TELECOM ParisTech

Institut Mines-Télécom

When optimal means optimal

Finding optimal distinguishers from the mathematical theory of communication

Annelie Heuser, Olivier Rioul, Sylvain Guilley

Cryptarchi 2014

Outlook

- side-channel ⬌ communication channel
- optimal distinguisher
  - known model
  - partially known model
- empirical Results
- what comes next!

Motivation

- questions raised by the community

What distinguishes known distinguishers in terms of distinctive features?

Given a side-channel context what is the best distinguisher among all known ones?

- question we would like to answer

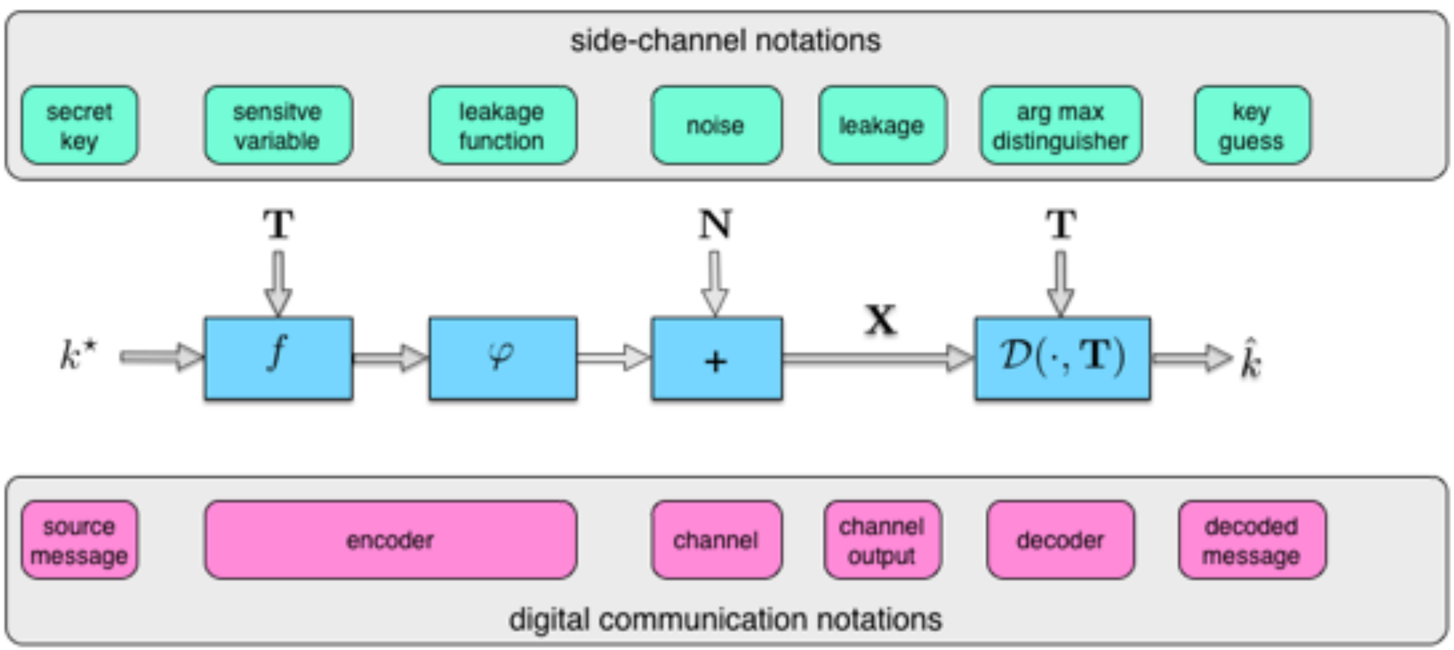
Given a side-channel scenario what is the best distinguisher among all possible ones?

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



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*

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

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

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

Template attack
[Chari+2002]

Optimal attack when the model is known

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

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

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

and for equiprobable keys this reduces to

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

Optimal Attack when the model is known

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

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

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

- most publications considered Gaussian noise
- furthermore we investigate in 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*

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

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

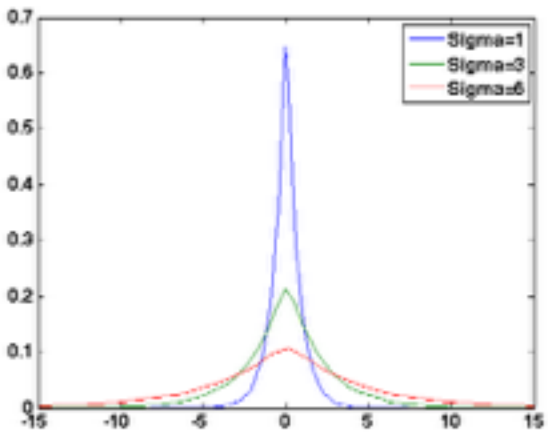
[Mangard+2011]

TELECOM ParisTech

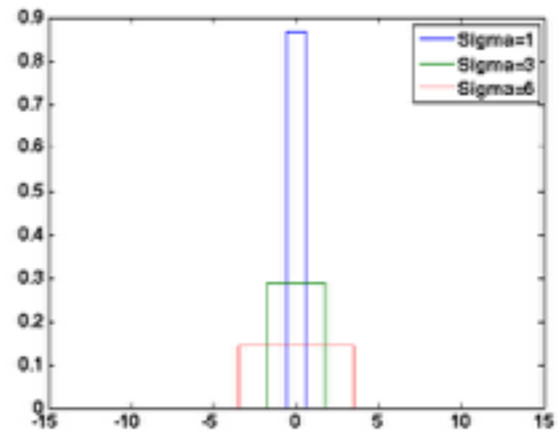
Uniform and Laplacian noise

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



Laplacian



uniform

Uniform and Laplacian noise

Theorem (Optimal expression for uniform and Laplacian noises) *When f and φ are known such that $Y(k) = \varphi(f(k,T))$, 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:* $\mathcal{D}_{opt}^{M,U}(\mathbf{x},\mathbf{t}) = \arg\max_k \; -\|\mathbf{x} - \mathbf{y}(k)\|_\infty$,
- *Laplace noise distribution:* $\mathcal{D}_{opt}^{M,L}(\mathbf{x},\mathbf{t}) = \arg\max_k \; -\|\mathbf{x} - \mathbf{y}(k)\|_1$.

- novel distinguishing rules
- cannot be approximated by correlation or covariance

Model known on a proportional scale

- Model only known on a proportional scale

$$X = aY(k^\star) + b + N$$

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*

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

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

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

Mono-bit leakage model

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

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

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

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

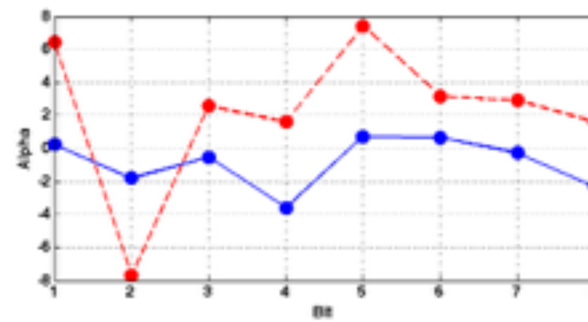
- nor to the t-test improvement [Coron+2000]

Model only partially known

- 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
- assumption about the weights
  - unknown
  - normally distributed
  - fixed over over one experiments/ over a set of traces



TELECOM ParisTech

Model only partially known

Theorem (Optimal expression when the model is partially unknown) *Let* $\mathbf{Y}_{\boldsymbol{\alpha}}(k) = \sum_{j=1}^{n} \alpha_j [f(\mathbf{T}, k)]_j$ *and* $\mathbf{Y}_j(k) = [f(\mathbf{T}, k)]_j$. *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*

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

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* $(\langle\mathbf{x}|\mathbf{y}(k)\rangle)_j = \langle\mathbf{x}|\mathbf{y}_j(k)\rangle$, $Z(k)$ *is the* $n \times n$ *Gram matrix with entries* $Z_{j,j'}(k) = \langle\mathbf{y}_j(k)|\mathbf{y}_{j'}(k)\rangle$, $\mathbf{1}$ *is the all-one vector, and* I *is the identity matrix.*

- if ESNR is small we recover the distinguisher when the model is known
- in contrast to linear regression the weights are not explicitly estimated

Empirical evaluation: known model

- known model, only stochastic noise
  $X = \text{HW}[\texttt{Sbox}[T \oplus k^\star]] + N \quad Y = \text{HW}[\texttt{Sbox}[T \oplus k]]$
- Compared distinguisher

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

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

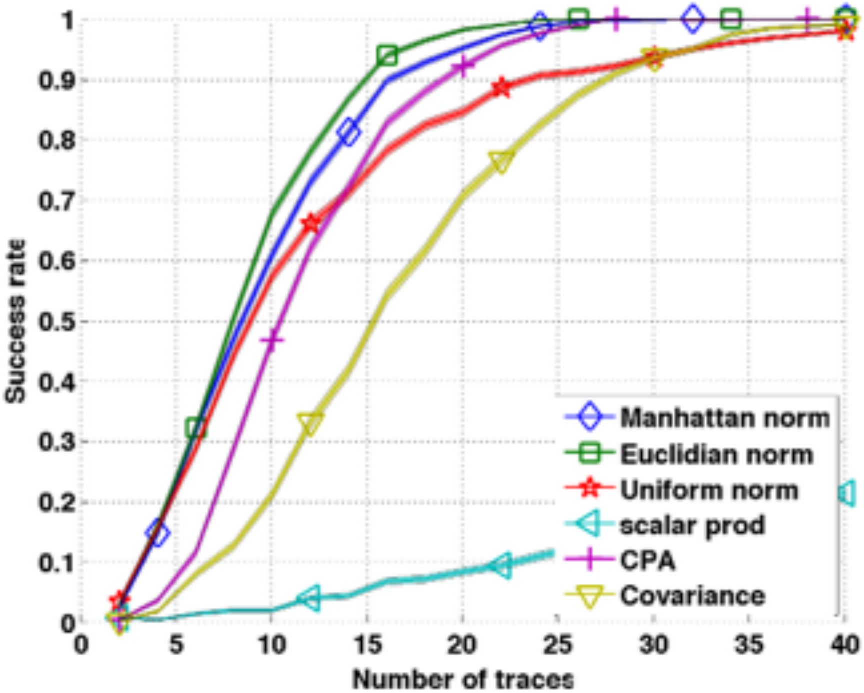
$$\mathcal{D}_{opt}^{M,L}(\mathbf{x},\mathbf{t}) = \arg\max_k -\|\mathbf{x} - \mathbf{y}(k)\|_1, \quad \text{(Manhattan norm)}$$

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

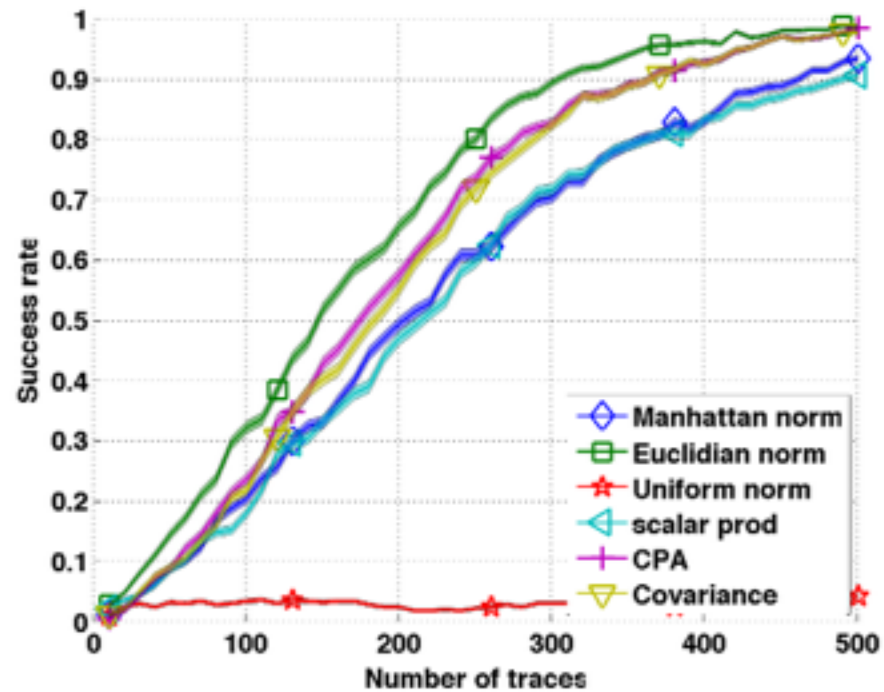
$$\mathcal{D}_{Cov}(\mathbf{x},\mathbf{t}) = \arg\max_k |\langle \mathbf{x} - \overline{\mathbf{x}}|\mathbf{y}(k)\rangle|, \quad \text{(Covariance)}$$

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

Gaussian noise

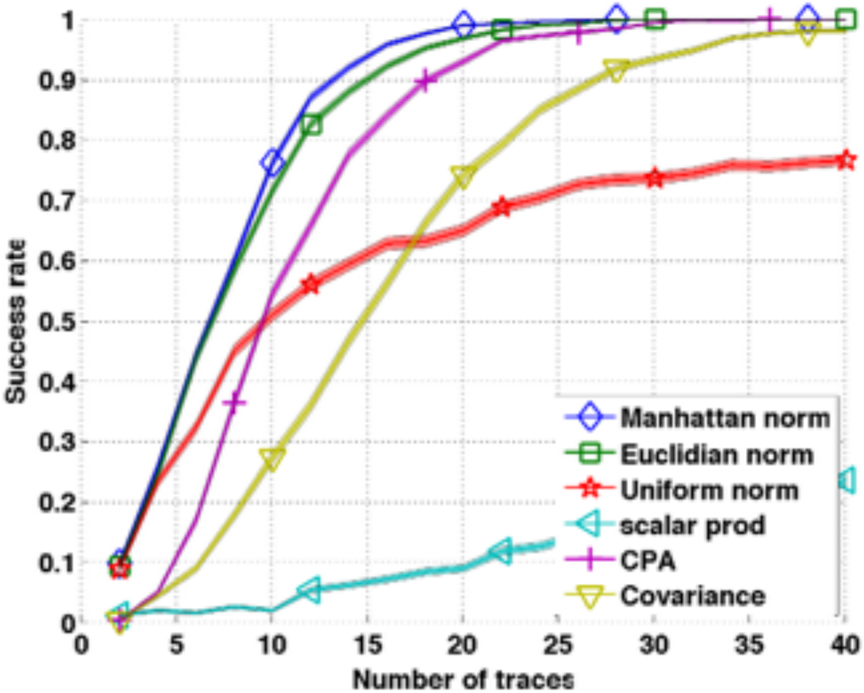


sigma = 1

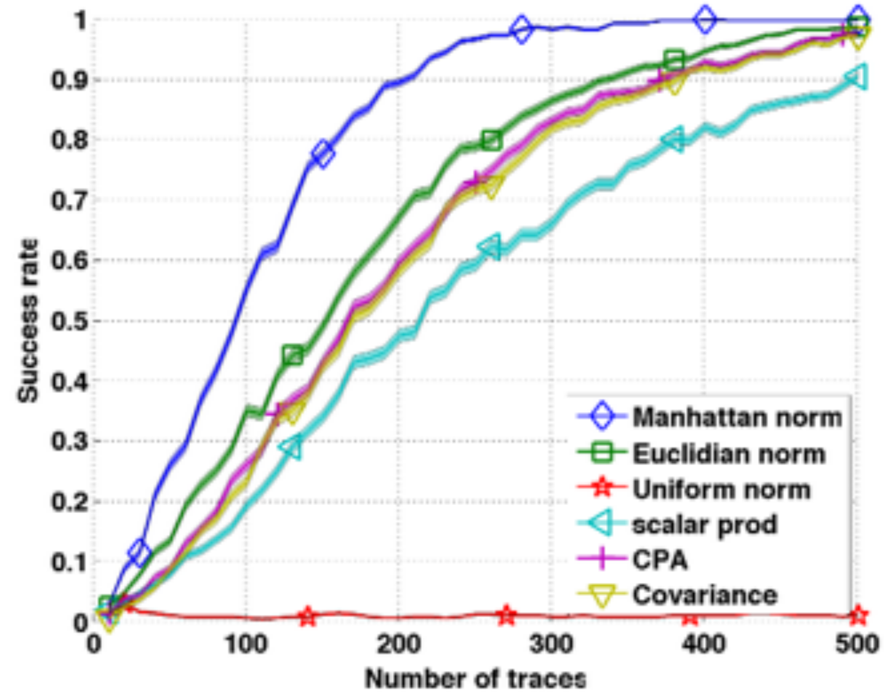


sigma = 6

Laplacian noise

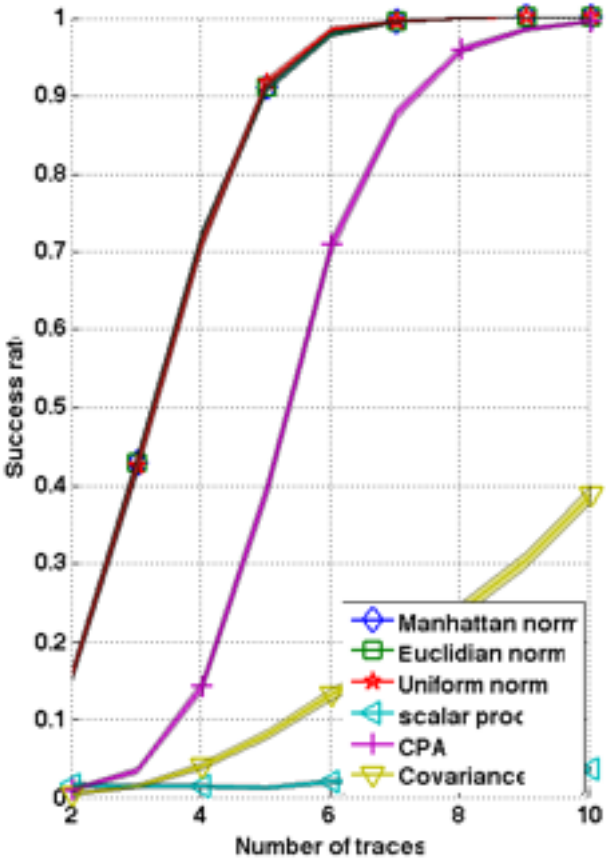


sigma = 1

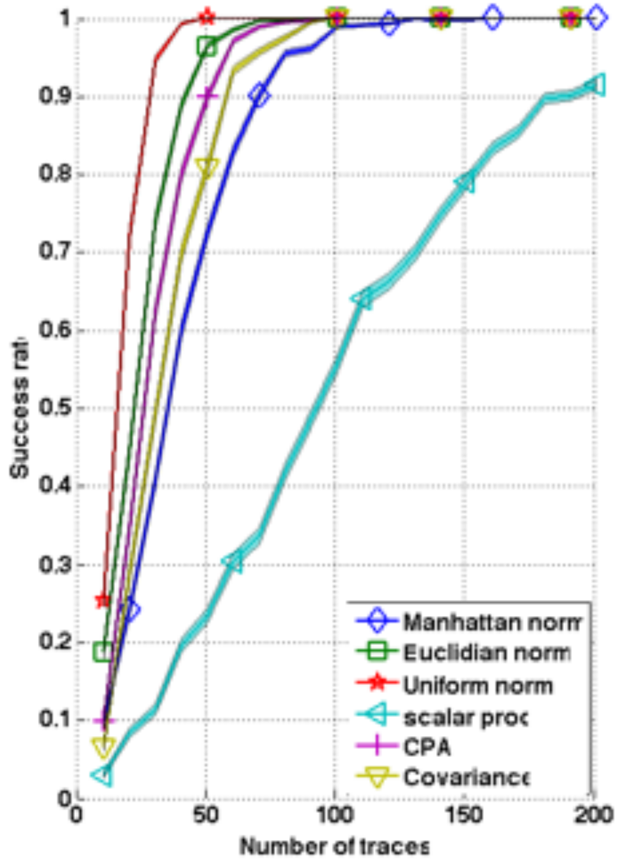


sigma = 6

Uniform noise



sigma = 1



sigma = 6

Gaussian noise: partially unknown model

- stochastic scenario

$$Y_j = [\mathtt{Sbox}[T \oplus k]]_j \text{ for } j = 1, \dots, 8$$

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

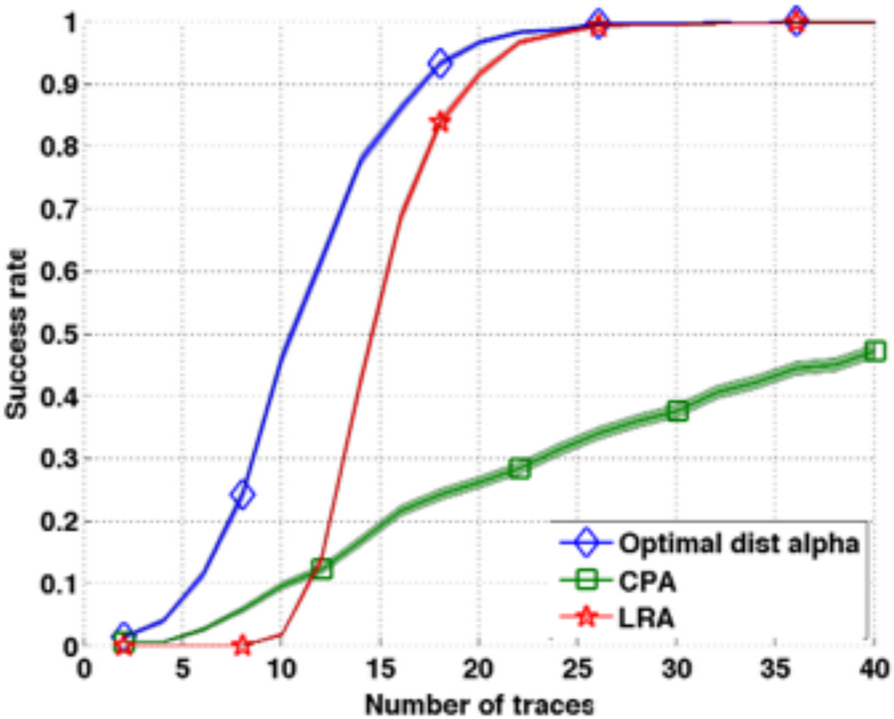
$$\alpha_j \sim \mathcal{N}(1, \sigma_\alpha)$$

- optimal distinguisher compared with Linear regression attack (LRA)

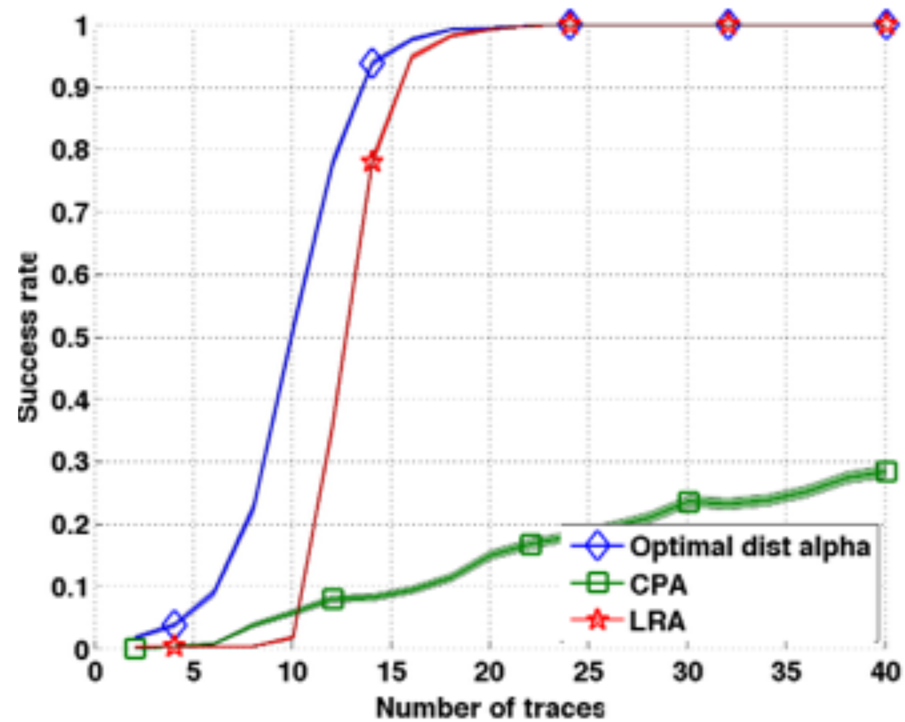
$$\mathcal{D}_{LRA}(\mathbf{x}, \mathbf{t}) = \arg\max_k \frac{\|\mathbf{x} - \mathbf{y}'(k) \cdot \boldsymbol{\beta}(k)\|_2^2}{\|\mathbf{x} - \overline{\mathbf{x}}\|_2^2},$$

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

Gaussian noise: partially unknown model

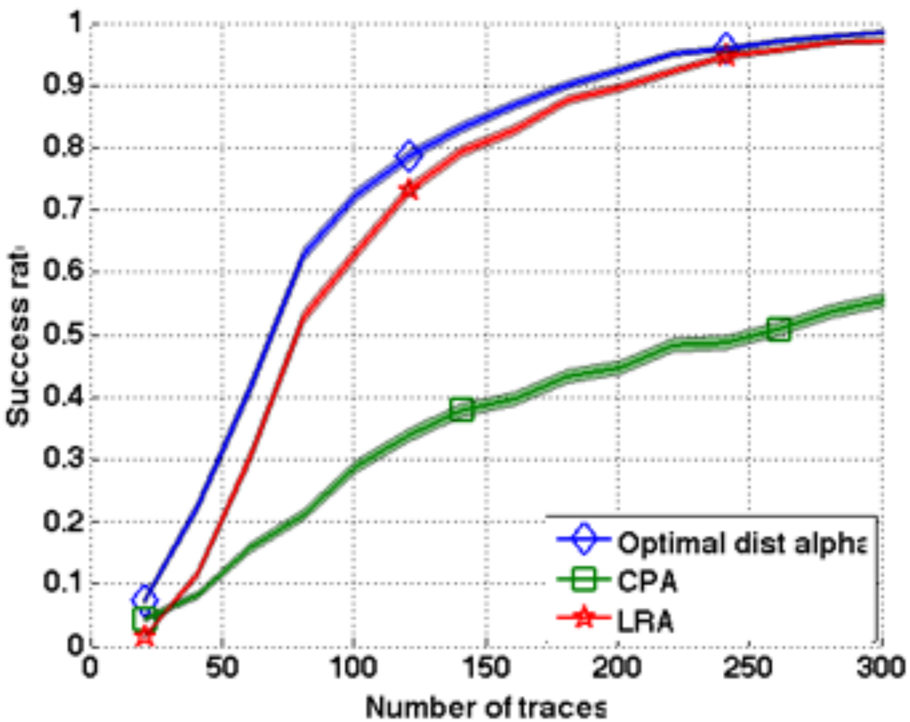


$\sigma_\alpha = 2, \sigma = 1$

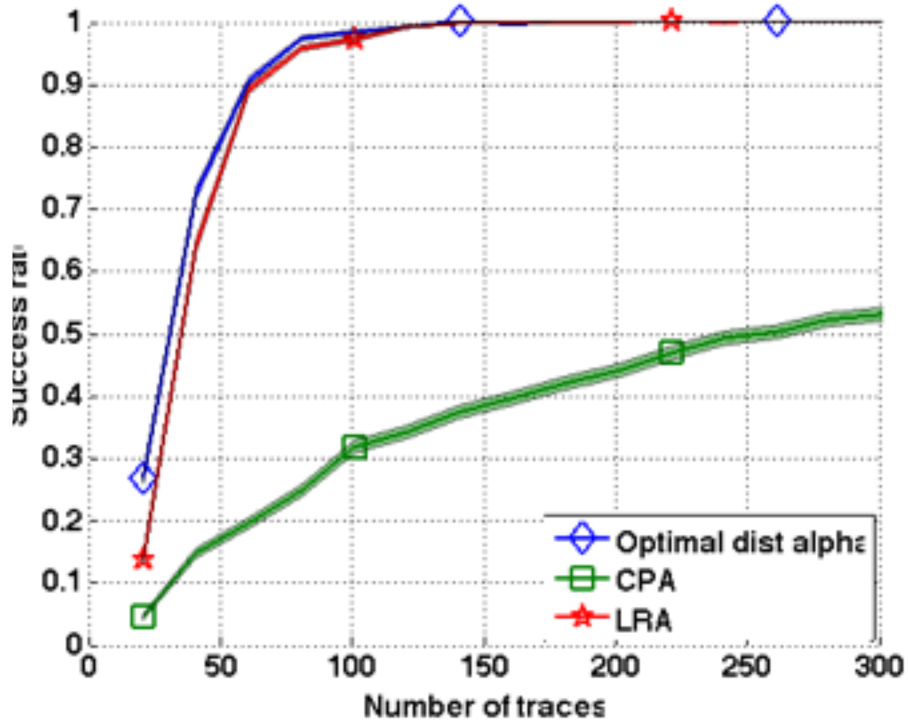


$\sigma_\alpha = 4, \sigma = 1$

Gaussian noise: partially unknown model



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



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

Conclusion

- Transformed the problem of SCA into a communication theory problem to derive optimal distinguisher in a given context
- known leakage model:
  - Gaussian noise: optimal distinguisher close to CPA for low SNR
  - apart from Gaussian noise: optimal distinguisher differ from any known distinguisher
- partially unknown leakage model: optimal distinguisher performs better than LRA in the given context

A mathematical study should prevail in side-channel analysis!

Future work

- Quantify the gain in terms of numbers of traces required to break the key, in concrete setups (feasibility OK on DPA contest v4).
- preliminary step to determine the underlying scenario
- application to higher-order attack (under submission)

[Chari+2002] Suresh Chari, Josyula R. Rao, and Pankaj Rohatgi. Template Attacks. In CHES, volume 2523 of LNCS, pages 13–28. Springer, August 2002. San Francisco Bay(Redwood City), USA.

[Coron+2000] Jean-S´ebastien Coron, Paul C. Kocher, and David Naccache. Statistics and Secret Leakage. In Financial Cryptography, volume 1962 of Lecture Notes in Computer Science, pages 157–173. Springer, February 20-24 2000. Anguilla, British West Indies.

[Kocher+1999] Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential Power Analysis. In Proceedings of CRYPTO'99, volume 1666 of LNCS, pages 388–397. Springer-Verlag, 1999.

[Margard+2011] Stefan Mangard, Elisabeth Oswald, and Franc¸ois-Xavier Standaert. One for All - All for One: Unifying Standard DPA Attacks. Information Security, IET, 5(2):100–111, 2011. ISSN: 1751-8709 ; Digital Object Identifier: 10.1049/iet-ifs.2010.0096.

Thank you!!

Questions?

to appear in CHES 2014, extended paper on eprint

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