A Comprehensive Evaluation of Mutual Information Analysis Using a Fair Evaluation Framework

Carolyn Whitnall, Elisabeth Oswald

carolyn.whitnall@bris.ac.uk
Department of Computer Science, University of Bristol

16th August 2011

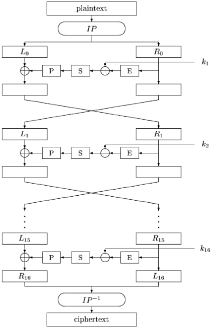


Photo from Iain Tate on Flickr

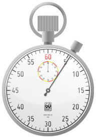


Photo from Casey Marshall on Flickr

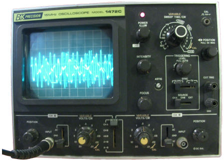
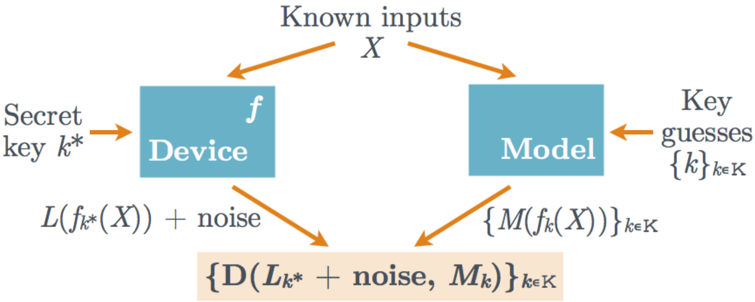


Photo from Becky Stern on Flickr

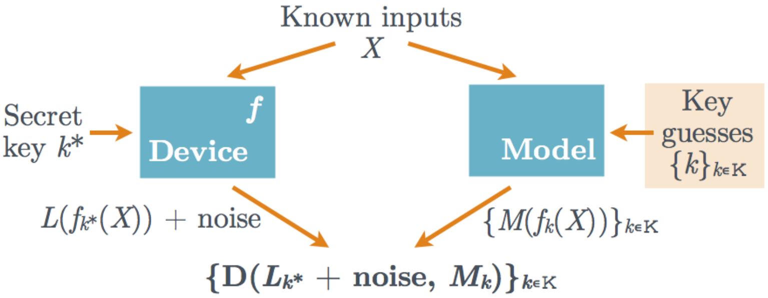
Algorithm + Device = Measurements!

But how to make the most of those measurements?

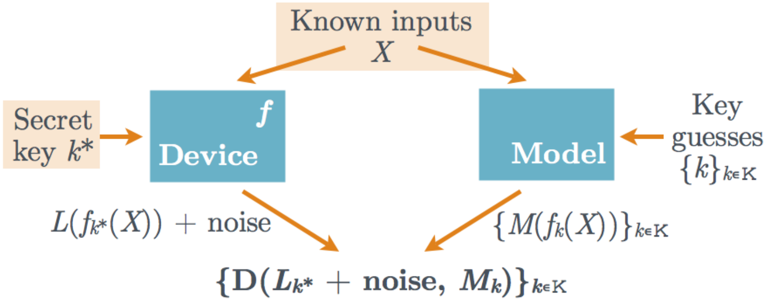
WHAT IS A SIDE-CHANNEL DISTINGUISHER?



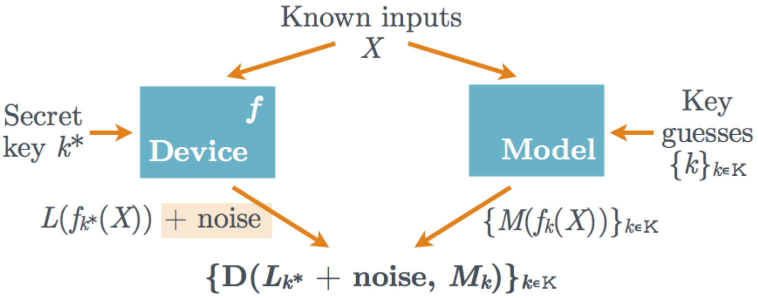
What is a Side-Channel Distinguisher?



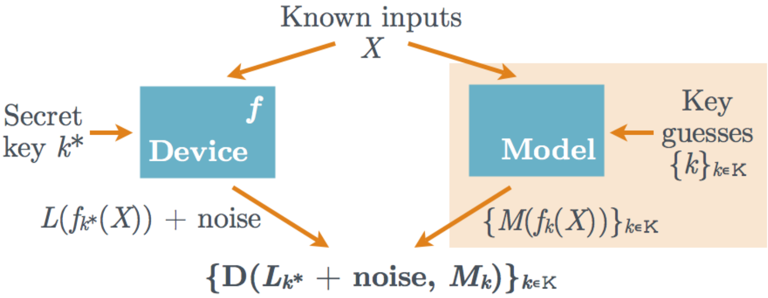
What is a Side-Channel Distinguisher?



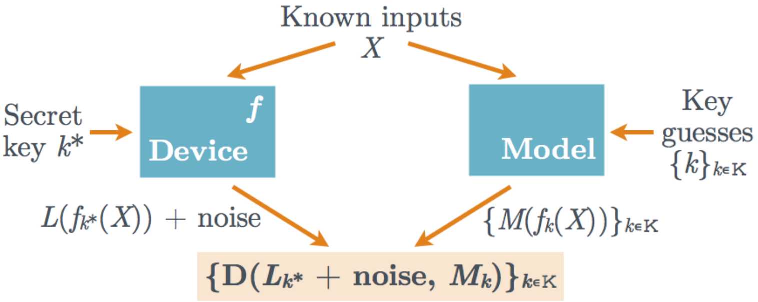
What is a Side-Channel Distinguisher?



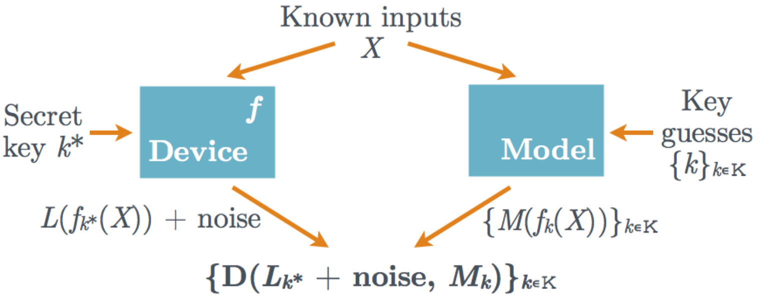
What is a Side-Channel Distinguisher?

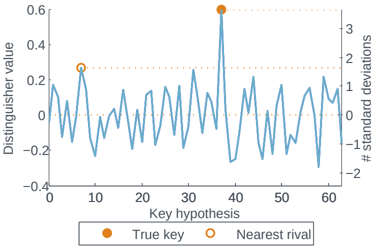


What is a Side-Channel Distinguisher?

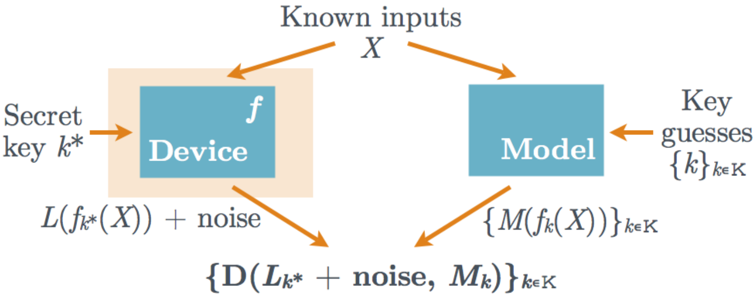


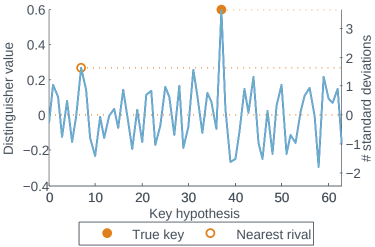
What is a Side-Channel Distinguisher?



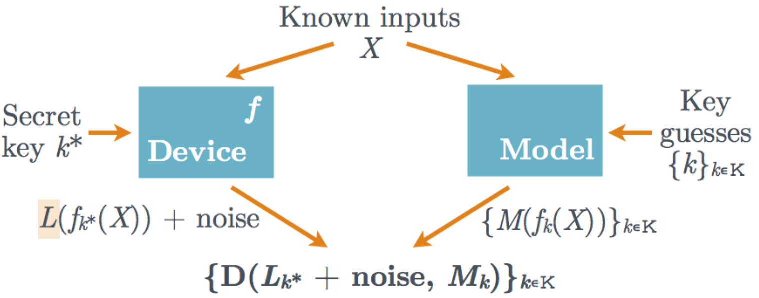


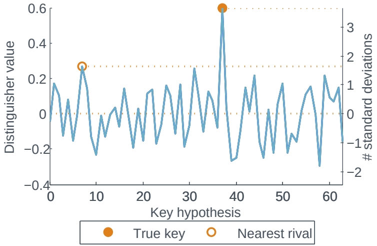
What is a Side-Channel Distinguisher?



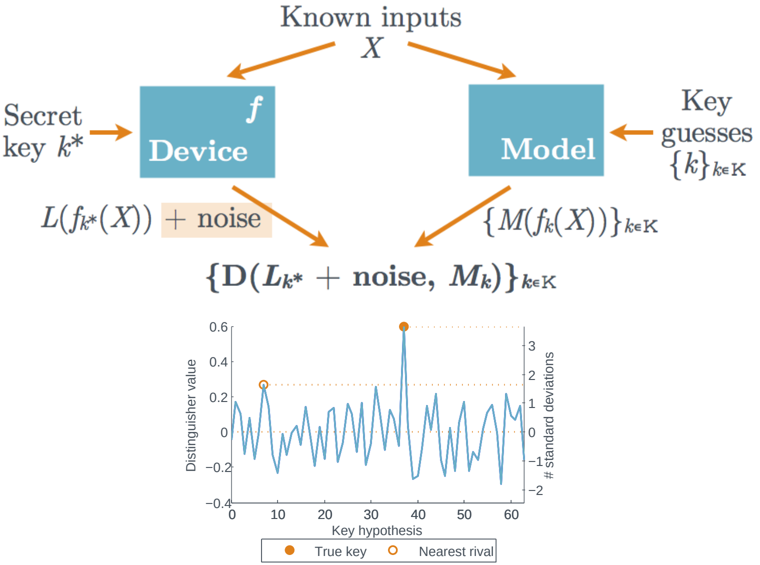


WHAT IS A SIDE-CHANNEL DISTINGUISHER?





What is a Side-Channel Distinguisher?



What Makes a Good Distinguisher?

The usual approach...

Desirable metric: "# of trace measurements required for key recovery"

- Not like-for-like: Practical outcomes highly sensitive to estimator choice
- Not computable: Sampling distributions (usually) unknown

Our contribution

'True' distinguishing vectors can be directly computed for well-defined hypothetical scenarios

Theoretic advantages $\not\Longrightarrow$ practical advantages (unequal estimation costs)
BUT
Certain characteristics have a strong bearing on likely practical outcomes

What features of the *theoretic* distinguishing vectors most contribute to its estimatability?

What Makes a Good Distinguisher?

The usual approach...

Desirable metric: "# of trace measurements required for key recovery"

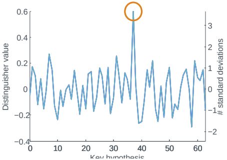
- Not like-for-like: Practical outcomes highly sensitive to estimator choice
- Not computable: Sampling distributions (usually) unknown

Our contribution

'True' distinguishing vectors can be directly computed for well-defined hypothetical scenarios

Theoretic advantages $\not\Longrightarrow$ practical advantages (unequal estimation costs)
BUT
Certain characteristics have a strong bearing on likely practical outcomes

What features of the *theoretic* distinguishing vectors most contribute to its estimatability?



Correct key ranking in the theoretic vector

- Distinguisher must isolate key in theory to stand a chance in practice

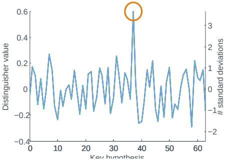
Nearest-rival distinguishing score – # s.d. between correct key value and highest ranked alternative

- The smaller the margin, the fewer the traces needed for estimation!

Average minimum support – how large an input support does the distinguisher need?

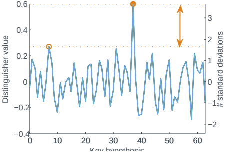
- An attack which needs to ‘see more inputs’ will inevitably need more traces

‘A Fair Evaluation Framework’



Correct key ranking in the theoretic vector

- Distinguisher must isolate key in theory to stand a chance in practice

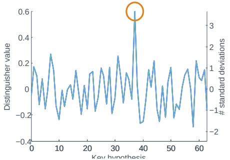


Nearest-rival distinguishing score – # s.d. between correct key value and highest ranked alternative

- The smaller the margin, the fewer the traces needed for estimation!

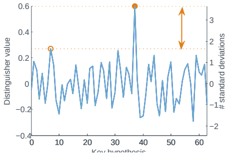
Average minimum support – how large an input support does the distinguisher need?

- An attack which needs to ‘see more inputs’ will inevitably need more traces



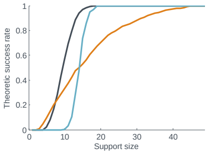
Correct key ranking in the theoretic vector

- Distinguisher must isolate key in theory to stand a chance in practice



Nearest-rival distinguishing score – # s.d. between correct key value and highest ranked alternative

- The smaller the margin, the fewer the traces needed for estimation!



Average minimum support – how large an input support does the distinguisher need?

- An attack which needs to ‘see more inputs’ will inevitably need more traces

The Distinguishers at a Glance...

MIA: Mutual information

- Defined as: $D(k) = \mathrm{I}(L_{k^*} + \varepsilon; M_k) = \mathrm{H}(L_{k^*} + \varepsilon) - \mathrm{H}(L_{k^*} + \varepsilon | M_k)$, where H is the differential entropy: $\mathrm{H}(X) = -\int_{x \in \mathcal{X}} p_X(x) \log_2(p_X(x))$
- *Functional of the distribution*—estimation problematic
  - DPA outcomes extremely sensitive to estimator choice; no 'ideal' exists
  - No general results for the sampling distributions

CPA: Pearson's correlation coefficient

- Defined as: $D(k) = \rho(L_{k^*} + \varepsilon, M_k) = \frac{\mathrm{Cov}(L_{k^*} + \varepsilon, M_k)}{\sqrt{\mathrm{Var}(L_{k^*} + \varepsilon)}\sqrt{\mathrm{Var}(M_k)}}$
- *Function of distributional moments*—estimation simple
  - Sample correlation coefficient suits a broad range of assumptions
  - Lots of 'nice' results for its sampling distribution

THE DISTINGUISHERS AT A GLANCE...

MIA: MUTUAL INFORMATION

- Defined as: $D(k) = \mathrm{I}(L_{k^*} + \varepsilon; M_k) = \mathrm{H}(L_{k^*} + \varepsilon) - \mathrm{H}(L_{k^*} + \varepsilon | M_k)$, where H is the differential entropy: $\mathrm{H}(X) = -\int_{x \in \mathcal{X}} p_X(x) \log_2(p_X(x))$
- *Functional of the distribution*—estimation problematic
  - DPA outcomes extremely sensitive to estimator choice; no 'ideal' exists
  - No general results for the sampling distributions

CPA: PEARSON'S CORRELATION COEFFICIENT

- Defined as: $D(k) = \rho(L_{k^*} + \varepsilon, M_k) = \frac{\mathrm{Cov}(L_{k^*} + \varepsilon, M_k)}{\sqrt{\mathrm{Var}(L_{k^*} + \varepsilon)}\sqrt{\mathrm{Var}(M_k)}}$
- *Function of distributional moments*—estimation simple
  - Sample correlation coefficient suits a broad range of assumptions
  - Lots of 'nice' results for its sampling distribution

Why 'Mutual Information Analysis'?

Proposed (Gierlichs *et al.*, 2008) as an enhancement to correlation DPA:

- *Optimal* in an information theoretic sense – quantifies total dependence
- *Generic* – should work even without a good power model
- *However*... correlation DPA frequently performs better in empirical comparisons

What can we learn from a theoretic evaluation?

Distinguisher	**Power model**	**Abbreviation**
Correlation DPA	Hamming weight	CPA(HW)
Mutual Information Analysis	Hamming weight	MIA(HW)
	Identity	MIA(ID)

Why 'Mutual Information Analysis'?

Proposed (Gierlichs *et al.*, 2008) as an enhancement to correlation DPA:

- *Optimal* in an information theoretic sense – quantifies total dependence
- *Generic* – should work even without a good power model
- *However*… correlation DPA frequently performs better in empirical comparisons

What can we learn from a theoretic evaluation?

Distinguisher	Power model	Abbreviation
Correlation DPA	Hamming weight	CPA(HW)
Mutual Information Analysis	Hamming weight	MIA(HW)
	Identity	MIA(ID)

Why 'Mutual Information Analysis'?

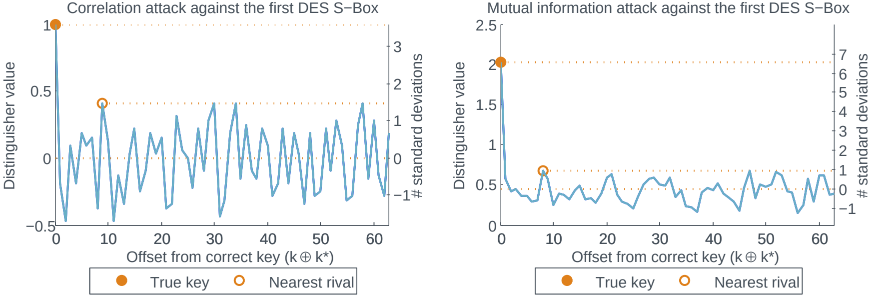
Proposed (Gierlichs *et al.*, 2008) as an enhancement to correlation DPA:

- *Optimal* in an information theoretic sense – quantifies total dependence
- *Generic* – should work even without a good power model
- *However*… correlation DPA frequently performs better in empirical comparisons

What can we learn from a theoretic evaluation?

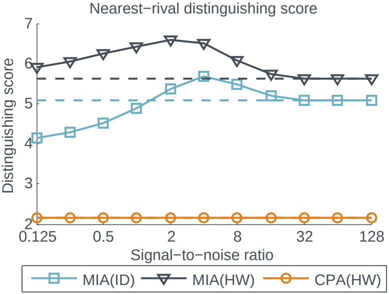
Distinguisher	**Power model**	**Abbreviation**
Correlation DPA	Hamming weight	CPA(HW)
Mutual Information Analysis	Hamming weight	MIA(HW)
	Identity	MIA(ID)

Noise-Free Hamming Weight Leakage



	CPA(HW)	MIA(HW)	MIA(ID)
Correct key ranking	1	1	1
Nearest-rival distinguishing score	2.14	5.61	5.08
Average minimum support	6	8	16

MIA Strangely Sensitive to Noise



Impact of noise on nearest rival distinguishing score:

- *Constant* for correlation-based distinguisher
- Evidence of *stochastic resonance* for MI-based distinguishers

(Note: no change in required support sizes throughout tested range)

MIA Shows Promise in Less Typical Scenarios...

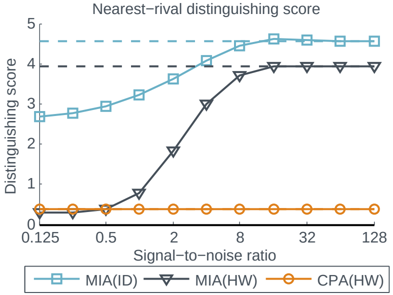
Candidate scenario: Hamming distance leakage from reference state $4_{(10)} = 0100_{(2)}$

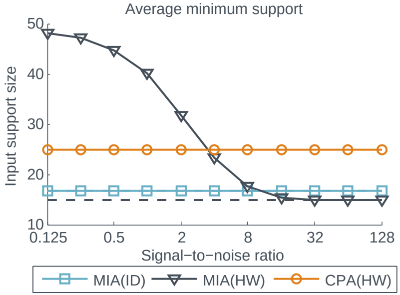
	$\lvert$CPA(HW)$\rvert$	MIA(HW)	MIA(ID)
Correct key ranking	1	1	1
Nearest rival distinguishing score	0.86	3.93	4.57
Average minimum support	34	15	17

- **Question 1:** Do these advantages persist in the presence of noise?
- **Question 2:** If so, can they be translated to practical advantages with standard estimation procedures?

. . . Still Looking Promising. . .

Question 1: Do the theoretic advantages in the 'pure signal' setting persist in the presence of noise?



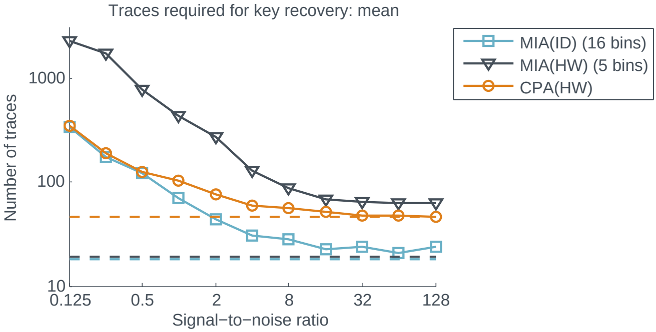


✗**MIA(HW)**
- Distinguishing score falls below that of CPA(HW)
- Hefty penalty in terms of required support size

✓**MIA(ID)**
- Maintains substantially larger distinguishing scores
- Required support size remains constant

. . . Experimental Results Confirm It!

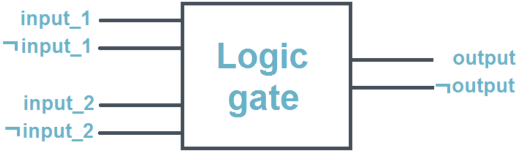
Question 2: Can the theoretic advantages be translated to practical advantages with standard estimation procedures?



✗**MIA(HW)** *Least efficient* in all but the pure-signal scenario

✓**MIA(ID)** *Comparable* to CPA(HW) when SNR $\leq$ 0.5, but *more efficient* thereafter

Bad News for Dual-Rail Precharge Logic?



- Unless output capacitances are *perfectly balanced* then some data-dependent signal will still leak
- Power consumption when *not* perfectly balanced can be likened to the HD from a constant reference state:
  - Reference state $\longleftrightarrow$ Bit-wise difference in the wire capacitances
- *Confirmed* by experimental attacks in Gierlichs *et al.*, 2008

MIA can be used to thwart countermeasures which resist correlation DPA!

In Conclusion

The problem: Empirical studies don't enable concrete, like-for-like comparisons between distinguishers

Our solution: A *theoretic* evaluation which bypasses the practical problems of estimation

Implications for MI-based distinguishers:

- There *are* scenarios where MI has a substantial *theoretic advantage* (e.g. Hamming distance leakage, DRP logic)
- Such advantages *can* be translated into practical advantages
- The (standardised) MI distinguishing vector exhibits a type of *stochastic resonance* as noise levels vary

Whitnall, C and Oswald, E: *A Fair Evaluation Framework for Comparing Side-Channel Distinguishers*. Journal of Cryptographic Engineering, 2011.

In Conclusion

The problem: Empirical studies don't enable concrete, like-for-like comparisons between distinguishers

Our solution: A *theoretic* evaluation which bypasses the practical problems of estimation

Implications for MI-based distinguishers:

- There *are* scenarios where MI has a substantial *theoretic advantage* (e.g. Hamming distance leakage, DRP logic)
- Such advantages *can* be translated into practical advantages
- The (standardised) MI distinguishing vector exhibits a type of *stochastic resonance* as noise levels vary

Whitnall, C and Oswald, E: *A Fair Evaluation Framework for Comparing Side-Channel Distinguishers*. Journal of Cryptographic Engineering, 2011.

Any questions?