



COSIC

How (not) to compare Side-Channel distinguishers



Benedikt Gierlichs, K.U. Leuven – COSIC

Part of this talk is based on: F.-X. Standaert, B. Gierlichs, I. Verbauwhede
"Partition vs. Comparison Side-Channel distinguishers", ICISC 2008

PASTIS workshop – December 2008, Gardanne

Motivation



- Given a cryptographic device that leaks sensitive information through a side-channel, many ways to exploit the side-channel leakage can be considered
 - DPA (Kocher, Correlation), single-bit, multi-bit, template attack, ...
- A natural question: which method is the **best**?
- Initial problem: what exactly does **best** mean?
 - Reliable? Robust? Efficient? Generic? ...
 - Probably a bit of all of that

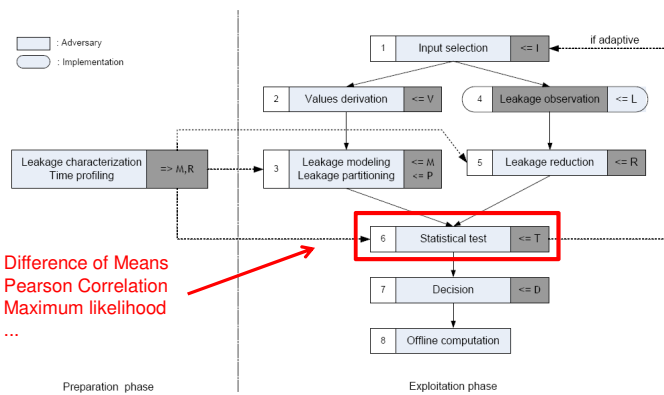
December 2008

Benedikt Gierlichs

PASTIS workshop, Gardanne

2

Focus: Distinguisher (statistical test)



December 2008

Benedikt Gierlichs

PASTIS workshop, Gardanne

3

Reliable / Robust

- Output of a distinguisher is deterministic
 - i.e. same inputs same outputs
- But, result of a DPA attack is typically probabilistic
 - It is based on sampling random variables with (a priori) unknown PDFs
 - Repeating the attack with a different sample set of the same size can lead to different results
 - The result of a single attack is most likely not representative
- We want to know how confident we can be about an attack's result
 - Evaluation context: compute the probability that the attack is successful (to be defined)
 - Practical attack: run the attack with independent data-sets and compare results

December 2008

Benedikt Gierlichs

PASTIS workshop, Gardanne

4

Efficient

- Situation is even worse as there is no common understanding of which quantities to measure
 - The number of power traces, i.e. samples from the random variables
 - Computational cost
 - Effort for device profiling
 - ...
- Literature typically focuses on number of power traces although this obviously is an incomplete metric
 - It may be possible to trade off computational and data complexity
 - How do we deal with template attacks?

Generic

- A distinguisher is supposed to detect patterns in samples of random variables
 - e.g. statistical dependence vs. independence
- Attack contexts (target devices, lab equipment, ...) can yield a broad variety of statistical dependencies
- Some distinguishers are specialized to detect particular (classes of) dependencies. Others are more generic.
- What are you interested in? A specialist or an all-rounder?

What did we do?

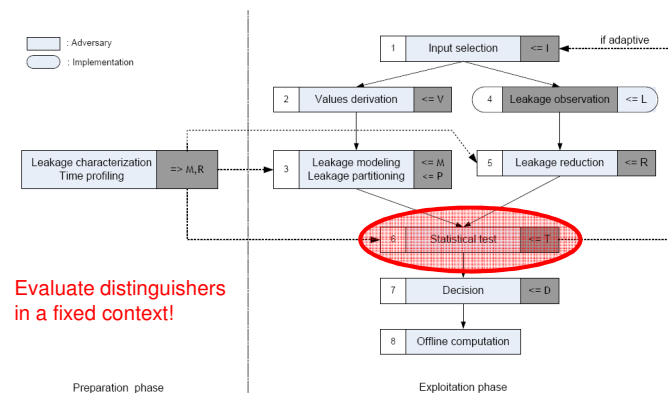
What's out there?

- Difference of Means test (KJJ99)
- Pearson correlation coefficient (BCO04)
- Bayesian analysis - templates (CRR02)
- Stochastic model (SLP05)
- Mutual Information Analysis (GBTP08)
- Usually applied to different devices hard to compare
- Single experiment as proof of concept vs. sound statistical evaluation
- ⇒ Our goal: discuss the fair empirical comparison and point out limitations

Our approach

- Compute a distinguisher's success probability as function of the number of measurements (details later)
- This metric allows to
 - Compare the distinguishing power of statistical tests given any fixed amount of input data
 - Evaluate how better sampling reduces uncertainty
- But, this metric only works in a fixed context, i.e.
 - Keep everything else (device, measurement setup) constant
 - Use the same (uniformly distributed) inputs
 - Use the same assumptions on the leakage model

Distinguisher = Statistical test



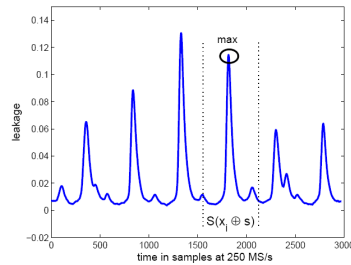
Our approach

- We evaluated 5 distinguishers: DoM, Pearson correlation coefficient, MIA, Bayesian analysis Variance test (a new proposal) that relates to MIA
- We repeated this evaluation on 2 target devices using 2 different measurement setups (one per device)
- Our results show that generic conclusions are difficult (impossible?)
- The question whether a distinguisher is generic remains unanswered. Need more tests using many different devices.

Target devices and implementations

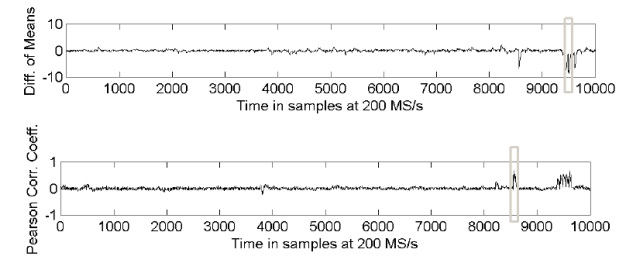
- AES-128 encryption in 8-bit RISC microcontrollers
 - PIC 16F877 running at 4 MHz
 - Atmel ATmega163 running at 3.57 MHz
 - Similar devices but substantially different leakage behavior
- All attacks target 8 first bits of the AES master key
- Power consumption measurements on 2 setups
 - Resistor in the supply circuit
 - Oscilloscopes: 1 GHz bandwidth
 - 250 MS/s sampling rate for the PIC
 - 200 MS/s sampling rate for the Atmel

Selection of time instants (PIC)



- All distinguishers applied to the same single time sample

Selection of time instants (Atmel)



- Individual selection for each distinguisher

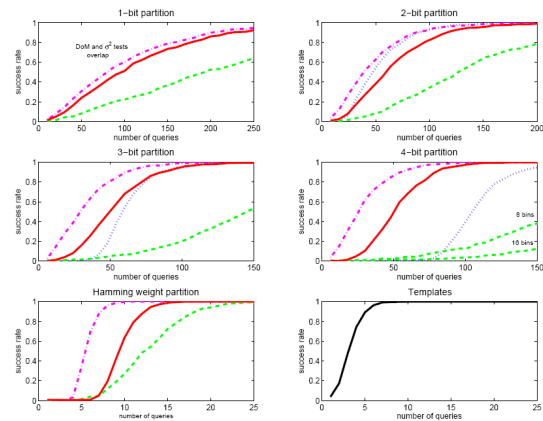
Evaluation metrics

- All distinguishers rank key candidates according to a score
- **Success rate**
 - Success rate of order o relates to the probability that the correct key is sorted among the first o key candidates by the adversary
 - $o=5$ means that the correct key must be among the 5 best candidates
- **Guessing entropy**
 - Guessing entropy relates to the number of keys that need to be tested after the DPA attack
 - Average position of the correct key in the sorted list of key candidates
 - Guessing entropy = 5 means that, on average, 5 candidates have to be tested to find the correct key

Experimental design

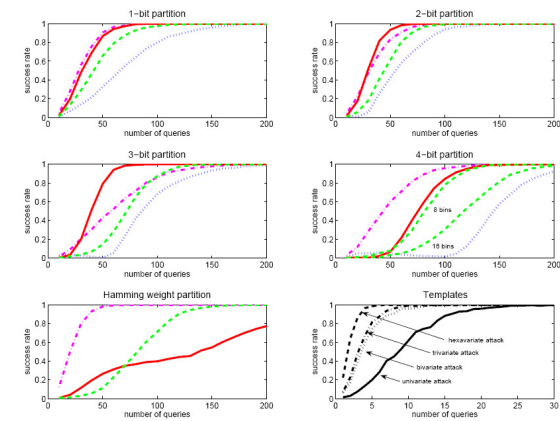
- For each statistical test and device, compute success rates and guessing entropies for:
 - various number of queries ($1 \leq q \leq 250$),
 - various partitions and models (1-bit, 2-bit, 3-bit, 4-bit, HW)
- Importance of statistical sampling
 - Each attack was repeated 1000 times using 1000 independent data-sets (independent in the strong sense)

1st order Success rate (PIC)



DoM
Pearson Corr.
MIA
Variance test
Templates

1st order Success rate (Atmel)



DoM
Pearson Corr.
MIA
Variance test
Templates

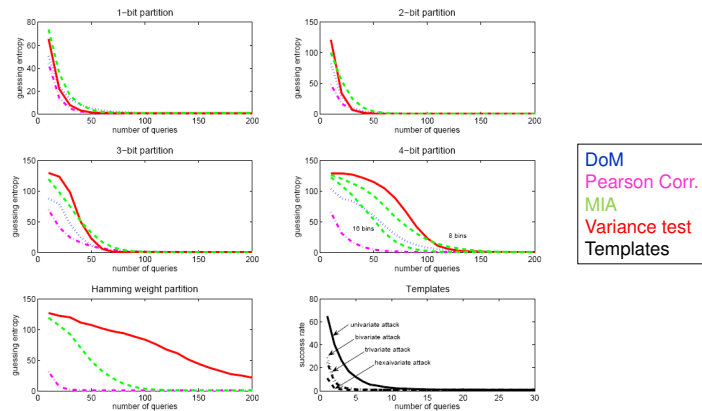
Observations

- Success rates are real numbers, i.e. for some fixed number of measurements the attacks work *sometimes*
 - Importance of statistical sampling
- Devices have different leakage behaviors
- PIC: \approx Hamming weight leakages ($\rho = 0.97$)
 - Hamming weight correlation very efficient
- Atmel: LSB leaks (much) more than other bits
 - 1-bit, 2-bit attacks very efficient
- Templates most efficient (unbounded profiling step, i.e. a large number of measurements in profiling)
- No general conclusions for non-profiled distinguishers

More specific comments

- DoM's weakness in multi-bit attacks:
 - measurements that are not assigned to one of the two sets (e.g. all-zeros or all-ones) are not used
- Number and selection of bins significantly impacts MIA's performance
- Other metrics bring different insights...

Guessing entropy (Atmel)



Summary

- Term **best** is not trivial to define
- We put forward a methodology for the fair empirical comparison of distinguishers
- We applied this methodology in 2 similar but different contexts
- Context-dependent results and conclusions
- Avoid wrong general claims
- Next: application to other platforms/distinguishers
 - Results for many target devices needed
 - Collision attacks
 - Impact of (un-)bounded profiling on Templates and Stochastic model

Thanks for your attention!

Questions



Slides and paper are available at <http://homes.esat.kuleuven.be/~bgierlic>
 benedikt.gierlich@esat.kuleuven.be