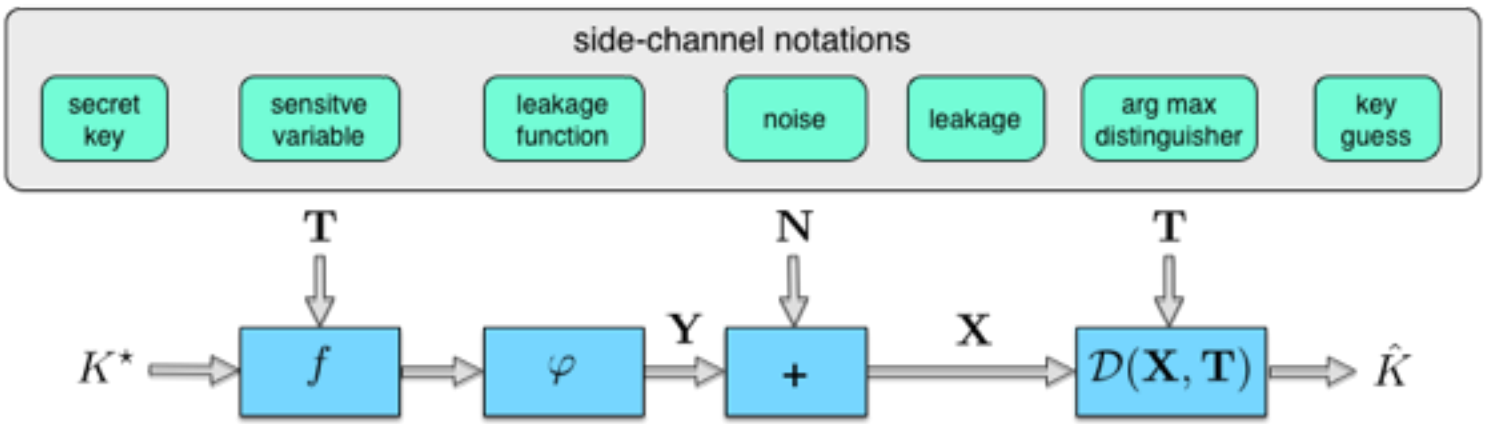
Annelie Heuser, Olivier Rioul, Sylvain Guilley

SCA as a communication channel

leakage input/output secret key noise

$$\mathbf{X} = \varphi(f(\mathbf{T}, K^{\star})) + \mathbf{N}$$

device-specific function algorithmic-specific function

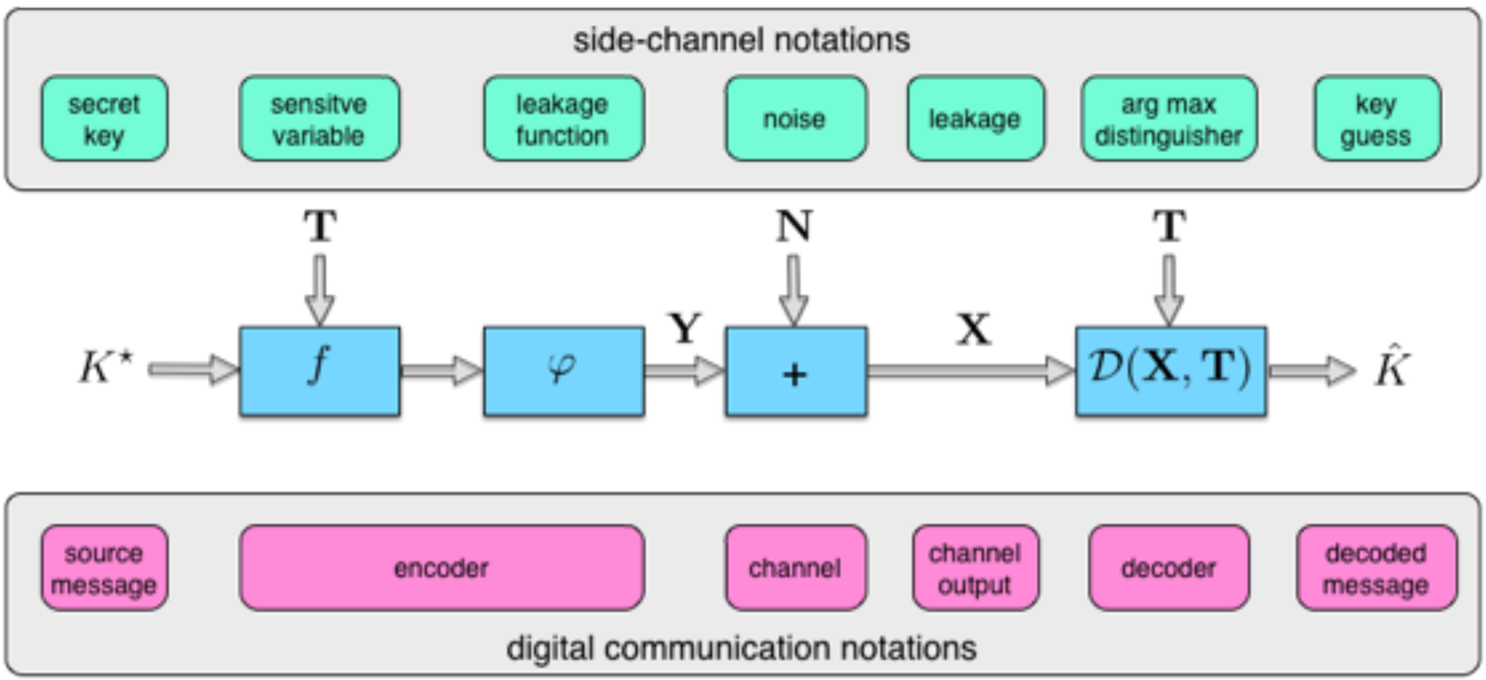


SCA as a communication channel

leakage — input/output — secret key — noise

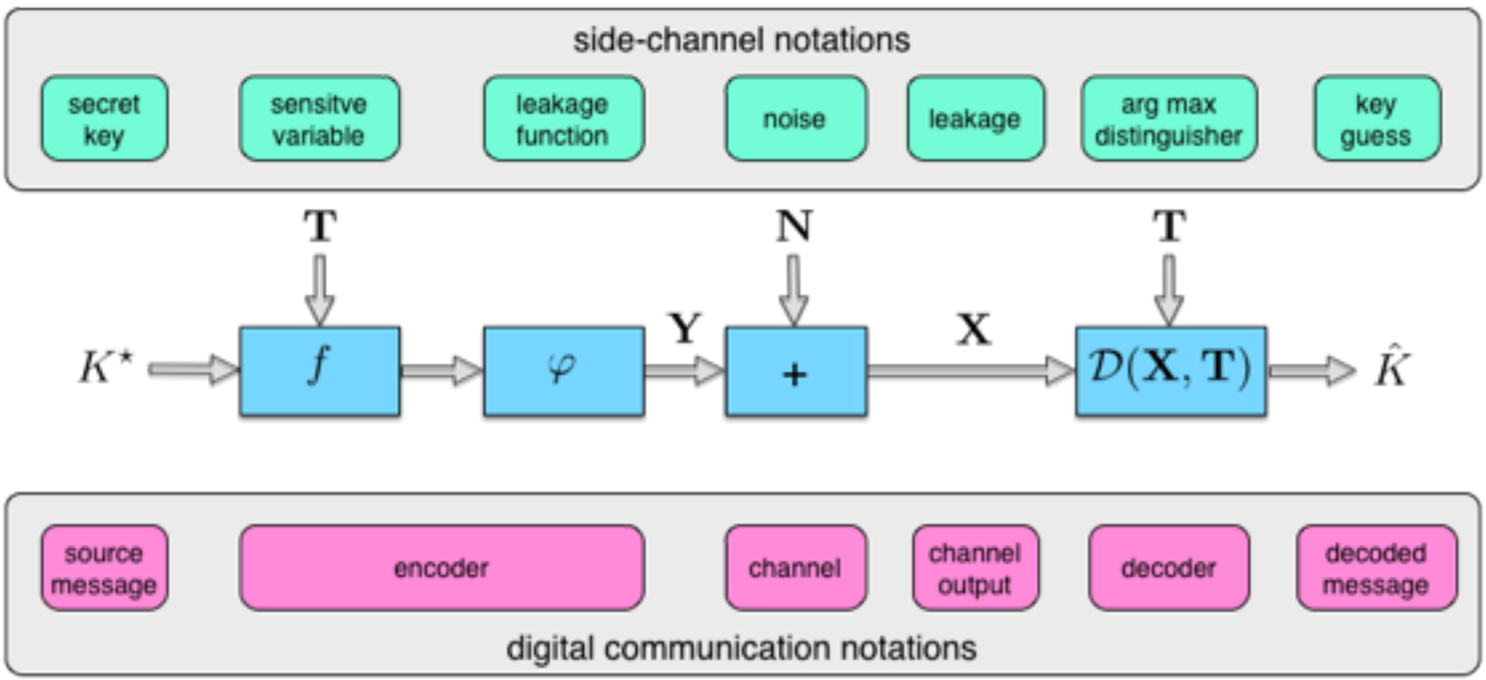
$$\mathbf{X} = \varphi(f(\mathbf{T}, K^{\star})) + \mathbf{N}$$

device-specific function — algorithmic-specific function



SCA as a communication channel

- secret key is fixed but unknown
- communication theory: modeled as random
- practice: equal for all messages



Optimal distinguishing rule

- Minimize the probability of error

$$\mathbb{P}_e = \mathbb{P}\{\hat{K} \neq K^\star\}$$

Theorem (Optimal distinguishing rule) *The optimal distinguishing rule is given by the* maximum a posteriori probability (MAP) *rule*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \Big(\mathbb{P}\{k^\star\} \cdot p(\mathbf{x}|\mathbf{t}, k^\star)\Big) \quad .$$

If the keys are assumed equiprobable, i.e. $\mathbb{P}\{k\} = 2^{-n}$, *the equation reduces to the* maximum likelihood distinguishing rule

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} p(\mathbf{x}|\mathbf{t}, k^\star) \quad .$$

Proof given in the paper!

Optimal distinguishing rule

- Minimize the probability of error

$$\mathbb{P}_e = \mathbb{P}\{\hat{K} \neq K^\star\}$$

Theorem (Optimal distinguishing rule) *The optimal distinguishing rule is given by the* maximum a posteriori probability (MAP) *rule*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \left(\mathbb{P}\{k^\star\} \cdot p(\mathbf{x}|\mathbf{t}, k^\star) \right) \ .$$

If the keys are assumed equiprobable, i.e. $\mathbb{P}\{k\} = 2^{-n}$, *the equation reduces to the* maximum likelihood distinguishing rule

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} p(\mathbf{x}|\mathbf{t}, k^\star) \ .$$

Template attack [Chari+2002]

Proof given in the paper!

Optimal attack when the model is known

knows

$$\mathbf{X} = \varphi(f(\mathbf{T}, K^\star)) + \mathbf{N}$$

Proposition (Maximum likelihood) *When f and φ are known to the attacker such that $\boxed{\mathbf{Y}(K^\star) = \varphi(f(\mathbf{T}, K^\star))}$, then the optimal decision becomes*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \Big(\mathbb{P}\{k^\star\} \cdot p(\mathbf{x}|\mathbf{y}(k^\star)) \Big) \ ,$$

and for equiprobable keys this reduces to

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \ p(\mathbf{x}|\mathbf{y}(k^\star)) \ .$$

Proof given in the paper!

Annelie Heuser, Olivier Rioul, Sylvain Guilley

Optimal attack when the model is known

knows

$$\mathbf{X} = \varphi(f(\mathbf{T}, K^\star)) + \mathbf{N}$$

Proposition (Maximum likelihood) *When f and φ are known to the attacker such that $\boxed{\mathbf{Y}(K^\star) = \varphi(f(\mathbf{T}, K^\star))}$, then the optimal decision becomes*

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \Big(\mathbb{P}\{k^\star\} \cdot \boxed{p(\mathbf{x}|\mathbf{y}(k^\star))} \Big) \quad ,$$

and for equiprobable keys this reduces to

$$\mathcal{D}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \boxed{p(\mathbf{x}|\mathbf{y}(k^\star))} \; .$$

Proof given in the paper!

Optimal Attack when the model is known

- Additive and i.i.d. noise

Proposition *When the leakage arises from* $\mathbf{X} = \mathbf{Y}(K^\star) + \mathbf{N}$, *then*

$$p(\mathbf{x}|\mathbf{y}(k^\star)) = p_{\mathbf{N}}(\mathbf{x} - \mathbf{y}(k^\star)) = \prod_{i=1}^{m} p_{N_i}(x_i - y_i(k^\star)) \quad .$$

This expression depends only on the noise probability distribution $p_{\mathbf{N}}$.

Proof given in the paper!

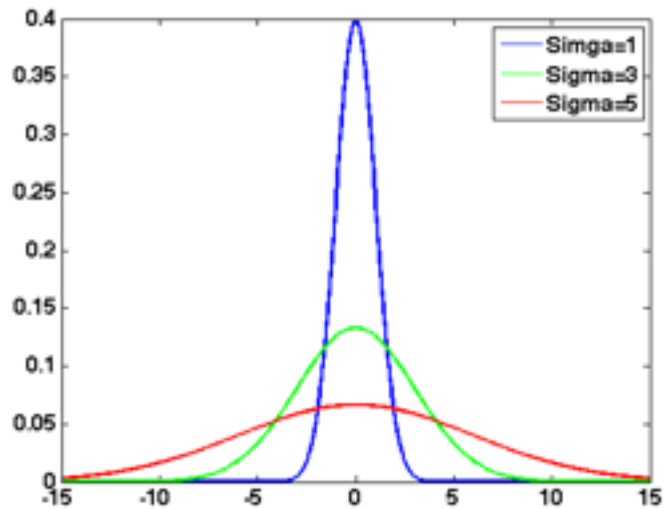
- Most publications considered Gaussian noise
- Furthermore investigate uniform and Laplacian noise

Gaussian noise distribution

Theorem (Optimal expression for Gaussian noise) *When the noise is zero mean Gaussian, $N \sim \mathcal{N}(0, \sigma^2)$, the optimal distinguishing rule is*

$$\mathcal{D}_{opt}^{M,G}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \langle \mathbf{x} | \mathbf{y}(k^\star) \rangle - \frac{1}{2} \|\mathbf{y}(k^\star)\|_2^2 \ .$$

Proof given in the paper!



Gaussian noise distribution

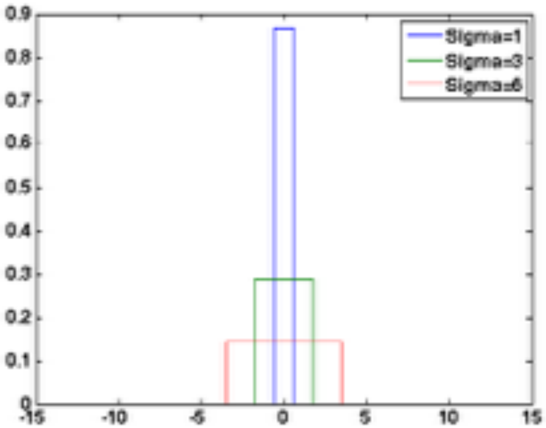
Theorem (Optimal expression for Gaussian noise) *When the noise is zero mean Gaussian, $N \sim \mathcal{N}(0, \sigma^2)$, the optimal distinguishing rule is*

$$\mathcal{D}_{opt}^{M,G}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \langle \mathbf{x} | \mathbf{y}(k^\star) \rangle - \boxed{\frac{1}{2} \|\mathbf{y}(k^\star)\|_2^2} .$$

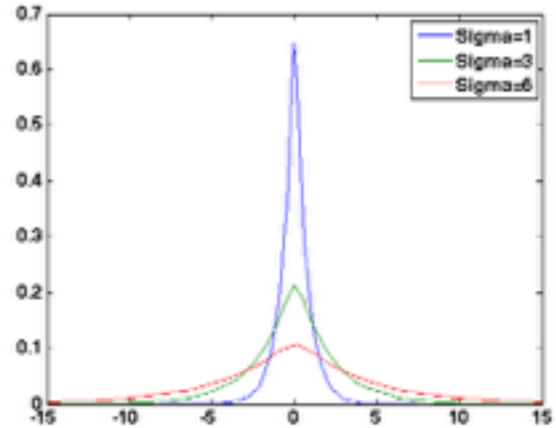
Proof given in the paper!

- For large number of measurements
  - the last term becomes key-independent but plays an important rule otherwise
  - the optimal distinguisher approximates to the covariance and the correlation
- But not with the absolute value!
- The optimal attack is independent on σ

Uniform and Laplacian noise



uniform



Laplacian

Definition (Noise distributions) *Let N be a zero-mean variable with variance σ^2 modeling the noise. Its distribution is:*

- *Uniform,* $N \sim \mathcal{U}(0,\sigma^2)$ *if* $p_N(n) = \begin{cases} \frac{1}{2\sigma\sqrt{3}} & \text{for } n \in [-\sqrt{3}\sigma, \sqrt{3}\sigma] \ , \\ 0 & \text{otherwise} \ . \end{cases}$
- *Laplacian,* $N \sim \mathcal{L}(0,\sigma^2)$ *if* $p_N(n) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{|n|}{\sigma/\sqrt{2}}}$.

Uniform and Laplacian noise

Theorem (Optimal expression for uniform and Laplacian noises) *When f and φ are known such that $Y(k) = \varphi(f(K^\star, T))$, and the leakage arises from $X = Y(K^\star) + N$ with $N \sim \mathcal{U}(0, \sigma^2)$ or $N \sim \mathcal{L}(0, \sigma^2)$, then the optimal distinguishing rule becomes*

- *Uniform noise distribution:* $\mathcal{D}_{opt}^{M,U}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \; -\|\mathbf{x} - \mathbf{y}(k^\star)\|_\infty$,
- *Laplace noise distribution:* $\mathcal{D}_{opt}^{M,L}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \; -\|\mathbf{x} - \mathbf{y}(k^\star)\|_1$.

Proof given in the paper!

- Novel distinguishing rules
- Cannot be approximated by correlation or covariance

Mono-bit leakage model

- W.l.o.g. $Y(K^\star) = \pm 1$
- Then $\|\mathbf{y}(k^\star)\|_2^2$ is equal to the number of measurements

$$\mathcal{D}^{M,G}_{opt(1\ \text{bit})}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} \langle \mathbf{x} | \mathbf{y}(k^\star) \rangle = \arg\max_{k^\star} \sum_{i | y_i(k^\star)=1} x_i - \sum_{i | y_i(k^\star)=-1} x_i$$

- Not equivalent to the difference-of-means test [Kocher+1999]

$$\mathcal{D}^{M,G}_{\text{KJJ}}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} \overline{\mathbf{x}_{+1}} - \overline{\mathbf{x}_{-1}}$$

- Nor to the t-test improvement [Coron+2000]

Model known on a proportional scale

- Model only known on a proportional scale

$$X = \boxed{a}Y(K^\star) + \boxed{b} + N$$

  where a and b are unknown and $a, b \in \mathbb{R}$
- One has to minimize $\|\mathbf{x} - a\mathbf{y}(k) - b\|_2$

Model known on a proportional scale

- Model only known on a proportional scale

$$X = \boxed{a} Y(K^\star) + \boxed{b} + N$$

where a and b are unknown and $a, b \in \mathbb{R}$

- One has to minimize $\|\mathbf{x} - a\mathbf{y}(k) - b\|_2$

Theorem (Correlation Power Analysis) *Where N is zero-mean Gaussian, the optimal distinguishing rule becomes*

$$\hat{k} = \arg\min_{k^\star} \min_{a,b} \|\mathbf{x} - a\mathbf{y}(k^\star) - b\|^2 \ ,$$

which is equivalent to maximizing the absolute value of the empirical Pearson's coefficient:

$$\hat{k} = \arg\max_{k^\star} |\hat{\rho}(k^\star)| = \frac{|\widehat{\mathrm{Cov}}(\mathbf{x}, \mathbf{y}(k^\star))|}{\sqrt{\widehat{\mathrm{Var}}(\mathbf{x})\widehat{\mathrm{Var}}(\mathbf{y}(k^\star))}}.$$

Proof given in the paper!

Model only partially known

knows

$$\mathbf{X} = \varphi(f(\mathbf{T}, k^\star)) + \mathbf{N}$$

- Leakage arising from a weighted sum of bits

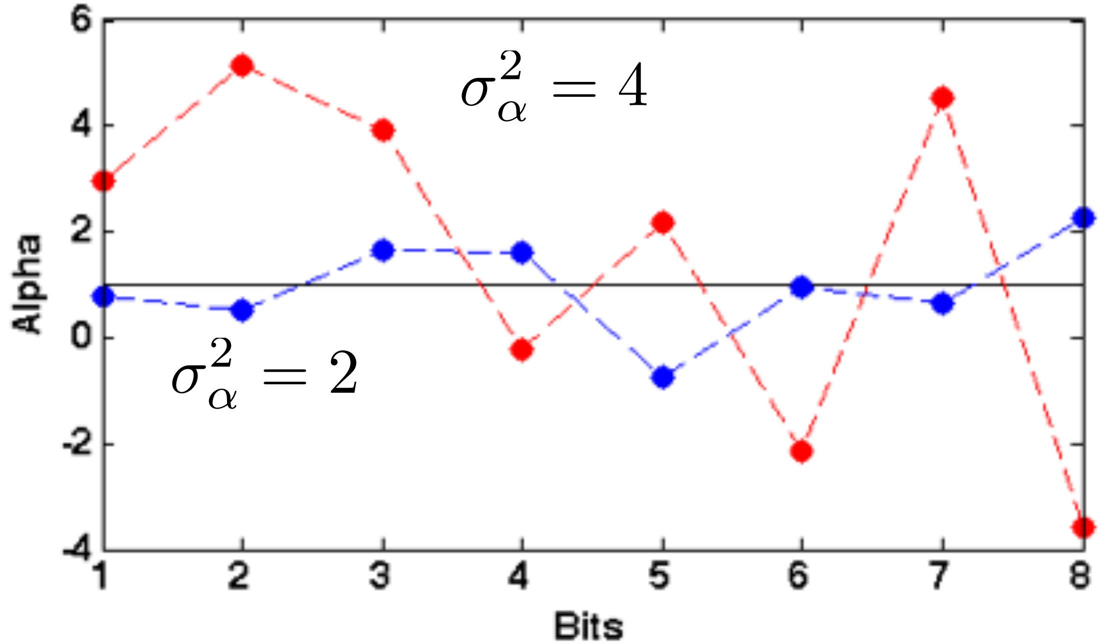
$$X = \sum_{j=1}^{n} \alpha_j [f(T, K^\star)]_j + N$$

- Weights are unknown, *epistemic* noise is present

Annelie Heuser, Olivier Rioul, Sylvain Guilley

Model only partially known

- Assumption about the weights
  - Unknown
  - Normally distributed $\alpha_j \sim \mathcal{N}(1, \sigma_\alpha^2)$
  - Fixed over one experiments



Model only partially known

Theorem (Optimal expression when the model is partially unknown) *Let* $\mathbf{Y}_{\boldsymbol{\alpha}}(K^\star) = \sum_{j=1}^{n} \alpha_j [f(\mathbf{T}, K^\star)]_j$ *and* $\mathbf{Y}_j(K^\star) = [f(\mathbf{T}, K^\star)]_j$*. When assuming that the weights are independently deviating normally from the Hamming weight model, i.e.,* $\forall j \in [\![1, 8]\!], \alpha_j \sim \mathcal{N}(1, \sigma_\alpha^2)$*, the optimal distinguishing rule is*

$$\mathcal{D}_{opt}^{\alpha,G}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \left(\gamma\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle + \mathbf{1}\right)^t \cdot \left(\gamma Z(k^\star) + I\right)^{-1} \cdot \left(\gamma\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle + \mathbf{1}\right) - \sigma_\alpha^2 \ln\det(\gamma Z(k) + I) \ ,$$

where $\gamma = \frac{\sigma_\alpha^2}{\sigma^2}$ *is the epistemic to stochastic noise ratio (ESNR),* $\langle \mathbf{x}|\mathbf{y}\rangle$ *is the vector with elements* $(\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle)_j = \langle \mathbf{x}|\mathbf{y}_j(k)\rangle$*,* $Z(k^\star)$ *is the* $n \times n$ *Gram matrix with entries* $Z_{j,j'}(k^\star) = \langle \mathbf{y}_j(k^\star)|\mathbf{y}_{j'}(k^\star)\rangle$*,* $\mathbf{1}$ *is the all-one vector, and* I *is the identity matrix.*

Proof given in the paper!

Model only partially known

Theorem (Optimal expression when the model is partially unknown) *Let* $\mathbf{Y}_{\boldsymbol{\alpha}}(K^\star) = \sum_{j=1}^{n} \alpha_j [f(\mathbf{T}, K^\star)]_j$ *and* $\mathbf{Y}_j(K^\star) = [f(\mathbf{T}, K^\star)]_j$. *When assuming that the weights are independently deviating normally from the Hamming weight model, i.e.,* $\forall j \in [\![1, 8]\!], \alpha_j \sim \mathcal{N}(1, \sigma_\alpha^2)$, *the optimal distinguishing rule is*

$$\mathcal{D}_{opt}^{\alpha,G}(\mathbf{x}, \mathbf{t}) = \arg\max_{k^\star} \left(\boxed{\gamma}\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle + \mathbf{1}\right)^t \cdot \left(\boxed{\gamma} Z(k^\star) + I\right)^{-1} \cdot \left(\boxed{\gamma}\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle + \mathbf{1}\right) - \sigma_\alpha^2 \ln\det\left(\boxed{\gamma} Z(k) + I\right) \ ,$$

where $\gamma = \frac{\sigma_\alpha^2}{\sigma^2}$ *is the epistemic to stochastic noise ratio (ESNR),* $\langle \mathbf{x}|\mathbf{y}\rangle$ *is the vector with elements* $(\langle \mathbf{x}|\mathbf{y}(k^\star)\rangle)_j = \langle \mathbf{x}|\mathbf{y}_j(k)\rangle$, $Z(k^\star)$ *is the* $n \times n$ *Gram matrix with entries* $Z_{j,j'}(k^\star) = \langle \mathbf{y}_j(k^\star)|\mathbf{y}_{j'}(k^\star)\rangle$, $\mathbf{1}$ *is the all-one vector, and* I *is the identity matrix.*

Proof given in the paper!

- In contrast to linear regression analysis the weights are not explicitly estimated

Empirical evaluation: known model

- Known model, only stochastic noise

$$X = \mathsf{HW}[\mathtt{Sbox}[T \oplus K^\star]] + N \qquad Y = \mathsf{HW}[\mathtt{Sbox}[T \oplus K^\star]]$$

- Compared distinguisher

$$\mathcal{D}_{opt}^{M,G}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} \langle \mathbf{x}|\mathbf{y}(k^\star)\rangle - \frac{1}{2}\|\mathbf{y}(k^\star)\|_2^2, \qquad \text{(Euclidean norm)}$$

$$\mathcal{D}_{opt\text{-}s}^{M,G}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} \langle \mathbf{x}|\mathbf{y}(k^\star)\rangle, \qquad \text{(Scalar product)}$$

$$\mathcal{D}_{opt}^{M,L}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} -\|\mathbf{x} - \mathbf{y}(k^\star)\|_1, \qquad \text{(Manhattan norm)}$$

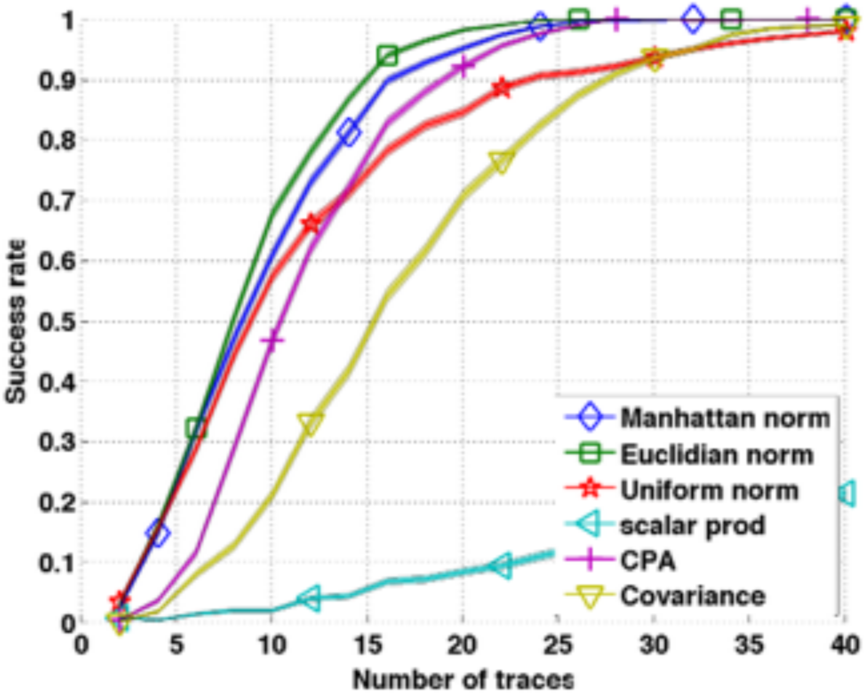
$$\mathcal{D}_{opt}^{M,U}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} -\|\mathbf{x} - \mathbf{y}(k^\star)\|_\infty, \qquad \text{(Uniform norm)}$$

$$\mathcal{D}_{Cov}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} |\langle \mathbf{x} - \overline{\mathbf{x}}|\mathbf{y}(k^\star)\rangle|, \qquad \text{(Covariance)}$$

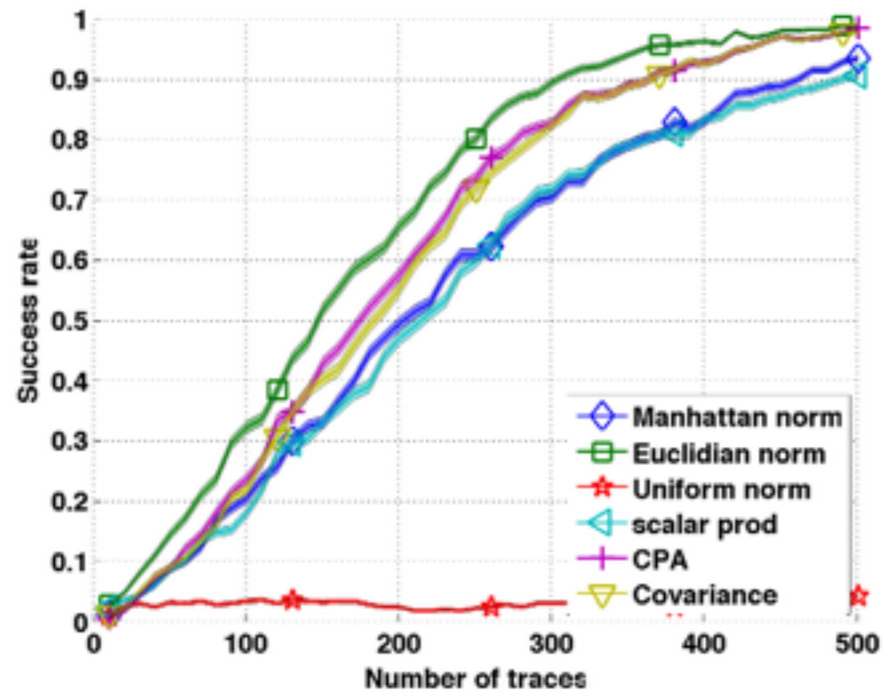
$$\mathcal{D}_{CPA}(\mathbf{x},\mathbf{t}) = \arg\max_{k^\star} \left| \frac{\langle \mathbf{x} - \overline{\mathbf{x}}|\mathbf{y}(k^\star)\rangle}{\|\mathbf{x} - \overline{\mathbf{x}}\|_2 \cdot \|\mathbf{y}(k^\star) - \overline{\mathbf{y}(k^\star)}\|_2} \right|. \qquad \text{(CPA)}$$

Annelie Heuser, Olivier Rioul, Sylvain Guilley

Gaussian noise

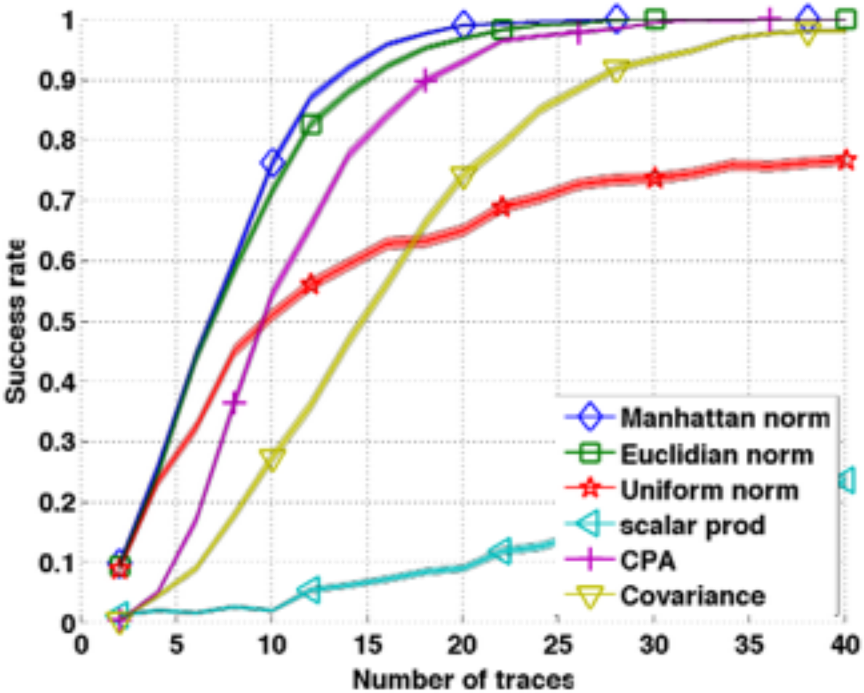


Sigma = 1

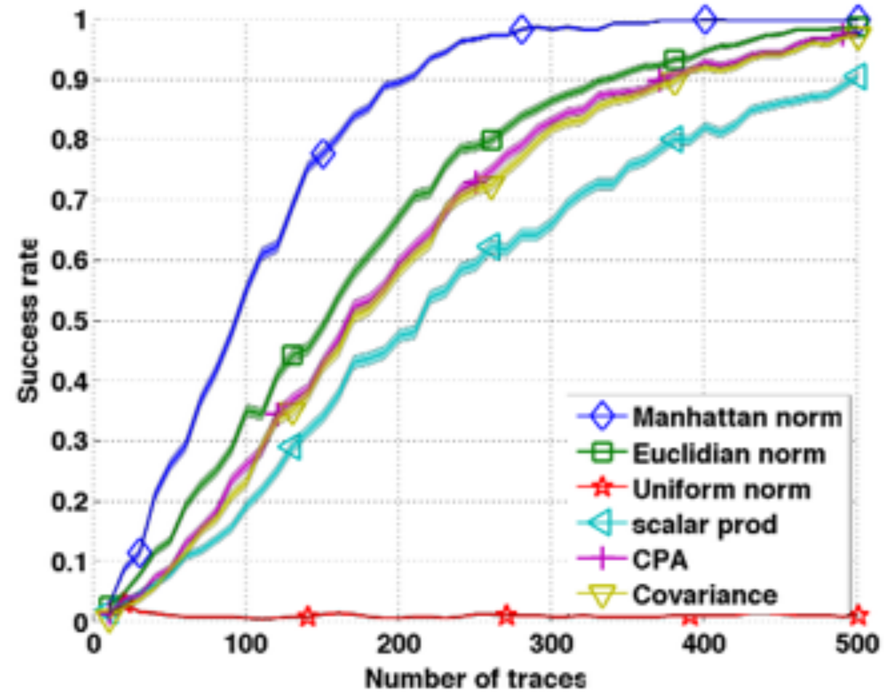


Sigma = 6

Laplacian noise

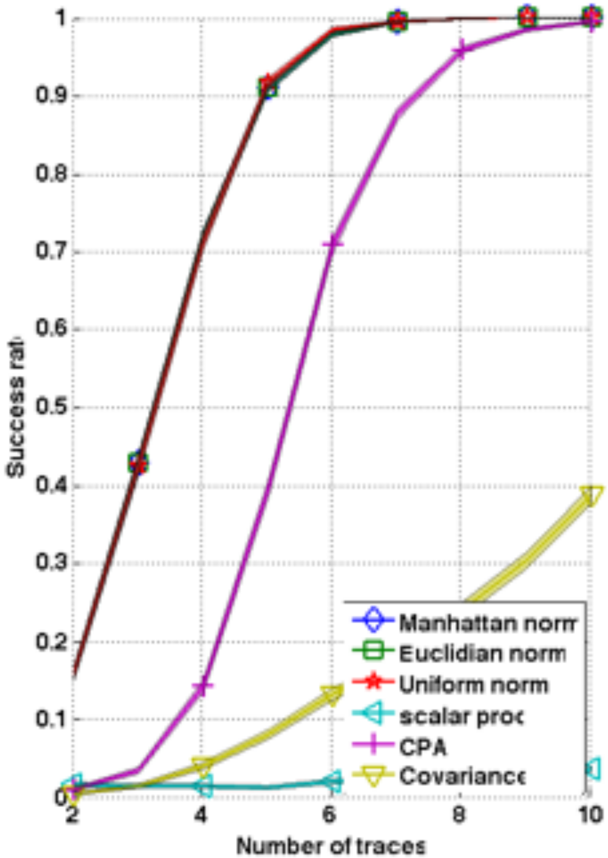


Sigma = 1

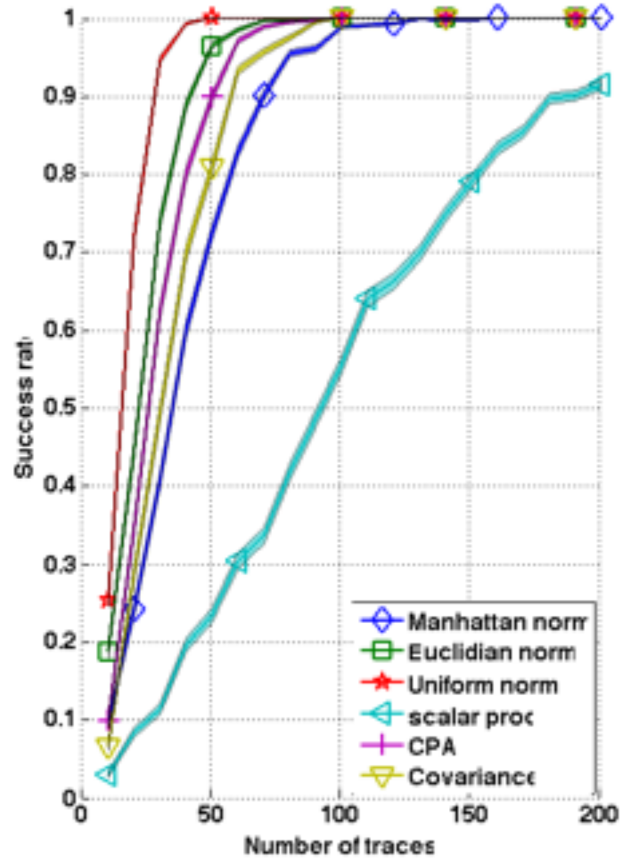


Sigma = 6

Uniform noise



Sigma = 1



Sigma = 6

Gaussian noise: partially unknown model

- Stochastic scenario

$$Y_j = [\mathtt{Sbox}[T \oplus K^\star]]_j \text{ for } j = 1, \ldots, 8$$

$$X = \sum_{j=1}^{8} \alpha_j Y_j(K^\star) + N$$

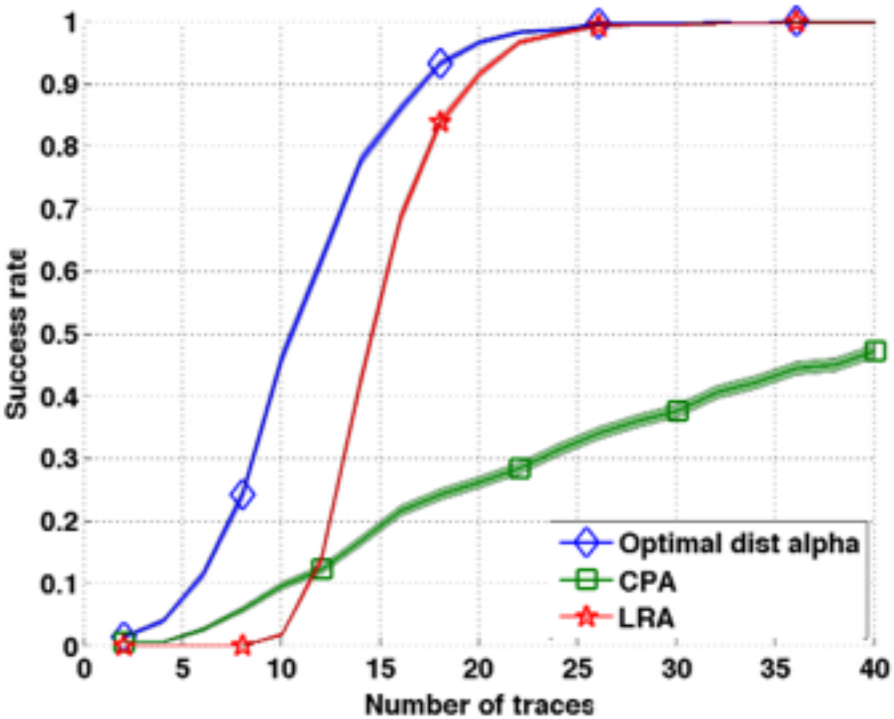
$$\alpha_j \sim \mathcal{N}(1, \sigma_\alpha^2)$$

- Optimal distinguisher compared with linear regression attack (LRA)

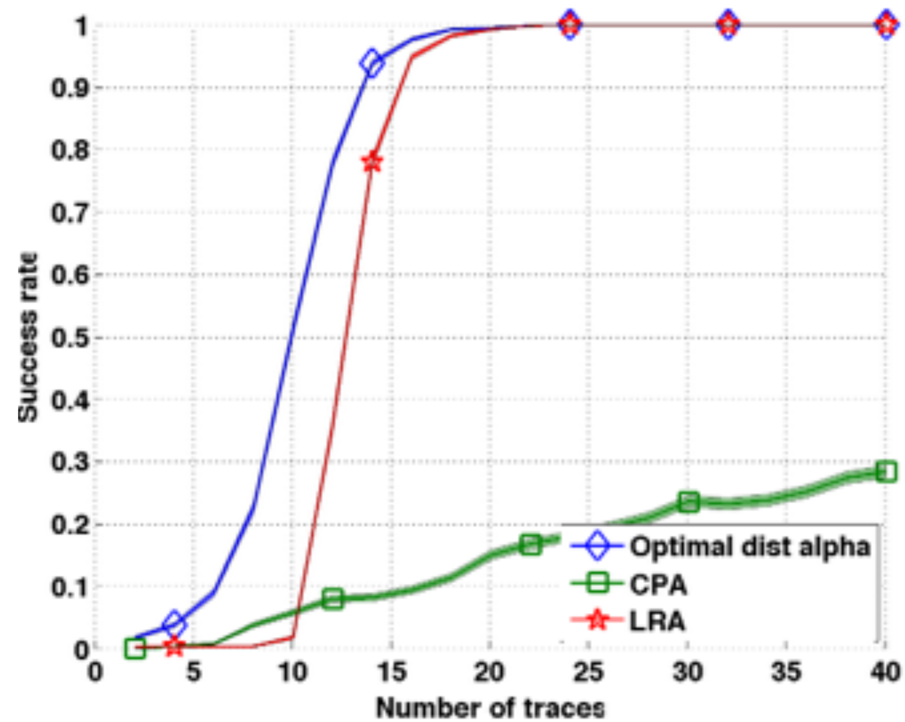
$$\mathcal{D}_{LRA}(\mathbf{x}, \mathbf{t}) = \arg\min_{k^\star} \frac{\|\mathbf{x} - \mathbf{y}'(k^\star) \cdot \hat{\boldsymbol{\alpha}}\|_2^2}{\|\mathbf{x} - \overline{\mathbf{x}}\|_2^2},$$

$$\mathbf{y}'(k) = (\mathbf{1}, \mathbf{y}_1(k), \mathbf{y}_2(k), \ldots, \mathbf{y}_8(k))$$

Gaussian noise: partially unknown model

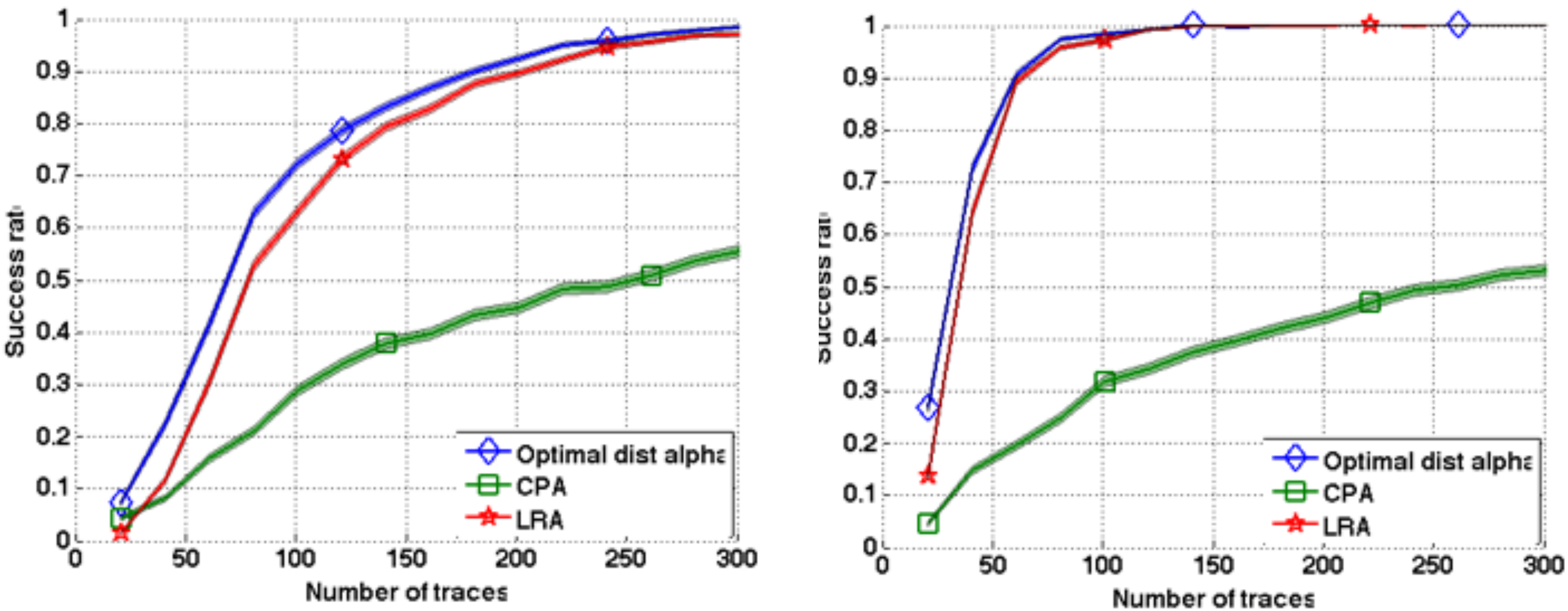


$$\sigma_\alpha = 2, \sigma = 1$$



$$\sigma_\alpha = 4, \sigma = 1$$

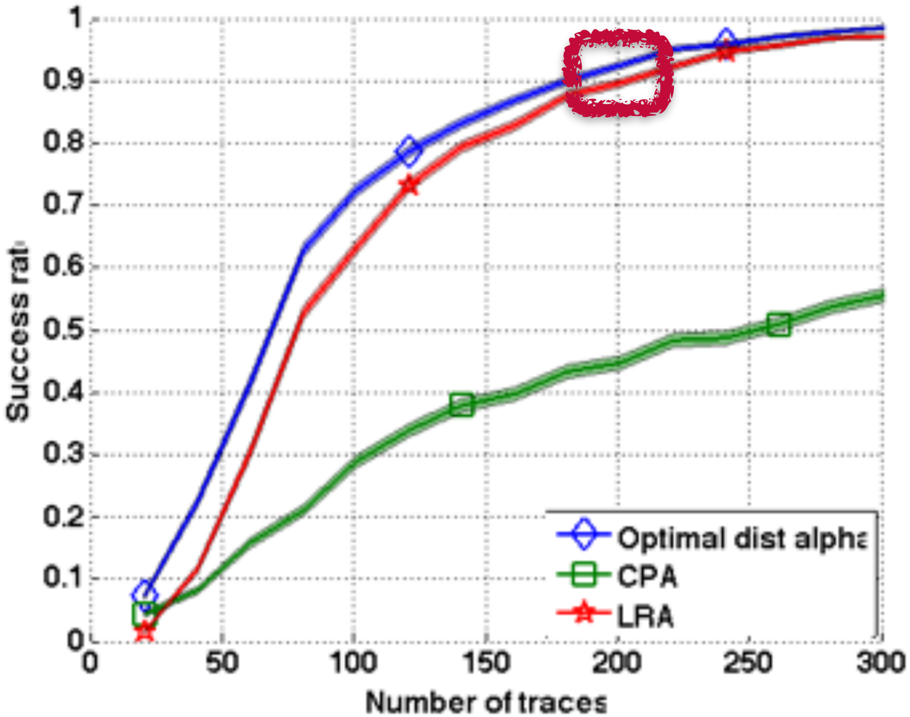
Gaussian noise: partially unknown model



$$\sigma_\alpha = 2, \sigma = 6 \qquad \sigma_\alpha = 4, \sigma = 6$$

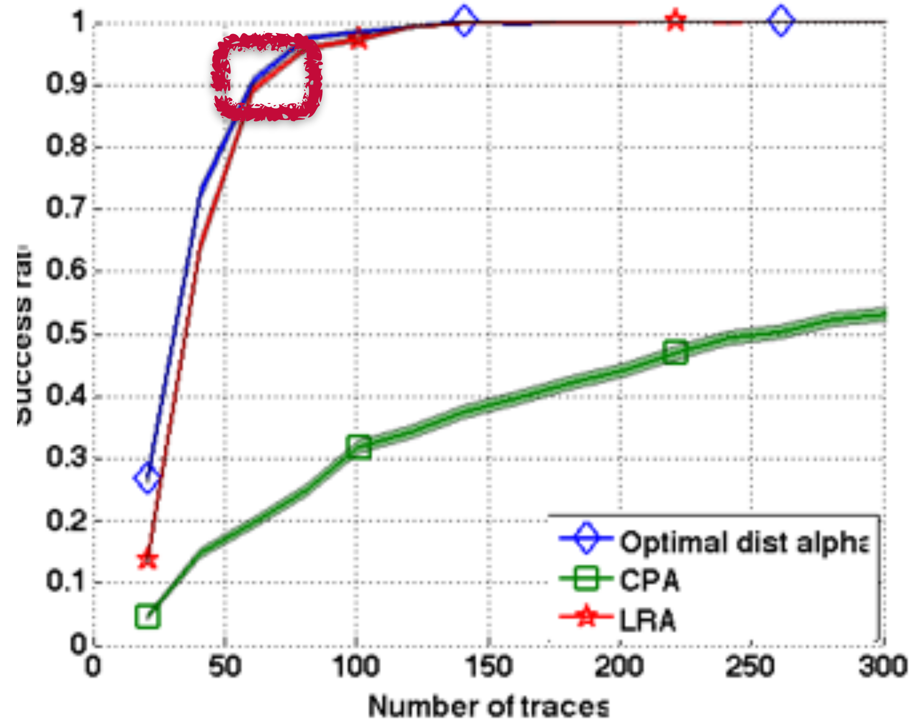
Gaussian noise: partially unknown model

200 traces



$$\sigma_\alpha = 2, \sigma = 6$$

60 traces



$$\sigma_\alpha = 4, \sigma = 6$$

Conclusion

- Transformation: SCA problem to communication theory problem
- **Known leakage model**
  - Gaussian noise: optimal distinguisher close to CPA for low SNR
  - Apart from Gaussian noise: optimal distinguishers differ from any known distinguisher
- **One-bit models**: optimal distinguisher close to DoM
- **Proportional scale**: CPA is optimal
- **Partially unknown leakage model**: optimal distinguisher performs better than LRA in the given context

Future work

- Application to real measurements
  - Preliminary step to determine the underlying scenario
  - Quantify the gain in terms of numbers of traces required to break the key, in concrete setups (feasibility OK on DPA contest v4).
- First-order optimal distinguisher (FOOD) to higher-order optimal distinguisher (HOOD) - accepted at ASIACRYPT

Thank you!!

Good Is Not Good Enough
Deriving Optimal Distinguishers from Communication Theory

Questions?

annelie.heuser@telecom-paristech.fr

Annelie Heuser is a Google European fellow in the field of privacy and is partially founded by this fellowship.