Mutual Information Analysis[⋆]
A Generic Side-Channel Distinguisher

Benedikt Gierlichs[1], Lejla Batina[1], Pim Tuyls[1,2], and Bart Preneel[1]

[1] K.U. Leuven, ESAT/SCD-COSIC and IBBT
Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium
firstname.lastname@esat.kuleuven.be
[2] Philips Research Europe, Eindhoven, The Netherlands
pim.tuyls@philips.com

Abstract. We propose a generic information-theoretic distinguisher for differential side-channel analysis. Our model of side-channel leakage is a refinement of the one given by Standaert *et al.* An embedded device containing a secret key is modeled as a black box with a leakage function whose output is captured by an adversary through the noisy measurement of a physical observable. Although quite general, the model and the distinguisher are practical and allow us to develop a new differential side-channel attack. More precisely, we build a distinguisher that uses the value of the Mutual Information between the observed measurements and a hypothetical leakage to rank key guesses. The attack is effective without any knowledge about the particular dependencies between measurements and leakage as well as between leakage and processed data, which makes it a universal tool. Our approach is confirmed by results of power analysis experiments. We demonstrate that the model and the attack work effectively in an attack scenario against DPA-resistant logic.

Keywords: Differential Side-Channel Analysis (DSCA), Information Theory, Mutual Information, DPA-resistant logic.

1 Introduction

Pervasive devices such as smart cards, mobile phones, PDAs and more recently RFIDs and sensor nodes are now closely integrated into our lives. The devices typically operate in hostile environments and hence the data contained might be relatively easy compromised. This physical accessibility has led to a number of very powerful attacks targeting implementations. As an example we mention Differential Power Analysis (DPA) [9] which demonstrates that by monitoring the power dissipation of a smart card, the cryptographic keys can be rather efficiently extracted if no special countermeasures are taken. In the last decade many

[⋆] This work was supported in part by the IAP Programme P6/26 BCRYPT of the Belgian State (Belgian Science Policy), by FWO projects G.0475.05, and G.0300.07, by the European Comission through the IST Programme under Contract IST-2002-507932 ECRYPT NoE, and by the K.U. Leuven-BOF.

E. Oswald and P. Rohatgi (Eds.): CHES 2008, LNCS 5154, pp. 426–442, 2008.

other side-channels have been described such as electromagnetic emanation [15], timing [8], acoustics [16] *etc.* Both theory and practice have been developed and as a consequence several more advanced power analysis attacks such as correlation [2], template [3], and higher-order attacks [11] have been proposed as well as a broad range of countermeasures [4,6,7,10]. For all side-channels we use the terminology Differential Side-Channel Analysis (DSCA) when we refer to differential attacks.

DSCA attacks as introduced by Kocher *et al.* [9] use a boolean partitioning function to sort a set of power curves into two subsets. The function is usually defined on an intermediate value which can be predicted on the basis of a key hypothesis and known data. The difference between the averages of the power consumption curves of the two subsets shows a clear peak for the correct key guess. In this context, we refer to a statistical test, *e.g.* difference of means [9], Pearson correlation coefficient [2], Bayesian classification [3], as a side-channel distinguisher.

Micali and Reyzin propose theoretical models for side-channel security in [14]. In the model, the assumptions are very strong and in particular the adversary is the strongest possible, which makes their model hard to work with in practice. This was the motivation for the work of Standaert *et al.* [17]. They also use theoretical concepts such as Mutual Information to investigate side-channel leakage and attacks. In their work, the Mutual Information only measures the average amount of information present in measurements.

We introduce a Mutual Information-based distinguisher that constitutes the core of a new and generic differential side-channel attack: Mutual Information Analysis (MIA). In contrast to [17], we apply information theory to develop a powerful attack without any device characterization. The distinguisher uses only generic assumptions and is therefore more effective. Yet, the lack of assumptions may sometimes result in less efficient attacks. Further on, our model and the attack are successfully tested in practice. In general, while previous side-channel attacks tried to keep reducing the number of measurements needed by ever more sophisticated power consumption models, we take the opposite direction: we attempt to produce attacks that are still effective in more realistic attack scenarios, at the cost of a limited increase in the number of measurements.

This paper is organized as follows. Section 2 recalls the basic notions of information theory and introduces our information-theoretic model for side-channel leakage and analysis. In Sect. 3 we outline the construction of a distinguisher and we give a theoretical reasoning for our approach. Sect. 4 discusses practical aspects of MIA. In Sect. 5 we compare MIA with other known distinguishers. Sect. 6 gives empirical evidence for the correctness of our model and for the effectiveness of the proposed attack. We conclude our work in Sect. 7.

2 Preliminaries

2.1 Information Theory

We introduce some notions of information theory. For more details we refer to [5] and to the Appendix.

Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables on the (discrete) spaces $\mathcal{X}$ and $\mathcal{Y}$ with probability distributions $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{P}_{\mathbf{Y}}$ respectively. The reduction in uncertainty on $\mathbf{X}$ that is obtained by having observed $\mathbf{Y}$, is exactly equal to the information that one has obtained on $\mathbf{X}$ by having observed $\mathbf{Y}$. Hence the formula for the Mutual Information $\mathbf{I}(\mathbf{X};\mathbf{Y})$ is given by

$$\mathbf{I}(\mathbf{X};\mathbf{Y}) = \mathsf{H}(\mathbf{X}) - \mathsf{H}(\mathbf{X}|\mathbf{Y}) = \mathsf{H}(\mathbf{X}) + \mathsf{H}(\mathbf{Y}) - \mathsf{H}(\mathbf{X},\mathbf{Y}) = \mathbf{I}(\mathbf{Y};\mathbf{X})\,. \tag{1}$$

The Mutual Information satisfies $0 \leq \mathbf{I}(\mathbf{X};\mathbf{Y}) \leq \mathsf{H}(\mathbf{X})$. The lower bound is reached if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. The upper bound is achieved when $\mathbf{Y}$ fully determines $\mathbf{X}$. Hence, the larger the Mutual Information, the more close the relation between $\mathbf{X}$ and $\mathbf{Y}$ is to a one-to-one relation.

2.2 Side-Channel Model

In this section we introduce our information-theoretic model for side-channel leakage of cryptographic devices, which is a refinement of the model proposed by Standaert *et al.* in [17].

We consider a device (*e.g.* an IC) that carries out a cryptographic operation E_k, depending on a secret key k from a key space $\mathcal{K} = \{0,1\}^m$. The unknown key is modeled as a random variable $\mathbf{K}$ on $\mathcal{K}$. In order to analyze the impact of an adversary who can (up to a certain extent) observe the device's internal state, the device's side-channel leakage is modeled by a side-channel leakage function L. We assume that the values $\mathbf{L}$ of the leakage function depend on state transitions $\mathbf{W}$ (*e.g.* bit flips) in the device. The physical observable $\mathbf{O}$ represents (possibly noisy) measurements of $\mathbf{L}$.

Summarizing, we have the following model which consists of a cascade of two channels (see also Fig. 1): $\mathbf{W} \rightarrow \mathbf{L} \rightarrow \mathbf{O}$.

1. $\mathbf{W} \rightarrow \mathbf{L}$: The leakage channel through which information on the words $\mathbf{W}$ is leaked in $\mathbf{L}$.
2. $\mathbf{L} \rightarrow \mathbf{O}$: The (possibly noisy) observation channel through which $\mathbf{O}$ provides information on $\mathbf{L}$. An adversary has access to the output of this channel.

In the following we make these ideas more precise. We assume that the values of L are determined by state transitions (*e.g.* bit flips) in the device. These state transitions are provoked by a pair of words $(v_1, v_2) \in \{0,1\}^n \times \{0,1\}^n = \mathcal{W}$, where n is the device's word length, (*e.g.* previous and next state) being processed by the device. When a cryptographic operation E_k is executed, the pair (v_1, v_2) of words usually depends on the secret key k and is randomly distributed from an adversary's point of view. Therefore we model the occurring pairs as the random variable $\mathbf{W}$ on $\mathcal{W}$. The values of the leakage function L contain information on $\mathbf{W}$ and hence, while E_k is executed, information on the secret key k used in the device. Therefore we model the images of $\mathbf{W}$ under L as a random variable $\mathbf{L}$ on a discrete space $\mathcal{L}$

$$\mathsf{L} : \mathcal{W} \rightarrow \mathcal{L}\,;\;\; \mathbf{W} \mapsto \mathbf{L} = \mathsf{L}(\mathbf{W})\,. \tag{2}$$

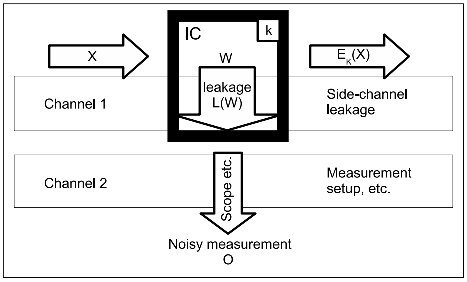

Fig. 1. Schematic illustration of the cascaded channels

Later, we will make the dependency of $\mathbf{L}$ on the key k explicit and denote it by $\mathbf{L}_k$. It is furthermore assumed that $\mathcal{L}$ is at most of size 2^{2n}, *i.e.* the leakage function L is surjective. For example, the Hamming weight model implies $\mathcal{L} = \{0, 1, \ldots, 7, 8\}$. The random variable $\mathbf{L}$ is observed by measuring a physical observable (voltage, radiation, *etc.*). The physical observable is modeled as the random variable $\mathbf{O}$ on a space $\mathcal{O}$.

Before an attack, the adversary obtains $q > 0$ measurement traces $o_{x_i}(t)$, $i = 1 \ldots, q$, by measuring $\mathbf{O}(t)$ while the device processes known data x_i with the cryptographic operation E_k over time t. During the attack, the adversary uses the information on $\mathbf{L}$ contained in $\mathbf{O}$ and aims at reconstructing the word sequence $\mathbf{W}$, which would allow to discriminate the secret key k.

The real side-channel leakage function of the device might not be known to the adversary. We thus denote her guess, *i.e.* the hypothetical leakage function, by $\hat{\mathsf{L}}$. For the sake of explanation, we assume $\hat{\mathsf{L}} = \mathsf{L}$ for the moment and address this issue later in Sect. 4.1. The adversary makes a guess $\hat{k} \in \mathcal{K}$ on the key k stored in the device. This implies a guess $\mathbf{W}_{\hat{k}}$ on the occurred pairs of words $\mathbf{W}$. The guess $\mathbf{W}_{\hat{k}}$ in turn implies a guess $\hat{\mathbf{L}}_{\hat{k}} = \hat{\mathsf{L}}(\mathbf{W}_{\hat{k}})$ on the output values $\mathbf{L}_k$ of the real leakage function. In the last step the adversary checks whether her guess $\hat{\mathbf{L}}_{\hat{k}}$ is compatible with the observed measurement values $\mathbf{O}$.

In order to explain the attack, we first restrict ourselves to the interesting point(s) in time $t = \tau_j$ when the pair of words $\mathbf{W}$ being processed depends on the result of a function $f_k : \{0,1\}^m \to \{0,1\}^n, \mathbf{X} \mapsto f_k(\mathbf{X})$ applied to a known input $\mathbf{X}$. We assume that the cryptographic primitive E_k and its implementation are known to the adversary, that $f_k(\cdot)$ is a suitable intermediate result of $E_k(\cdot)$, and that the inputs $\mathbf{X}$ are chosen uniformly at random from $\{0,1\}^m$. Further, we assume that the key space is $\{0,1\}^m$.

2.3 Side-Channel Attack

We denote by $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$ the multiset of q measurements of the physical observable $\mathbf{O}$ obtained when the known inputs $x_1, \ldots, x_q$ were processed by the device. Our side-channel adversary uses a distinguisher $\mathcal{D}$, which takes as input the measurements $o_{x_1}, \ldots, o_{x_q}$ and the inputs $x_1, \ldots, x_q$, and outputs the key guess k^*. The adversary's advantage of using this distinguisher is defined as the probability that the distinguisher's key guess k^* is indeed the correct key k.

3 The Information-Theoretic Distinguisher

In this section we derive our distinguisher and analyze it formally in our attack scenario.

3.1 Construction

Let $L_0, \ldots, L_l$ be subsets of the space $\mathcal{L}$. The set $\{L_0, \ldots, L_l\}$ is a partition of $\mathcal{L}$ and the elements $L_i, i = 0, \ldots, l$ are called atoms.

To each possible key guess $\hat{k} \in \mathcal{K}$, which implies a guess $\mathbf{W}_{\hat{k}}$ on the pairs of words, we associate a partition $\{L_0^{\hat{k}}, \ldots, L_l^{\hat{k}}\}$ of $\mathcal{L}$ which is defined by $L_i^{\hat{k}} = \{x \in \{0,1\}^m \quad | \quad \hat{\mathsf{L}}(\mathbf{W}_{\hat{k}}) = i \wedge \mathbf{W}_{\hat{k}} = (v_1, f_{\hat{k}}(x))\} \quad \text{for} \quad i = 0, \ldots, l$. That is, we associate all inputs values $\mathbf{X} = x$ that leak $\hat{\mathbf{L}}_{\hat{k}} = i$ under the key guess $\hat{k}$ to $L_i^{\hat{k}}$. Each partition $\{L_0^{\hat{k}}, \ldots, L_l^{\hat{k}}\}$ of $\mathcal{L}$ induces a subdivision[1] of the measurement space $\mathcal{O}$, since each measurement is associated with an input x.

Let $\mathbb{P}_{\hat{\mathbf{L}}_{\hat{k}}}$ and $\mathbb{P}_{\mathbf{O}}$ denote the probability distributions of the random variables $\hat{\mathbf{L}}_{\hat{k}}$ and $\mathbf{O}$ respectively.

Given the multiset of measurements $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$ and a partition of $\mathcal{L}$, we define the following set of conditional distributions $\{\mathbb{P}_{\mathbf{O}|L_i^{\hat{k}}}\}_{i=0}^{l}$. The distributions $\mathbb{P}_{\mathbf{O}|L_i^{\hat{k}}}$ describe the random variable $\mathbf{O}$ given the atoms $L_i^{\hat{k}}$ for a hypothetical key $\hat{k}$. They represent a (possibly noisy) observation channel $\hat{\mathbf{L}}_{\hat{k}} \rightarrow \mathbf{O}$ which depends on the hypothetical key $\hat{k}$ and the actual key k. The attacker will look for the distribution that is most likely compatible with the measurement results.

We compute an estimation of the Mutual Information $\mathbf{I}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O})$ under the key guess $\hat{k}$ while the actual key is k as

$$\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}) = \tilde{\mathsf{H}}(\mathbf{O}) - \tilde{\mathsf{H}}(\mathbf{O}|\hat{\mathbf{L}}_{\hat{k}}), \tag{3}$$

where $\tilde{\mathsf{H}}(\cdot)$ denotes an estimated entropy.

The distributions $\mathbb{P}_{\mathbf{O}|L_i^{\hat{k}}}$ are determined empirically by generating the histograms of the measurements $o_{x_1}, \ldots, o_{x_q}$ associated to the atoms of the partition $\{L_0^{\hat{k}}, \ldots, L_l^{\hat{k}}\}$. They are estimated by

$$\tilde{\mathbb{P}}_{\mathbf{O}|L_i^{\hat{k}}} = \frac{|\{o_{x_j} = o |\, x_j \in L_i^{\hat{k}}\}|}{|L_i^{\hat{k}}|}$$

[1] In contrast to a partition, the atoms of a subdivision do not necessarily have an empty intersection.

where $|\{\cdot\}|$ denotes the cardinality of a set. The distribution $\mathbb{P}_{\mathbf{O}}$ is determined empirically as $\tilde{\mathbb{P}}_{\mathbf{O}} = |\{o_{x_j} = o\}|/q$.

We define our distinguisher $\mathcal{D} : \mathcal{O}^q \times \{0,1\}^m \rightarrow \mathcal{K}$ as follows: given a multiset $\mathcal{M} = \{o_{x_1}, \ldots, o_{x_q}\}$ of observations and the corresponding plaintexts $x_1, \ldots, x_q$, it outputs the key guess k^* that maximizes the mutual information between the observations and the hypothetical leakage values,

$$\mathcal{D}(o_{x_1}, \ldots, o_{x_q}; x_1, \ldots, x_q) \rightarrow k^* \quad \text{iff} \quad \tilde{\mathbf{I}}(\hat{\mathbf{L}}_{k^*}; \mathbf{O}) = \max_{\hat{k}} \tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}). \tag{4}$$

We extend the distinguisher $\mathcal{D}$ defined above to retrieve also the interesting point(s) in time $t = \tau_j$ when the intermediate result $f_k(\cdot)$ is computed. It takes as input the multiset of observed traces $\mathcal{M} = \{o_{x_1}(t), \ldots, o_{x_q}(t)\}$ and the inputs $x_1, \ldots, x_q$. The extended distinguisher is defined by,

$$\begin{aligned} \mathcal{D}(o_{x_1}(t), \ldots, o_{x_q}(t); x_1, \ldots, x_q) \rightarrow (k^*, \tau_j) \quad \text{iff} & \\ \tilde{\mathbf{I}}(\hat{\mathbf{L}}_{k^*}; \mathbf{O}(\tau_j)) = \max_{(\hat{k}, t)} \tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t)). & \end{aligned} \tag{5}$$

Note that there may exist additional points in time where $\mathbf{O}(t)$ (partially) depends on $\mathbf{L}_k$ but where $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_k; \mathbf{O}(t))$ is not maximal. To cover this case we denote τ_j as all instants when $\mathbf{O}(t)$ (partially) depends on $\mathbf{L}_k$.

3.2 Theoretical Reasoning

We consider the Mutual Information between the output of a guessed leakage function $\hat{\mathbf{L}}_{\hat{k}}$ and an observable $\mathbf{O}(t)$, *i.e.* the reduction in the uncertainty on $\hat{\mathbf{L}}_{\hat{k}}$ due to the knowledge of $\mathbf{O}(t)$ for a key hypothesis $\hat{k}$. There exist four interesting combinations of time instants and key candidates to study.

1) incorrect key hypotheses $\hat{k} \neq k$ and wrong time instants $t \neq \tau_j$
In this case $\mathbf{I}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t)) = 0$ because the two variables are independent (see Appendix). However, the equality holds only theoretically. In practice we compute $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t))$ close to 0 as we are working with estimates of entropy.
2) correct key guess $\hat{k} = k$ and wrong time instants $t \neq \tau_j$
In this case $\mathbf{I}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t)) = 0$ because the two variables are independent. Recall that $t \neq \tau_j$ implies independence by definition. Again, in practice we obtain values only close to 0.
3) correct key guess $\hat{k} = k$ and correct instant(s) $t = \tau_j$
In this case $\mathbf{I}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(\tau_j)) = \mathsf{H}(\mathbf{O}(\tau_j)) - \mathsf{H}(\mathbf{O}(\tau_j)|\hat{\mathbf{L}}_{\hat{k}}) > 0$ because the variables are dependent by definition. The value of $\mathsf{H}(\mathbf{O}(\tau_j)|\hat{\mathbf{L}}_{\hat{k}})$ is minimized. So, at the right point(s) in time $t = \tau_j$, the correct key guess $\hat{k} = k$ leads to the highest Mutual Information. In practice, high values of Mutual Information can appear for several points in time if the targeted intermediate result is computed, stored, and reused later. Both facts are empirically confirmed in Sect. 6.
4) incorrect key guess $\hat{k} \neq k$ and correct time instants $t = \tau_j$
In this case $\mathbf{I}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t)) = 0$ if and only if an incorrect key guess leads to random

hypothetical leakage values. In practice we might observe Mutual Information values greater than zero. These “ghost peaks” occur if a wrong key guess does not lead to hypothetical leakage values that are independent of the real leakage values. This phenomenon is also observed for other distinguishers and studied in detail in [2].

4 Practical Aspects of Mutual Information Analysis

In this section we address aspects of Mutual Information Analysis that are of importance for its practical application.

The Mutual Information distinguisher, as most statistical tests, is bounded in its efficiency to recover keys by the hypothetical leakage function $\hat{\mathsf{L}}$. The closer the partition induced by $\hat{\mathbf{L}}_{\hat{k}}$ is to the *a priori unknown* physical data-dependency inherent in $\mathbf{O}(t)$, the more efficient and effective the statistical test will be.

Hence, a side-channel analyst faces several problems which we will summarize using our model’s notation. The flow of information from k to $\mathbf{O}(t)$ has to be examined via the transition caused by $\mathbf{W}$. It involves the hypothetical leakage function $\hat{\mathsf{L}}$ and the electrical properties of the observation channel. The choice of $f_{\hat{k}}(\cdot)$ is usually an easy task and can be performed device independently. Any intermediate result that combines a small part of the (constant) unknown key and a known varying value may be chosen (here “small” means that exhausting all $\hat{k}$ should be feasible). On the other hand, the choice of $\hat{\mathsf{L}}$ as well as the abstraction of the observation channel pose a non-trivial task. Typically, the latter is modeled as a (linear) one-to-one relation (one-to-many if noise is considered) such that the model’s complexity is concentrated in $\hat{\mathsf{L}}$. Based on the choice of $\hat{\mathsf{L}}$ and a key guess $\hat{k}$ an adversary predicts the device’s power dissipation and uses a statistical test to quantify the fitness of her simulation. However, obviously this approach requires an engineer’s insight into the device’s leakage behavior if the goal is to obtain significant results. As long as the target device has been built in standard CMOS technology, this behavior can be *approximated* by the Hamming weight [13] or Hamming distance [2] model. Then, the complexity is shifted to the architecture level as one has to define the exact transition (v_1, v_2) that leaks, which usually involves previously computed values, counters, conditional branches, or memory addresses (*cf.* [2]).

The approach for our attack follows the opposite idea. Instead of crafting an attack for a specific device and implementation, we propose to shift the complexity from the modeling step into the distinguisher. Rather than trying to model the leakage function and the system’s electrical properties as good as possible and measuring the (linear) correlation between the simulated and the observed power dissipation, we propose the following. Assume a one-to-many relation between the leaked and observed values, *i.e.* do not average measurements unless the Gaussian assumption is justified. Assume a suitable leakage function $\hat{\mathsf{L}}$. Compute an estimation of the Mutual Information $\tilde{\mathbf{I}}(\mathbf{L}_{\hat{k}}, \mathbf{O}(t))$ between the hypothetical leakage and the observations and use it as a distinguisher to discriminate keys.

4.1 Hypothetical Side-Channel Leakage

Up to now we assumed that $\hat{\mathsf{L}} = \mathsf{L}$ which reflects a powerful adversary that knows the exact side-channel leakage function of the device under attack. Although this assumption might be justified in some cases, *e.g.* it might be known that the target device leaks the Hamming distance of v_1 and v_2, we relax the assumption and hence also cover cases where the real leakage function is unknown.

There exist two important restrictions for the adversary'y choice of $\hat{\mathsf{L}}$. By assumption, L is a surjective mapping $\mathcal{L} : \mathcal{W} \rightarrow \mathcal{L}\,;\;\; \mathbf{W} \mapsto \mathbf{L} = \mathsf{L}(\mathbf{W})$. The best the adversary can do in order not to deliberately loose information and to ensure that the distinguisher $\mathcal{D}$ works as expected is to ensure that $\hat{\mathsf{L}}$ does not produce collisions where L does not. Since L is unknown, the only way to guarantee this property is to choose $\hat{\mathsf{L}}$ as a bijective mapping of $\mathbf{W}$. Such a setting suggests that $\hat{\mathsf{L}}$ might produce less collisions than L, which makes our distinguisher less efficient but does not tackle its effectiveness.

The second restriction arises due to the generic character of the distinguisher. $\hat{\mathsf{L}}$ must be chosen such that different key hypotheses $\hat{k}$ do not yield a permutation of $\hat{\mathbf{L}}_{\hat{k}}$. If this would happen, $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O})$ would be constant and more important, independent of the guess $\hat{k}$. The distinguisher would not be able to discriminate key candidates.

In the following example, the choice of $\hat{\mathsf{L}}$ does *not* allow to discriminate key candidates using our distinguisher: suppose that E_k is AES encryption and that the targeted transition $\mathbf{W}$ is $(v_1, f_{\hat{k}}(\cdot))$ for a constant reference state $v_1 \in \{0,1\}^n$ and for $f_{\hat{k}}(\cdot)$ being a Sbox lookup during the first round. The AES Sbox is a bijective map. Therefore, different key candidates $\hat{k}$ lead to permutations of the guess $\mathbf{W}_{\hat{k}}$. Choosing $\hat{\mathsf{L}}$ as a bijective map of $\mathbf{W}_{\hat{k}}$ implies that the partition $\{L_i^{\hat{k}}\}_{i=0}^{l}$ is merely permuted, which has no effect on the entropy $\tilde{\mathsf{H}}(\mathbf{O}|\hat{\mathbf{L}}_{\hat{k}})$ and thus no effect on $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O})$. A simple workaround for this problem is to choose $\hat{\mathsf{L}}$ as a bijective map of a subspace of $\mathcal{W}$, *e.g.* one could choose $\hat{\mathbf{L}}_{\hat{k}} :=$ the seven least significant bits of $\mathbf{W}_{\hat{k}}$. In the same context, the DES Sboxes do not lead to a problem since they are not bijective.

Another interesting property of bijective hypothetical leakage functions is, that the sometimes unknown reference state v_1 is transparent to them and can simply be ignored, as long as it is constant.

4.2 Estimation of Probability Densities

In practice, an adversary does not know the probability distributions $\mathbb{P}_{\mathbf{O}|\hat{\mathbf{L}}_{\hat{k}}}$ and $\mathbb{P}_{\mathbf{O}}$ and has to estimate them. Since all successive computations are based on these estimations, the estimation of probability densities is a key issue.

The estimation technique we use relies on histograms. In our experience, it is a simple and efficient technique to address the issue. A histogram estimates the probability distribution of data in a given sample set by counting how many samples fall into a certain bin.

The arising questions are: “How many bins should be used?” and “Should all bins be equally wide?”. As far as we know, there exists no strategy that leads to the best estimation in all scenarios. By applying this technique in numerous side-channel attack scenarios we extracted the following basic guidelines.

1) The first design principle of Mutual Information Analysis is the exploitation of information. We thus aim at estimating the probability distributions as good as possible. This means to use as many bins as there are distinct values in the domain covered by the sample set. This approach may require a limited increase of measurements, but it ensures that no information is lost.
2) Generating histograms is different depending on whether the observations of the random variable are deterministic or probabilistic (noisy). In the deterministic case, we can at least be sure about the value of an observed datum while this does not hold in the probabilistic case.
3) We usually work with bins of equal width. In general, less bins imply less information and vice versa. If we work with noisy observations, choosing less bins may have the effect of noise reduction. In practice this means that several distinct samples can fall into the same bin, which reflects the assumption that they stem from the same datum.

5 Contrasting MIA and Other Distinguishers

In the seminal paper on Differential Power Analysis [9], Kocher *et al.* suggest to use a single-bit partitioning function. In our notation this is the hypothetical leakage function. An advantage of a single-bit approach is that it does not require an assumption on the real leakage function. One merely assumes that different bit values leak differently. A disadvantage is the loss of information due to ignoring all other bits.

The extension to consider several bits at once was first proposed by Messerges *et al.* in [12]. More precisely, the authors proposed to use a partitioning function based on more than one bit and to analyze those atoms that are maximal different (*e.g.* all zeros vs. all ones). However, this approach requires an assumption on the real leakage function to identify those two atoms. Further, it does not allow to exploit the available information in an optimal way, since only to atoms of the partition are considered.

Other methods, *e.g.* the Hamming models, require even more sophisticated assumptions on the real leakage function and try to estimate it as good as possible. A disadvantage of these methods is, that they can only be applied if the assumptions are justified.

Independent of single- or multi-bit partitioning functions, an adversary can choose amongst several distinguishers. Kocher et *al.* suggested the difference of means test. Later publications suggested further distinguishers including the t-test [1] and the Pearson correlation coefficient [2]. These distinguishers analyze a probability distribution at most by its mean and variance (Gaussian assumption). Hence they do not exploit all information available and are inappropriate if the Gaussian assumption does not hold. Pearson's correlation coefficient requires the additional assumption of a linear relation between leakage and observation.

Template Attacks [3] are a different kind of attack. They assume a powerful adversary that fully controls a training device which is used to estimate the probability densities of the physical observable for each $L_i^{\hat{k}}$. Effectively this is equivalent to knowing L. For a side-channel measurement from a target device, the maximum-likelihood test derives which previously estimated probability density is the most likely origin of the sample. Template attacks constitute the strongest form of side-channel attacks, if the Gaussian assumption holds. A disadvantage of the approach is the need for a training device.

In contrast, MIA requires neither a training device, nor a restrictive assumption about the real leakage function, nor the Gaussian assumption. MIA estimates the full probability density for each $L_i^{\hat{k}}$ from observations of the target device's leakage. Due to the lack of reference data, *e.g.* templates, MIA cannot apply the maximum-likelihood test. Instead, MIA uses our information-based distinguisher. An important advantage of MIA is, that it can exploit arbitrary relationships between $\mathbf{L}_k$ and $\mathbf{O}$.

The work of Standaert *et al.* [17] is different from ours in the following sense: they propose a Mutual Information-based metric for measuring an amount of side-channel leakage. That is, they do not propose an attack but a leakage analysis/evaluation tool.

6 Experimental Results for Mutual Information

In this section, we apply the theoretical framework from Sect. 2 and 3 and provide experimental results based on power measurements from an AT90S8515 micro controller ($n = 8$ bit) performing $E_k :=$ AES-128 encryption in software[2]. The measurements $\mathbf{O}(t)$ represent the voltage drop over a 50Ω resistor inserted in the smart card's ground line. We sample the power consumption at instants $t = 1, \ldots, 1800$ during the first round of the AES-128 encryption of randomly chosen plaintexts with a constant key. Our experiments focus on the first key byte denoted by $\mathbf{K} \in \{0,1\}^8$ and the first plaintext byte denoted by $\mathbf{X} \in \{0,1\}^8$.

6.1 Mutual Information Applied to Side Channel Leakage

We empirically confirm that Mutual Information Analysis is indeed effective using relaxed assumptions with the following experiment:

- population size $q = 1000$ power curves $o_{x_i}(t)$, $i = 1, \ldots, q$
- $f_{\hat{k}}(\mathbf{X}) = \mathrm{Sbox}(\mathbf{X} \oplus \hat{k})$, $\mathbf{W}_{\hat{k}} = (v_1, f_{\hat{k}}(\mathbf{X}))$, v_1 constant and unknown
- $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}}) :=$ the r^{th} bit of $f_{\hat{k}}(\mathbf{X})$, where $r = 0$ denotes the LSB.

Hence, each $o_{x_i}(t)$ is associated to an atom of $\{L_i^{\hat{k}}\}_{i=0}^{1}$ by $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}})$ which is the rth bit of $\mathrm{Sbox}(\mathbf{X} \oplus \hat{k})$. For the dependence between leaked value and observed power dissipation we assume a one-to-many relation due to noise, *i.e.* each distinct value of $\mathbf{L}_k$ leads to exactly one power consumption value under noise-free conditions,

[2] We would like to point out that the AES encryption terminates in constant time.

but in reality it might lead to differing observations. The probability densities $\tilde{\mathbb{P}}_{\mathbf{O}}$ and $\tilde{\mathbb{P}}_{\mathbf{O}|\hat{\mathbf{L}}_{\hat{k}}}$ are empirically determined by sampling the distributions of $\hat{\mathbf{L}}_{\hat{k}}$ and $\mathbf{O}$ with histograms. The number of bins for the histograms is chosen according to size of $\mathcal{L}$, *i.e.* the number of distinguishable values in $\hat{\mathbf{L}}_{\hat{k}}$, which is two.

We compute the Mutual Information $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}; \mathbf{O}(t))$ for $\hat{k} = k$ according to Eq. (1) and (3) for each t. Figure 2 shows the resulting Mutual Information traces for $r = 0, 1, 2$. The obvious peaks in the upper plot ($r = 0$) appear during the jointly implemented SubBytes and ShiftRows operations as well as during the MixColumn operation, which involve the targeted value $\mathbf{W}_{\hat{k}}$ several times. These peaks are less significant in the plots for $r = 1$ and $r = 2$ which clearly indicates that the single bits leak different amounts of information.

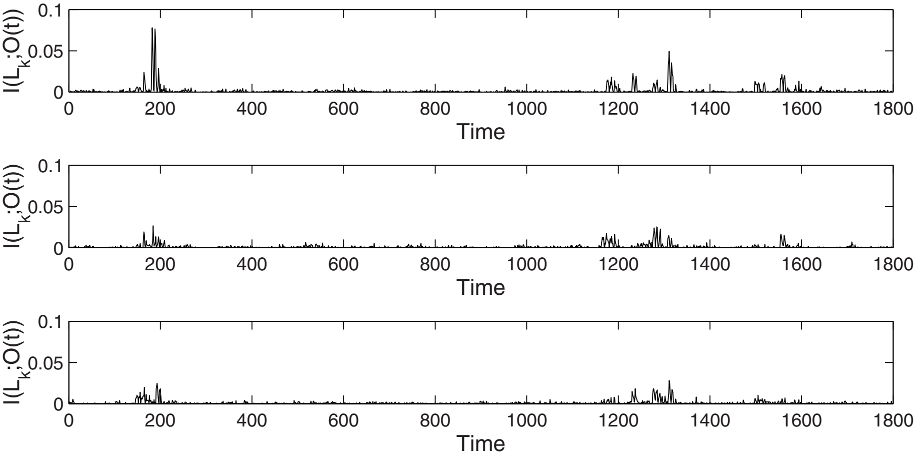

Fig. 2. Mutual Information of 1-bit leakages for bit $r = 0, 1, 2$, from top to bottom

However, the information leaked adds up as shown in Fig. 3 which depicts the Mutual Information trace of the 2-bit leakage function $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}}) :=$ the two LSBs of $f_{\hat{k}}(\mathbf{X})$. We used four bins to estimate the probability distributions.

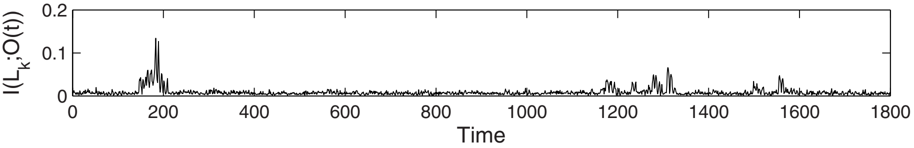

Fig. 3. Mutual Information for 2-bit leakages

6.2 Empirical Evidence

This section provides empirical evidence showing that the attack and the distinguisher are effective and hence confirms the theoretical considerations of

Sect. 3.2. We empirically verify that our distinguisher is effective in our general setting with the following experiment:

- population size $q = 1000$ power curves $o_{x_i}(t)$, $i = 1, \ldots, q$
- $f_{\hat{k}}(\mathbf{X}) = \text{Sbox}(\mathbf{X} \oplus \hat{k})$, $\mathbf{W}_{\hat{k}} = (v_1, f_{\hat{k}}(\mathbf{X}))$, v_1 constant and unknown
- $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}}(\cdot)) :=$ the three MSBs of $f_{\hat{k}}(\mathbf{X})$.

As before, we assume a one-to-many relation between leaked and observed values due to noise. Based on a key guess $\hat{k} \in \mathcal{K}$, each $o_{x_i}(t)$ is associated to an atom of $\{L_i^{\hat{k}}\}_{i=0}^{7}$ by $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}})$ which is equal to the three MSBs of $\text{Sbox}(\mathbf{X} \oplus \hat{k})$. We estimate the probability distributions $\mathbb{P}_{\mathbf{O}}$ and $\mathbb{P}_{\mathbf{O}|\hat{\mathbf{L}}_{\hat{k}}}$ with histograms for which we use eight bins and compute the Mutual Information of $\hat{\mathbf{L}}_{\hat{k}}$ and $\mathbf{O}(t)$ according to Eq. (1) and (3). Figure 4 depicts the resulting Mutual Information trace for the correct key guess $\hat{k} = k$.

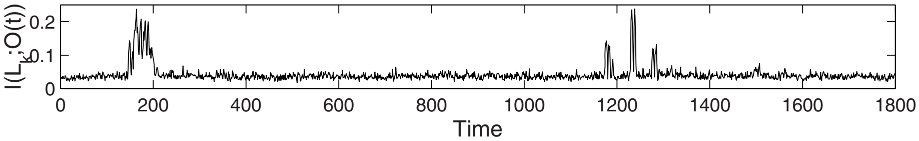

Fig. 4. Mutual Information over time for the correct key hypothesis

As can be seen when comparing to Fig. 2 and 3 the trace shows clear peaks at the points of interest $t = \tau_j$ where the targeted intermediate result $f_{\hat{k}}(\cdot)$ is processed. Next, we compute the same Mutual Information trace for all other key hypotheses $\hat{k}$ and test, if the highest derived Mutual Information value for any wrong $\hat{k}$ is lower than the one for $\hat{k} = k$. More formally that is: $\text{argmax}_{t,\hat{k}=k}$ $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}, \mathbf{O}(t)) > \text{argmax}_{t,\hat{k}\neq k}$ $\tilde{\mathbf{I}}(\hat{\mathbf{L}}_{\hat{k}}, \mathbf{O}(t))$. Figure 5 shows the highest Mutual Information value (selected from the whole time frame t) for every key hypothesis. The peak for the correct key hypothesis $\hat{k} = k$ is clearly distinguishable.

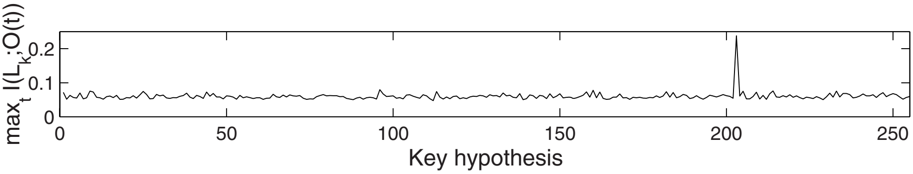

Fig. 5. Maximum Mutual Information per key hypothesis

6.3 MIA and Dual Rail Precharge Logic

In this section we apply our distinguisher in a scenario for which it seems particularly promising: special logic, *e.g.* Wave Dynamic Differential Logic (WDDL)

[18], designed to resist differential side-channel analysis. While the assumption of the Hamming weight or distance leakage function is justified and leads to efficient attacks against devices implemented in standard CMOS, the situation is very different when facing dual rail precharge (DRP) logic. Let us look back to the initial foundations of those models. In standard CMOS, the instantaneous dynamic power dissipation of a logic or sequential gate is directly linked to whether the bit-pattern on the inputs lead to a transition (bit-flip) in the gate and/or on the output wire(s) or not. Although the energy required to perform the flip is not equal amongst cell types and not even amongst equal cells spread over the silicon area with process variations, the differences are usually negligible in a power analysis attack context. In particular, this is the case for sets of gates that drive large capacitive loads, *e.g.* bus lines in a microcontroller. This is why attacking a microcontroller when performing a memory lookup instruction with a correlation attack and the Hamming weight or distance leakage model can lead to correlation coefficients of almost one.

However, these models and assumptions do not hold for DRP logic. The fundamental idea of DRP logic is to encode one bit of information in a differential pair, *e.g.* 0 = (0,1) and 1 = (1,0), that is signaled over a wire pair. Further, the entire circuit is precharged to a constant pair (0,0) or (1,1) in the first half of each clock cycle. During the evaluation phase, the second half of each clock cycle, the logic evaluates and each wire pair takes either (0,1) or (1,0) depending on the bit value that is encoded. Doing so ensures that, whether the logical input to a gate changes or not, the gate performs exactly one bit flip in each evaluation phase. Obviously, hypothetical leakage functions that relate to the *number* of "logical" bit flips only are meaningless in this context. The circuit performs a constant number of bit flips per cycle, independent of the logical data values. Still, DRP logic leaks information. The relatively small differences that we neglected in the standard CMOS context now have a major impact.

For simplicity, consider two gates in DRP logic that each drive two differential outputs with capacities (α, β) and (γ, δ). If $\alpha > \beta$ and $\gamma > \delta$ holds, the Hamming models do not describe the power dissipation well, but they will work because the direction of the differential is the same for both logical output bits. The same holds if we replace $>$ with $<$. In the case that the directions of the differentials are not equal, the Hamming models no longer represent effective estimators of power dissipation behavior and side-channel leakage. This discrepancy increases with the number of logical bits, starting from two bits.

Since our distinguisher does not rely on a restrictive assumption about the leakage function, it is the method of choice for an attack against DRP logic. We confirm the correctness of this statement with empirical evidence.

The experimental platform is an 8051 microcontroller implemented in a DRP variant with differentially routed wire pairs. We implemented a simple yet representative test program which consists of a single table lookup of the Sbox S1 of the Data Encryption Standard. We obtained power measurements while the microcontroller performed lookups for randomly chosen plaintexts and a constant key. Before each measurement, the memory bus and the target register were reset to zero. Thus the previous state is zero. The measurements represent the voltage drop over a 50Ω resistor inserted in the microcontrollers V_{DD} line. We sampled the voltage drop at a rate of 2GS/s.

Let $\mathbf{X}$ be the six plaintext bits and k be the 6-bit subkey. Further experimental settings are:

- population size $q = 100\,000$ power curves $o_{x_i}(t)$, $i = 1, \ldots, q$
- $f_{\hat{k}}(\mathbf{X}) = \mathrm{S1}(\mathbf{X} \oplus \mathrm{k})$, $\mathbf{W}_{\hat{k}} = (\upsilon_1, f_{\hat{k}}(\mathbf{X}))$, $\upsilon_1 = 0$
- $\hat{\mathsf{L}}(\mathbf{W}_{\hat{k}}(\cdot)) :=$ all four bits of $f_{\hat{k}}(\mathbf{X})$.

Figure 6 shows the result of a standard correlation attack, where we used the Hamming weight of $\mathrm{S1}(\mathbf{X}\oplus\hat{k})$ as the predicted power dissipation. The correlation trace for the correct key is plotted in black, for all other key candidates in gray. As can be seen, the correct key hypothesis does not lead to a maximal or minimal correlation coefficient with respect to the whole period.

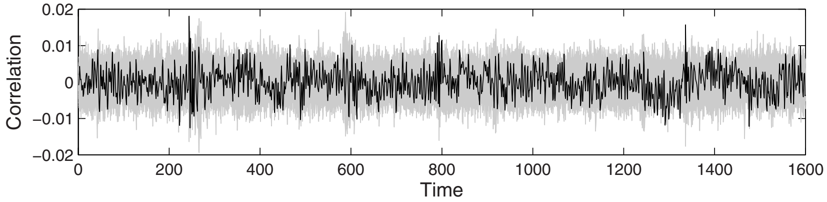

Fig. 6. Correlation traces: correct key in black, all other in gray

Figure 7 on the other hand shows the result of an attack with the Mutual Information-based distinguisher. The Mutual Information trace for the correct key is plotted in black, for all other key candidates in gray. At time index 600 the correct key hypothesis leads to a Mutual Information value that is maximal for all key hypotheses and the whole time.

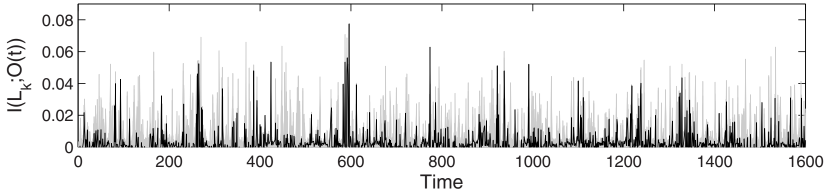

Fig. 7. Mutual Information traces, correct key in black, all other in gray

To emphasize the difference, we present in Fig. 8 plots of the maximal and minimal correlation values as well as of the maximal Mutual Information values per key hypothesis, chosen from the overall time frame.

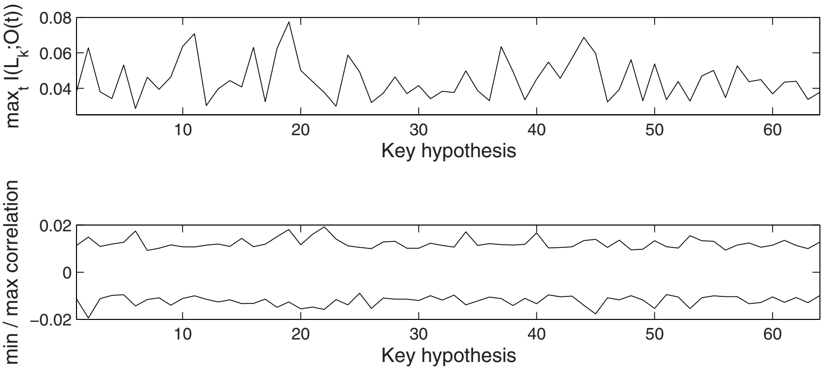

Fig. 8. Maximum Mutual Information per key hypothesis (upper plot); Maximal and minimal correlation coefficient per key hypothesis (lower plot)

7 Conclusion

We described a generic differential side-channel attack that is based on an information-theoretic distinguisher. The distinguisher uses the Mutual Information between the observed measurements and the values of a hypothetical leakage function to rank key guesses. We showed why the attack is particularly promising when the target device is implemented in dual rail precharge logic. The effectiveness of our approach is confirmed by results of power analysis experiments.

Acknowledgements

The authors would like to thank J.R. Rao for shepherding this paper, François-Xavier Standaert for fruitful discussions, and the anonymous reviewers of CHES 2007, AsiaCrypt 2007, Eurocrypt 2008, and CHES 2008 for their insightful comments. Most of them helped to improve the paper.

The information in this document reflects only the authors' views, is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

References

1. Aigner, M., Oswald, E.: Power Analysis Tutorial, `http://www.iaik.tugraz.at/aboutus/people/oswald/papers/dpa_tutorial.pdf`
2. Brier, E., Clavier, C., Olivier, F.: Correlation power analysis with a leakage model. In: Joye, M., Quisquater, J.-J. (eds.) CHES 2004. LNCS, vol. 3156, pp. 16–29. Springer, Heidelberg (2004)

3. Chari, S., Rao, J.R., Rohatgi, P.: Template attacks. In: Kaliski Jr., B.S., Koç, Ç.K., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 172–186. Springer, Heidelberg (2003)
4. Coron, J.-S., Goubin, L.: On Boolean and Arithmetic Masking against Differential Power Analysis. In: Paar, C., Koç, Ç.K. (eds.) CHES 2000. LNCS, vol. 1965, pp. 231–237. Springer, Heidelberg (2000)
5. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & Sons, Chichester (2006)
6. Golić, J.D., Tymen, C.: Multiplicative masking and power anaylsis of AES. In: Kaliski Jr., B.S., Koç, Ç.K., Paar, C. (eds.) CHES 2002. LNCS, vol. 2523, pp. 31–47. Springer, Heidelberg (2002)
7. Goubin, L.: A sound method for switching between boolean and arithmetic masking. In: Koç, Ç.K., Naccache, D., Paar, C. (eds.) CHES 2001. LNCS, vol. 2162, pp. 3–15. Springer, Heidelberg (2001)
8. Kocher, P.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS and other systems. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996)
9. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M.J. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
10. Messerges, T.S.: Securing the AES finalists against power analysis attacks. In: Schneier, B. (ed.) FSE 2000. LNCS, vol. 1978. Springer, Heidelberg (2001)
11. Messerges, T.S.: Using second-order power analysis to attack DPA resistant software. In: Koç, Ç.K., Paar, C. (eds.) CHES 2000. LNCS, vol. 1965, pp. 238–251. Springer, Heidelberg (2000)
12. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Investigations of power analysis attacks on smartcards. In: WOST 1999: Proceedings of the USENIX Workshop on Smartcard Technology on USENIX Workshop on Smartcard Technology, Berkeley, CA, USA, p. 17. USENIX Association (1999)
13. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examining smart-card security under the threat of power analysis attacks. IEEE Trans. Comput. 51(5), 541–552 (2002)
14. Micali, S., Reyzin, L.: Physically observable cryptography. In: Naor, M. (ed.) TCC 2004. LNCS, vol. 2951, pp. 278–296. Springer, Heidelberg (2004)
15. Quisquater, J.-J., Samyde, D.: ElectroMagnetic Analysis (EMA): Measures and Couter-Measures for Smard Cards. In: Attali, I., Jensen, T.P. (eds.) E-smart 2001. LNCS, vol. 2140, pp. 200–210. Springer, Heidelberg (2001)
16. Shamir, A., Tromer, E.: Acoustic cryptanalysis, http://theory.csail.mit.edu/~tromer/acoustic/
17. Standaert, F.-X., Malkin, T.G., Yung, M.: A formal practice-oriented model for the analysis of side-channel attacks. Cryptology ePrint Archive, Report 2006/139 (2006), http://eprint.iacr.org/
18. Tiri, K., Hwang, D., Hodjat, A., Lai, B.-C., Yang, S., Schaumont, P., Verbauwhede, I.: Prototype IC with WDDL and differential routing - DPA resistance assessment. In: Rao, J.R., Sunar, B. (eds.) CHES 2005. LNCS, vol. 3659, pp. 354–365. Springer, Heidelberg (2005)

Appendix

Let $\mathbf{X}$ be a random variable on a (discrete) space $\mathcal{X}$ with probability distribution $\mathbb{P}_{\mathbf{X}}$. The uncertainty that one has about the value of such a random variable when an experiment is performed, is expressed by the Shannon entropy of $\mathbf{X}$ which is usually denoted by $\mathsf{H}(\mathbf{X})$ or $\mathsf{H}(\mathbb{P}_{\mathbf{X}})$. It is defined by the following equation

$$\mathsf{H}(\mathbf{X}) = -\sum_{x \in \mathcal{X}} \mathbb{P}_{\mathbf{X}}[\mathbf{X} = x] \log_2 \mathbb{P}_{\mathbf{X}}[\mathbf{X} = x]\,. \tag{6}$$

$\mathsf{H}(\mathbf{X})$ expresses the uncertainty in bits. The entropy of the pair of random variables $(\mathbf{X}, \mathbf{Y})$ (where $\mathbf{Y}$ is a random variable on a space $\mathcal{Y}$) is denoted by $\mathsf{H}(\mathbf{X}, \mathbf{Y})$ and it expresses the uncertainty one has about both. We note that the entropy of two random variables is sub-additive *i.e.*

$$\mathsf{H}(\mathbf{X}, \mathbf{Y}) \leq \mathsf{H}(\mathbf{X}) + \mathsf{H}(\mathbf{Y}) \tag{7}$$

with equality if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent. Often one is interested in the uncertainty about $\mathbf{X}$ given that one has obtained the outcome of an experiment on a related random variable $\mathbf{Y}$. This is expressed by the conditional entropy $\mathsf{H}(\mathbf{X}|\mathbf{Y})$ which is defined as follows,

$$\mathsf{H}(\mathbf{X}|\mathbf{Y}) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}_{\mathbf{X},\mathbf{Y}}[\mathbf{X} = x, \mathbf{Y} = y] \log_2 \mathbb{P}_{\mathbf{X}|\mathbf{Y}}[\mathbf{X} = x|\mathbf{Y} = y], \tag{8}$$

where $\mathbb{P}_{\mathbf{X},\mathbf{Y}}$ denotes the joint probability distribution of $\mathbf{X}$ and $\mathbf{Y}$ and $\mathbb{P}_{\mathbf{X}|\mathbf{Y}}$ stands for the conditional probability distribution of $\mathbf{X}$ given $\mathbf{Y}$. When $\mathbf{Y}$ can be considered as an observation of $\mathbf{X}$ over a noisy channel, then one often characterizes the channel by its set of conditional distributions $\{\mathbb{P}_{\mathbf{Y}|\mathbf{X}=x}\}_{x \in \mathcal{X}}$.