Leakage Certification Revisited: Bounding Model Errors in Side-Channel Security Evaluations

Olivier Bronchain*, Julien M. Hendrickx*, Clément Massart*,
Alex Olshevsky†, François-Xavier Standaert*

* ICTEAM Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium.
† Dep. of Electrical and Computer Engineering, Boston Univ., Massachussetts, USA.

Abstract. Leakage certification aims at guaranteeing that the statistical models used in side-channel security evaluations are close to the true statistical distribution of the leakages, hence can be used to approximate a worst-case security level. Previous works in this direction were only qualitative: for a given amount of measurements available to an evaluation laboratory, they rated a model as "good enough" if the model assumption errors (i.e., the errors due to an incorrect choice of model family) were small with respect to the model estimation errors. We revisit this problem by providing the first quantitative tools for leakage certification. For this purpose, we provide bounds for the (unknown) *Mutual Information* metric that corresponds to the true statistical distribution of the leakages based on two easy-to-compute information theoretic quantities: the *Perceived Information*, which is the amount of information that can be extracted from a leaking device thanks to an estimated statistical model, possibly biased due to estimation and assumption errors, and the *Hypothetical Information*, which is the amount of information that would be extracted from an hypothetical device exactly following the model distribution. This positive outcome derives from the observation that while the estimation of the Mutual Information is in general a hard problem (i.e., estimators are biased and their convergence is distribution-dependent), it is significantly simplified in the case of statistical inference attacks where a target random variable (e.g., a key in a cryptographic setting) has a constant (e.g., uniform) probability. Our results therefore provide a general and principled path to bound the worst-case security level of an implementation. They also significantly speed up the evaluation of any profiled side-channel attack, since they imply that the estimation of the Perceived Information, which embeds an expensive cross-validation step, can be bounded by the computation of a cheaper Hypothetical Information, for any estimated statistical model.

1 Introduction

State-of-the-art. Side-Channel Attacks (SCAs) are among the most important threats against the security of modern embedded devices [20]. They leverage physical leakages such as the power consumption or electromagnetic radiation of an implementation in order to recover sensitive

data. Concretely, SCAs consist in two main steps: information extraction and information exploitation. In the first step, the adversary collects partial information about some intermediate computations of the leaking implementation. For this purpose, he generally compares key-dependent leakage models with actual measurements thanks to a distinguisher such as the popular Correlation Power Analysis (CPA) [2] or Template Attacks (TAs) [5]. In the second step, the adversary combines this partial information in order to recover the sensitive data in full (e.g., by performing a key recovery). For this purpose, the most frequent solution is to exploit a divide-and-conquer strategy (e.g., to recover each key byte independently), and to perform key enumeration if needed [22, 27, 34].[1]

Based on this description, the (worst-case) security evaluation of actual implementations and side-channel countermeasures requires estimating the amount of information leaked by a target device [33]. Fair evaluations ideally require exploiting a perfect leakage model (i.e., a model that perfectly corresponds to the leakage distribution) with a Bayesian distinguisher. Yet, such a perfect leakage model is in general unknown. Therefore, side-channel security evaluators (and adversaries) have to approximate the statistical distribution of the leakages using density estimation techniques. It raises the problem that security evaluations can become inaccurate due to estimation and assumption errors in the leakage model. Estimation errors are due to an insufficient number of measurements for the model parameters to converge. Assumption errors are due to incorrect choices of density estimation tools (e.g., assuming Gaussian leakages for non-Gaussian leakages).

The problem of ensuring that a leakage model is "good enough" so that it does not lead to over-estimating the security of an implementation has been formalized by Durvaux et al. as *leakage certification* [13]. In the first leakage certification test introduced at Eurocrypt 2014, a leakage model is defined as good enough if its assumption errors are small with respect to its estimation errors. Intuitively, it guarantees that given the amount of measurements used by the evaluator / adversary to estimate a model, any improvement of his (possibly incorrect) assumptions will not lead to noticeable degradations of the security level (since the impact of improved assumptions will be hidden by estimation errors). In a heuristic simplification proposed at CHES 2016, a model is considered as good enough if the statistical moments of the model do not notice-

[1] More advanced strategies, such as Algebraic Side-Channel Attacks (ASCA) [29] or Soft Analytical Side-Channel Attacks (SASCA) [35] can also be considered. Our following tools apply identically to these attacks.

ably deviate from the statistical moments of the actual leakage distribution [12]. In both cases, the certification tests are based on challenging the model against fresh samples in a cross-validation step. In both cases, the certification tests are qualitative and conditional to the number of measurements available to build the model. By increasing the number of measurements (and if the model is imperfect), one can make estimation errors arbitrarily small, which inevitably leads to the possible detection of assumption errors. As a result, a fundamental challenge in side-channel security evaluations (which we tackle in this paper) is to *bound the information loss due to model errors quantitatively.*

We note that from an information theoretic viewpoint, the risk of under-estimating the leakages due to model errors in side-channel security evaluations can be captured with the notion of Perceived Information (PI) initially introduced in [30] to analyze model variability in nanoscale devices. Informally, the PI corresponds to the amount of information that can be extracted from some data thanks to a statistical model possibly affected by estimation or assumption errors. If the model is perfect, the PI is identical to Shannon's standard definition of Mutual Information (MI). Otherwise, the difference between the MI and the PI provides a quantitative view of the information loss. (Yet, at this stage not a usable one since the MI is unknown, just as the perfect model).

Contribution. The main contributions of the paper are to provide simple and efficient information theoretic tools in order to bound the model errors in side-channel security evaluations, and to validate these tools empirically based on simulated leakages and actual measurements.

Our starting point for this purpose is a third information theoretic quantity that was introduced as part as a negative result on the way towards the CHES 2016 heuristic leakage certification test. Namely, the Hypothetical Information (HI), which is the amount of information that would be extractable from the samples if the true distribution was the statistical model. As discussed in [12], as such the HI seems useless since in case of incorrect model, it can be completely disconnected from the true leakage distribution (i.e., models with positive HI may not lead to successful attacks). Yet, we show next how it can be used in combination with the PI in order to enable quantitative leakage certification. In particular, our main results in this direction are twofold:

First, we show that – *under the assumption that the target random variable (e.g., the secret key) has constant (e.g., uniform) probability* – the empirical HI (eHI), which corresponds to the HI estimated directly based on the empirical leakage distribution, is in expected value an upper

bound for the MI and that it converges monotonically towards the true MI as the number of measurements used in order to estimate the leakage model increases. Second, we show that (under the same assumptions) the PI is a lower bound for the MI.

Our experiments then show that these tools can be concretely exploited in the analysis of actual leakage models and speed up side-channel security evaluations. They also sometimes illustrate the difficulty to obtain tight worst-case bounds in practice, and the interest of exploiting some additional (e.g., Gaussian) leakage assumptions in order to more efficiently obtain "close to worst-case" evaluations. In this case as well, we show that bounding the PI with the HI can lead to efficiency gains, especially for distributions with larger number of dimensions.

Related works. The fact that we may bound the MI is surprising since it is actually known to be impossible in general. As for example discussed by Paninski [26], there are no unbiased estimators for the MI (and the rate at which the error decreases depends on the data structure, for any estimator). This had led some works aiming at leakage detection to exploit more positive results for the distribution of the zero MI (i.e., the case with no information leakage) [6, 7, 24]. We follow a different path by observing that in the context of side-channel security evaluations, every key (or target intermediate variable) has a uniform distribution a priori, and it is easy for the evaluator to enforce that the number of leakages collected for every key (or target intermediate variable) is identical. In this case, where the probability of the key (or target intermediate variable) does not need to be estimated, we fall back on a situation where the maximum likelihood estimation of the MI is biased upwards everywhere. Combined with the good properties of the empirical distribution (which converges towards the true distribution) it leads to our first result. The result for the PI is even more direct, holds for any model, and is obtained by solving an optimization problem.

Besides, the problem of leakage certification shares strong similarities with the application of the bias-variance decomposition [9], introduced as a diagnosis tool for the evaluation of side-channel leakage models by Lerman et al. [18]. Note that we here mean the bias (and variance) of the leakage model, not the bias of the MI estimator as when previously referring to Paninski. Conceptually, evaluating the bias and variance of a leakage model can be viewed as similar to evaluating its estimation and assumption errors. Yet, the problem of this decomposition is again that it requires the knowledge of the perfect leakage model. Lerman et al. alleviate this difficulty by assuming that the perfect leakage model directly

provides the key (in one trace). However, this leads their estimation of the bias and variance to gradually become inaccurate as the target implementations become protected, so that this idealizing assumption becomes more and more incorrect.

2 Notations and background

In this section, we provide the background and definitions needed to describe our results, with a particular focus on the different metrics we suggest for side-channel security evaluations.

True distributions. Given a (discrete) secret key variable K and a (discrete or continuous) leakage variable L, we denote the true conditional Probability Mass Function (PMF) – which corresponds to discrete leakages – as $\Pr(L = l|K = k)$ and the true conditional Probability Density Function (PDF) – which corresponds to continuous leakages – as $\mathsf{f}(L = l|K = k)$.

Mutual Information (MI). For discrete leakages, it is defined as [8]:

$$\mathrm{MI}(K;L) = \mathrm{H}(K) + \sum_{l\in\mathcal{L}} \Pr(L=l) \cdot \sum_{k\in\mathcal{K}} \Pr(K=k|L=l) \cdot \log_2 \Pr(K=k|L=l), \quad (1)$$

$$= \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \Pr(K=k) \cdot \sum_{l\in\mathcal{L}} \Pr(L=l|K=k) \cdot \log_2 \Pr(K=k|L=l). \quad (2)$$

Using the simplified notation $\Pr(X = x) := \mathsf{p}(x)$, it leads to:

$$\mathrm{MI}(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|k) \cdot \log_2 \mathsf{p}(k|l). \quad (3)$$

Assuming uniformly distributed keys, $\mathsf{p}(k|l)$ is computed as $\frac{\mathsf{p}(l|k)}{\sum_{k^*\in\mathcal{K}} \mathsf{p}(l|k^*)}$ and $\mathrm{H}(K) = \log_2(|\mathcal{K}|)$. Similarly, in the case of continuous leakages, we can define the MI as follows:

$$\mathrm{MI}(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \Pr(k) \cdot \int_{l\in\mathcal{L}} \mathsf{f}(l|k) \cdot \log_2 \mathsf{p}(k|l)\ dl. \quad (4)$$

MI and statistical inference attacks. We are interested in the MI in the context of side-channel analysis because it is a good predictor of the success probability of a continuous "statistical inference attack", where an adversary uses his leakages in order to recover a secret key.[2] Precisely,

[2] We consider so-called noisy leakages, where the adversary can observe a noisy function of secret variables [28].

it is shown in [11] that a higher MI generally implies a more efficient maximum likelihood attack where the adversary selects the most likely key $\tilde{k}$ among all the candidates k^* as:

$$\tilde{k} = \underset{k^* \in \mathcal{K}}{\operatorname{argmax}} \prod_{l \in \mathcal{L}} \mathsf{p}(k^*|l). \tag{5}$$

Note that this implication only holds independently for each key k manipulated by the leaking device. That is, a higher "MI per key" $\mathrm{MI}(k; L)$ implies a higher probability of success $\Pr(\tilde{k} = k)$.

Intuitively, the link between such an attack and $\mathrm{MI}(k; L)$ comes from the similarity between the product of probabilities in the attack and the sum of log probabilities in the metric.

Sampling process. The true distributions are generally unknown, but we can sample them in order to produce data sets for estimating leakage models and testing these models. We denote these sampling processes as $\mathcal{M} \xleftarrow{n} \mathsf{p}(l|k)$ and $\mathcal{T} \xleftarrow{n_t} \mathsf{p}(l|k)$ in the discrete case, with n and n_t (resp., $n(k)$ and $n_t(k)$) the number of i.i.d. samples measured and stored (resp., per key) in the multisets of samples $\mathcal{M}$ and $\mathcal{T}$ (which have repetitions). We replace p by f for the continuous case.

Computing the MI by sampling. The MI metric can be computed directly thanks to Equations 3 or 4. It can also be computed "by sampling" (for discrete and continuous leakages) as:

$$\widehat{\mathrm{MI}}(K; L) = \mathrm{H}(K) + \sum_{k \in \mathcal{K}} \mathsf{p}(k) \cdot \sum_{i=1}^{n_t(k)} \frac{1}{n_t(k)} \cdot \log_2 \mathsf{p}(k|l_k(i)), \tag{6}$$

where $l_k(i) \in \mathcal{T}$ is the ith leakage sample observed for the key k. In the discrete case, it is easy to see that the blue part of the equation corresponds to the empirical distribution. So Equation 6 essentially replaces the true distribution $\mathsf{p}(l|k)$ by the empirical one, and the hat sign is used to reflect that the MI is computed by sampling. Since the empirical distribution converges towards the real one as $n_t \to \infty$, $\widehat{\mathrm{MI}}(K; L)$ also tends towards $\mathrm{MI}(K; L)$. In the continuous case, the convergence requires more elaboration (details are given in the full version of the paper [3]). For simplicity, we next refer to the blue part of Equation 6 as the empirical in both the discrete and continuous cases.

Note that the PMF after the log in Equation 6 is fixed (i.e., it is not an estimate). So this equation does not describe an estimation of the MI

in the usual sense, where the joint probability of two random variables has to be estimated: it only provides an alternative way to compute the MI of some known distribution. Hence it does not suffer from the bias issues discussed in [26].

Model estimation. Given a set of n modeling samples $\mathcal{M}$, we denote the process of estimating the conditional leakage distribution as $\tilde{\mathsf{m}}_n(l|k) \leftarrow \mathcal{M}$, where we use the red color to highlight the model and the tilde sign to reflect that it is the result of a statistical estimation.

We will consider two types of models: *exhaustive models* where we directly estimate the empirical distribution (e.g., in the discrete case they correspond to histograms on the full support of the observations); *simplified models* which may for example correspond to histograms with reduced numbers of bins in the discrete case, or to parametric (e.g., Gaussian) PDF estimation in the continuous case. Simplified models are aimed to converge faster (i.e., to require lower n values before becoming informative), possibly at the cost of some information loss when $n \to \infty$. In other words, exhaustive models (sometimes slowly) converge towards the real distribution as $n \to \infty$, while simplified models may be affected by assumption errors appearing for large n's (i.e., bad choices of parametric estimation such as assuming Gaussian noise for non-Gaussian leakages).

Finally, we use the term *model* for the (parametric or non-parametric) estimation of a distribution from a given number of profiling leakages n, and the term *model family* for the set of all the models that can be produced with a defined set of parameters. For example, the (univariate) Gaussian model family denotes all the models that can be produced by estimating a sample mean and a sample variance, and a Gaussian model corresponds to one estimation given n leakages.

Hypothetical and Perceived information. Given that the true distributions $\mathsf{p}(l|k)$ or $\mathsf{f}(l|k)$ are unknown, we cannot directly compute the MI. One option to get around this impossibility is to estimate it, which is known to be a hard problem (i.e., there are no unbiased and distribution-independent estimators [26]). We next study an alternative approach which is to analyze the information that is revealed by estimated models thanks to two previously introduced and easy-to-compute quantities. First the *Perceived Information* (PI), which is the amount of information that can be extracted from some data thanks to an estimated model, possibly affected by estimation or assumption errors [13]. Second the *Hypothetical Information* (HI), which is the amount of information that would be revealed by (hypothetical) data following the model distribution [12].

Informally, the PI predicts the concrete success probability of a maximum likelihood attack exploiting an estimated model just as the (unknown) MI predicts the theoretical success probability of a worst-case maximum likelihood attack exploiting the true leakage distribution [15]. It can be negative if the estimated model is too different from the true distribution, and therefore can underestimate the information available in the leakages. By contrast, the HI is a purely hypothetical value that is always non-negative and can therefore overestimate the information available in the leakages. We next aim to formalize their properties, and in particular to show that they can be used to (lower and upper) bound the worst-case security level captured by the unknown MI.

The HI is defined as follows in the discrete case:

$$\mathrm{HI}_n(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{l\in\mathcal{L}} \tilde{\mathsf{m}}_n(l|k) \cdot \log_2 \tilde{\mathsf{m}}_n(k|l). \tag{7}$$

(Replace $\sum$ by $\int$ in the continuous case) For an estimated model $\tilde{\mathsf{m}}_n(l|k)$, the HI can be computed based on Equation 7, or by sampling (just as for the MI). In the latter case, we use the notation $\widehat{\mathrm{HI}}_n(K;L)$:

$$\widehat{\mathrm{HI}}_n(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{i=1}^{n_t(k)} \frac{1}{n_t(k)} \cdot \log_2 \tilde{\mathsf{m}}_n(k|l_k(i)), \tag{8}$$

with as main difference from the MI case that the test samples come from a set $\mathcal{T}_{\mathrm{m}}$ which has been picked up from the model distribution rather than the true distribution. We denote this process as $\mathcal{T}_{\mathrm{m}} \xleftarrow{n_t} \tilde{\mathsf{m}}_n(l|k)$, and use the green color to denote the empirical distribution of the model.

Note that, as in Equation 6, the model after the log in Equation 8 is fixed. Similarly to the MI estimation process, the value of the estimation $\widehat{\mathrm{HI}}(K;L)$ when $n_t \to \infty$ equals $\mathrm{HI}(K;L)$. *In most practical cases, the HI will be estimated directly via Equations 7 (which is simpler and faster).*

Next, the PI is theoretically defined as follows in the discrete case:

$$\mathrm{PI}_n(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|k) \cdot \log_2 \tilde{\mathsf{m}}_n(k|l), \tag{9}$$

and as follows in the continuous case:

$$\mathrm{PI}_n(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \int_{l\in\mathcal{L}} \mathsf{f}(l|k) \cdot \log_2 \tilde{\mathsf{m}}_n(k|l)\ dl. \tag{10}$$

In contrast with the HI, these equations cannot be computed directly since they require the knowledge of the true distributions $\mathsf{p}(l|k)$ and $\mathsf{f}(l|k)$ which are unknown. *So concretely, the PI will always be computed thanks to the following sampling process* (where we keep the red color code for the model and the blue color code for the true empirical distribution):

$$\widehat{\mathrm{PI}}_n(K;L) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{i=1}^{n_t(k)} \frac{1}{n_t(k)} \cdot \log_2 \tilde{\mathsf{m}}(k|l_k(i)). \tag{11}$$

This is feasible in practice since, even though the analytical form of the true distributions is unknown to the evaluator, he can sample these distributions, by measuring his target implementation.

Note again that, as in Equation 6, the model after the log in Equation 11 is fixed. So the PI captures the amount of information that can be extracted from some fixed model (usually obtained by estimation in an earlier phase). In other words, the PI computation is a two-step process: first a model is estimated, second the amount of information it provides is estimated. This is captured in our equations with the tilde and hat notations: the first one is for the estimation of the model, the second one for the computation of the information theoretic metrics by sampling.

Other useful facts. We next list a few additional former results.

• *A sufficient condition for successful (maximum likelihood) attacks.* As previously mentioned, the PI can be negative, indicating an estimated model that is too different from the true distribution. Also, the link between information theoretic metrics and the success rate of maximum likelihood attacks only holds per key. A sufficient condition for successful maximum likelihood attacks, first stated in [33], can therefore be given based on the "PI per key". For this purpose, and again assuming uniformly distributed keys, we first define a PI matrix (PIM) as follows:

$$\widehat{\mathrm{PIM}}_n(k,k^*) = \mathrm{H}(K) + \sum_{i=1}^{n_t(k)} \frac{1}{n_t(k)} \cdot \log_2 \tilde{\mathsf{m}}_n(k^*|l). \tag{12}$$

It captures the correlation between a key generating leakages k and a key candidate in a maximum likelihood attack k^*. The sufficient condition of successful attack against this key k is:

$$k = \underset{k^*\in\mathcal{K}}{\operatorname{argmax}}\ \widehat{\mathrm{PIM}}_n(k,k^*). \tag{13}$$

The PI is connected to the PIM: $\widehat{\mathrm{PI}}_n(K;L) = \mathop{\mathsf{E}}_{k\in\mathcal{K}}\Big(\widehat{\mathrm{PIM}}_n(k,k)\Big)$.

• *Key equivalence in the standard DPA setting.* In the usual (divide-and-conquer) side-channel analysis context, formalized in [21] as the standard DPA setting that we consider next, the adversary can continuously accumulate information about the key thanks to multiple input plaintexts x. Information theoretic metrics such as the MI, HI and PI therefore have to include another sum over these inputs to be reflective of this setting. For example in the discrete MI case, it yields:

$$\mathrm{MI}(K;L,X) = \mathrm{H}(K) + \sum_{k\in\mathcal{K}} \mathsf{p}(k) \cdot \sum_{x\in\mathcal{X}} \mathsf{p}(x) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|k,x) \cdot \log_2 \mathsf{p}(k|l,x). \quad (14)$$

Concretely, the adversary exploits the leakages after a first group operation between uniformly distributed plaintexts x and a key k took place. For example, he can target an intermediate operation $y = x \oplus k$ or $y = \mathsf{S}(x \oplus k)$ with S a block cipher S-box.[3] As a result, one can leverage the "key equivalence property" also proven in [21], which states that $\mathrm{MI}(K;L,X) = \mathrm{MI}(k;L,X) = \mathrm{MI}(Y;L)$ (i.e., there are no weak keys with respect to standard DPA and all the information exploited depends on the target intermediate computation Y).[4] Again, we use the $\mathrm{MI}(k;L,X)$ notation for a "MI per key" (i.e., Equation 14 for a fixed value of K, which is the same for all k's). The same type of result holds with the HI and PI. In the following, and in order to keep notations concise, we will therefore state our results for $\mathrm{MI}(Y;L)$, $\mathrm{HI}_n(Y;L)$ and $\mathrm{PI}_n(Y;L)$:

$$\mathrm{MI}(Y;L) = \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|y) \cdot \log_2 \mathsf{p}(y|l), \quad (15)$$

$$\mathrm{HI}_n(Y;L) = \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \tilde{\mathsf{m}}_n(l|y) \cdot \log_2 \tilde{\mathsf{m}}_n(y|l), \quad (16)$$

$$\mathrm{PI}_n(Y;L) = \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|y) \cdot \log_2 \tilde{\mathsf{m}}_n(y|l), \quad (17)$$

where the n subscript is the amount of leakages used to estimate the model.

• *Cross-validation.* When computing a metric by sampling, one generally uses cross-validation in order to better take advantage of the collected

[3] It is shown in [36] that their adaptive selection only marginally improves the attacks, and in [10,11] how this average metric can be used to state a sufficient condition for secure masked implementations.

[4] The second equality is turned into an inequality in case of non-bijective S-boxes.

data. As detailed in [13], it allows all the measured leakages to be used both as profiling and as test samples (but not both at the same time).

• *Metrics convergence and confidence intervals.* When estimating a metric by sampling, one is generally interested in knowing whether the computed value is close enough to the asymptotic one. In the context of side-channel analysis considered here, the amount of collected data is generally sufficient to build a "convergence plot" (see the experimental section) enabling to gain simple (visual) confidence that the metric is well estimated. If needed (e.g., in case of limited amount of data available), the bootstrap confidence intervals proposed in [17] can be used.

• *Outliers.* We finally note that outliers may prevent the PI metric computed from real data to converge (e.g., in case a probability zero is assigned to the correct y, leading to a $\log(0)$ in the PI equation). The treatment of these outliers will be discussed in the next section.

3 Theoretical bounds for the MI metric

Given the motivation that the MI metric is a good predictor of the success probability of a worst-case side-channel attack using the true leakage model, and the impossibility to compute it directly for unknown distributions, we now provide our main theoretical results and show how the HI and PI metrics can be used to bound the MI. We first state our results for discrete leakages and discuss the continuous case in Section 3.4. We will consider three quantities for this purpose:

– The previously defined MI with $\mathsf{p}(y|l)$ computed thanks to Bayes assuming uniform y's (uniform y's are typically encountered in the aforementioned standard DPA setting):

$$\begin{aligned} \mathrm{MI}(Y;L) &= \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|y) \cdot \log_2 \mathsf{p}(y|l), \qquad (18) \\ &= \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \mathsf{p}(l|y) \cdot \log_2 \frac{\mathsf{p}(l|y)}{\sum_{y^*\in\mathcal{Y}} \mathsf{p}(l|y^*)}. \end{aligned}$$

– The PI (i.e., Equation 17) under a similar uniformity assumption.
– The empirical HI (eHI), which is Equation 16 taking as model $\tilde{\mathsf{m}}_n(l|y)$ the empirical distribution, that we denote by $\tilde{\mathsf{e}}_n(l|y)$, under a similar uniformity assumption:

$$\mathrm{eHI}_n(Y;L) = \mathrm{H}(Y) + \sum_{y\in\mathcal{Y}} \mathsf{p}(y) \cdot \sum_{l\in\mathcal{L}} \tilde{\mathsf{e}}_n(l|y) \cdot \log_2 \tilde{\mathsf{e}}_n(y|l). \qquad (19)$$

Note that the eHI is exactly the biased maximum likelihood estimator of the MI that is used in the leakage detection test of Chatzikokolaki et al. [6], applied in the SCA setting by Mather et al [24]. As detailed next, under our uniformity assumption this estimator of the MI is biased upwards everywhere, which explains why the eHI provides an upper bound of the unknown MI.

3.1 Technical lemmas

We start with a few technical lemmas that we need to prove our two main theorems. Note that some of them are variations of well-known results given in textbooks such as [8]. We provide the proofs for the sake of completeness and for readers not familiar with information theory. Considering a discrete random variable taking values $1, 2, \ldots, t$, we next denote the actual probability of a value v as $\mathsf{p}(v)$, and the t-dimensional vector containing these probabilities as p.

Lemma 1. *Denoting by $\tilde{\mathsf{e}}_n$ the empirical distribution estimated from n i.i.d. leakage samples indexed $1, 2, \ldots, n$, and by $\tilde{\mathsf{e}}_n^j$ the empirical distribution estimated from the same samples excluding the sample j, the following equality holds:*

$$\tilde{\mathsf{e}}_n = \sum_{j=1:n} \frac{1}{n}\, \tilde{\mathsf{e}}_n^j,$$

and each empirical distribution $\tilde{\mathsf{e}}_n^j$ follows the same distribution as $\tilde{\mathsf{e}}_{n-1}$.

Proof. Let $x \in \{1, 2, \ldots, t\}^n$ be the random i.i.d. samples. For any subset $\mathcal{S}$ of $\{1, \ldots, n\}$, we denote by $\tilde{\mathsf{e}}_{\mathcal{S}}$ the empirical distribution of the sample whose indices are in $\mathcal{S}$. Observe that:

$$\tilde{\mathsf{e}}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} I_{x_i},$$

with I_{x_i} the indicator function taking the value 1 for the entry x_i and 0 otherwise. We then have:

$$\sum_{j=1:n} \frac{1}{n}\, \tilde{\mathsf{e}}_n^j = \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i \in \{1:n\} \setminus \{j\}} \frac{1}{n-1} I_{x_i} \right),$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\left(\sum_{i=1}^{n} I_{x_i} \right) - I_{x_j} \right), \\
&= \frac{1}{n(n-1)} \left(n \left(\sum_{i=1}^{n} I_{x_i} \right) - \sum_{j=1}^{n} I_{x_j} \right), \\
&= \frac{1}{n(n-1)} (n-1) \left(\sum_{i=1}^{n} I_{x_i} \right), \\
&= \frac{1}{n} \sum_{i=1}^{n} I_{x_i} = \tilde{\mathsf{e}}_n,
\end{aligned}
$$

which proves the equality in the lemma. Moreover, since the samples are i.i.d., all $\tilde{\mathsf{e}}_n^j$ follow the same distribution, and in particular the same distribution as $\tilde{\mathsf{e}}_n^n = \tilde{\mathsf{e}}_{n-1}$. □

Lemma 2. *Let $\gamma : [0,1]^t \to \mathbb{R}$ be a convex function. Then for any $n > 1$, we have:*

$$\gamma(\mathsf{p}) \leq \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_n)\Big) \leq \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_{n-1})\Big).$$

Moreover, if γ is continuous at p and bounded from above on $[0,1]^t$, then:

$$\mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_n)\Big) \to \gamma(\mathsf{p}),$$

monotonically with n. Similarly, if γ is concave and under the assumption that it is continuous and bounded from below, the same result holds with reverse inequalities.

Proof. We focus on the convex case and begin with the first inequality. Observe that:

$$\mathsf{p} = \mathsf{E}\left(\tilde{\mathsf{e}}_n\right). \tag{20}$$

Indeed, by linearity of the expected value, we have $\mathsf{E}\left(\tilde{\mathsf{e}}_n\right) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}(I_{x_i})$, with I_{x_i} the indicator function, whose t-dimensional value is 1 for the entry x_i and 0 otherwise. Therefore, for any i and entry $v \in \{1, \ldots, t\}$:

$$\mathsf{E}(I_{x_i})_v = 1 \cdot \Pr(x_i = v) + 0 \cdot \Pr(x_i \neq v) = \mathsf{p}(v),$$

from which (20) follows. Hence, due to the convexity of γ, we have:

$$\gamma(\mathsf{p}) = \gamma\Big(\mathsf{E}\left(\tilde{\mathsf{e}}_n\right)\Big) \leq \mathsf{E}\Big(\gamma\left(\tilde{\mathsf{e}}_n\right)\Big).$$

For the second inequality, it follows from Lemma 1 that:

$$\tilde{\mathsf{e}}_n = \sum_{j=1:n} \frac{1}{n}\, \tilde{\mathsf{e}}_n^j.$$

Hence we have:

$$\gamma(\tilde{\mathsf{e}}_n) = \gamma\left(\sum_{j=1:n} \frac{1}{n}\, \tilde{\mathsf{e}}_n^j\right) \leq \sum_{j=1:n} \frac{1}{n} \gamma\left(\tilde{\mathsf{e}}_n^j\right).$$

Moreover, each $\tilde{\mathsf{e}}_n^j$ has the same distribution as $\tilde{\mathsf{e}}_{n-1}$. Hence:

$$\begin{aligned}
\mathsf{E}\Big(\gamma\left(\tilde{\mathsf{e}}_n\right)\Big) &\leq \mathsf{E}\left(\sum_{j=1:n} \frac{1}{n}\, \gamma\left(\tilde{\mathsf{e}}_n^j\right)\right), \\
&= \sum_{j=1:n} \frac{1}{n}\, \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_n^j)\Big), \\
&= \sum_{j=1:n} \frac{1}{n}\, \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_{n-1})\Big) = \mathsf{E}\gamma(\tilde{\mathsf{e}}_{n-1}).
\end{aligned}$$

Let us now show the convergence under the assumption that γ is continuous at p and uniformly bounded by some M. By continuity of γ at p, for every ϵ there is a δ such that $||\tilde{\mathsf{e}}_n - \mathsf{p}|| \leq \delta$ implies $|\gamma(\tilde{\mathsf{e}}_n) - \gamma(\mathsf{p})| \leq \epsilon$. Moreover, $\tilde{\mathsf{e}}_n$ converges in probability to p, meaning that for every (δ, ϵ') there is a n' such that $\Pr(||\tilde{\mathsf{e}}_n - \mathsf{p}|| > \delta) < \epsilon'$ for any $n > n'$. As a consequence, for $n > n'$, we have:

$$\Pr(|\gamma(\tilde{\mathsf{e}}_n) - \gamma(\mathsf{p})| > \epsilon) < \epsilon'.$$

Remembering that $\gamma(.) < M$, we then have that for every $n > n'$:

$$\begin{aligned}
\mathsf{E}\Big(\gamma(f^n)\Big) - \gamma(\mathsf{p}) &= \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_n) - \gamma(\mathsf{p})\Big), \\
&\leq \epsilon \Pr\Big(|\gamma(\tilde{\mathsf{e}}_n) - \gamma(\mathsf{p})| \leq \epsilon\Big) + \Big(M - \gamma(\mathsf{p})\Big) \Pr\Big(|\gamma(\tilde{\mathsf{e}}_n) - \gamma(\mathsf{p})| > \epsilon\Big), \\
&\leq \epsilon(1 - \epsilon') + \epsilon'\Big(M - \gamma(\mathsf{p})\Big),
\end{aligned}$$

for every $n > n'$. Combining this with $\gamma(\mathsf{p}) \leq \mathsf{E}\Big(\gamma(\tilde{\mathsf{e}}_n)\Big)$ yields the desired convergence result. □

Lemma 3. *Let $y \in \mathbb{R}_+^m$ be a vector of positive entries. Then for any positive $x \in \mathbb{R}_+^m$, we have:*

$$\sum_i y_i \log_2 \frac{x_i}{\sum_j x_j} \leq \sum_i y_i \log_2 \frac{y_i}{\sum_j y_j},$$

with equality if and only if $x_i = k y_i$ for some $k > 0$.

Proof. Let $x' = x/(\sum_j x_j)$ and $y' = y/(\sum_j y_j)$. These vectors can be viewed as probability distributions since they are non-negative and sum to 1. Hence we can compute the following KL-divergence, which is always non-negative, and zero if and only if $x' = y'$:

$$0 \leq D_{KL}(y'||x') = \sum_i \left(y'_i \log \left(\frac{y'_i}{x'_i} \right) \right).$$

Using $\log(y'_i/x'_i) = \log y'_i - \log x'_i$, we obtain:

$$\sum_i (y'_i \log x'_i) \leq \sum_i (y'_i \log y'_i),$$

from which the result follows by replacing x'_i, y'_i and multiplying by $\sum_j y_j$. Equality holds if and only if $x' = y'$, that is, if $x = ky$ for some $k > 0$. □

3.2 Bound from the HI

We first recall the following standard result from Cover and Thomas:

Theorem 1 (Cover & Thomas, 2.7.4 [8]). *The mutual information* $\mathrm{MI}(Y; L)$ *is a concave function of* $\mathsf{p}(y)$ *for fixed* $\mathsf{p}(l|y)$ *and a convex function of* $\mathsf{p}(l|y)$ *for fixed* $\mathsf{p}(y)$.

Combined with the technical Lemma 2, it leads to our main result:

Theorem 2. *On average over the profiling sets* $\mathcal{M}$ *used to estimate the eHI and assuming that the target random variable* Y *has (constant) uniform probability, we have:*

$$\mathop{\mathsf{E}}_{\mathcal{M} \overset{n}{\leftarrow} \mathsf{p}(l|y)} \Big(\mathrm{eHI}_n(Y; L) \Big) \geq \mathop{\mathsf{E}}_{\mathcal{M} \overset{n-1}{\leftarrow} \mathsf{p}(l|y)} \Big(\mathrm{eHI}_{n-1}(Y; L) \Big) \geq \mathrm{MI}(Y; L).$$

Moreover, $\lim_{n \to \infty} \mathrm{eHI}_n(Y; L) = \mathrm{MI}(Y; L)$ *(i.e., the eHI monotonically converges towards the MI).*

Proof. Observe that $\mathrm{eHI}_n(Y; L)$ is the mutual information between Y and the empirical distribution of the leakages. Hence (thanks to Theorem 1), it is convex in $\tilde{\mathsf{e}}_n(l|y)$ for a fixed distribution of y (which we have by assumption). The result then follows from Lemma 2. □

3.3 Bound from the PI

Theorem 3. *Assuming that the target random variable Y has (constant) uniform probability and given any model $\tilde{\mathsf{m}}_n(l|y)$ for the conditional probabilities $\mathsf{p}(l|y)$, we have:*

$$\mathrm{PI}_n(Y;L) := \mathrm{H}(Y) + \sum_y \mathsf{p}(y) \sum_l \mathsf{p}(l|y) \log_2 \frac{\tilde{\mathsf{m}}_n(l|y)}{\sum_{y^*} \tilde{\mathsf{m}}_n(l|y^*)} \leq \mathrm{MI}(Y;L).$$

Proof. Since $\mathsf{p}(y)$ is a constant c, we have:

$$\mathrm{PI}_n(Y;L) = \mathrm{H}(Y) + c \sum_l \left(\sum_y \mathsf{p}(l|y) \log_2 \frac{\tilde{\mathsf{m}}_n(l|y)}{\sum_{y^*} \tilde{\mathsf{m}}_n(l|y^*)} \right). \tag{21}$$

Now for any l, it follows from Lemma 3 that:

$$\sum_y \mathsf{p}(l|y) \log_2 \frac{\tilde{\mathsf{m}}_n(l|y)}{\sum_{y^*} \tilde{\mathsf{m}}_n(l|y^*)} \leq \sum_y \mathsf{p}(l|y) \log_2 \frac{\mathsf{p}(l|y)}{\sum_{y^*} \mathsf{p}(l|y^*)}. \tag{22}$$

Re-introducing this in Equation 21 leads to:

$$\begin{aligned} \mathrm{PI}_n(Y;L) &\leq \mathrm{H}(Y) + c \sum_l \left(\sum_y \mathsf{p}(l|y) \log_2 \frac{\mathsf{p}(l|y)}{\sum_{y^*} \mathsf{p}(l|y^*)} \right), \\ &= \mathrm{H}(Y) + \sum_y \mathsf{p}(y) \sum_l \mathsf{p}(l|y) \log_2 \frac{\mathsf{p}(l|y)}{\sum_{y^*} \mathsf{p}(l|y^*)}, \\ &= \mathrm{MI}(Y;L). \end{aligned} \tag{23}$$

□

Additional observation. It would be nice to know that $\mathrm{PI}_n(Y;L) = \mathrm{MI}(Y;L)$ if and only if $\tilde{\mathsf{m}}_n(l|y) = \mathsf{p}(l|y)$. However, this is not true in general. Suppose for example that l and y only take two values l_1, l_2 and y_1, y_2, and that $\mathsf{p}(l_i|y_j) = 1/2$ for all four cases. Then consider the model defined by $\tilde{\mathsf{m}}_n(l_1|y_j) = \alpha$ and $\tilde{\mathsf{m}}_n(l_2|y_j) = 1 - \alpha$ for both y_j and some $\alpha \in [0,1]$. Again assuming a constant $\mathsf{p}(y) = 1/2$, the perceived information of any such model would be:

$$\mathrm{PI}_n(Y;L) = \mathrm{H}(Y) + \frac{1}{2} \sum_l \sum_y \frac{1}{2} \log_2 \frac{\tilde{\mathsf{m}}_n(l|y)}{\sum_{y^*} \tilde{\mathsf{m}}_n(l|y^*)},$$

$$
\begin{aligned}
&= \mathrm{H}(Y) + \frac{1}{4}\Big(\log_2 \frac{\tilde{\mathsf{m}}_n(l_1|y_1)}{\tilde{\mathsf{m}}_n(l_1|y_1)+\tilde{\mathsf{m}}_n(l_1|y_2)} + \log_2 \frac{\tilde{\mathsf{m}}_n(l_1|y_2)}{\tilde{\mathsf{m}}_n(l_1|y_1)+\tilde{\mathsf{m}}_n(l_1|y_2)} \\
&\quad + \log_2 \frac{\tilde{\mathsf{m}}_n(l_2|y_1)}{\tilde{\mathsf{m}}_n(l_2|y_1)+\tilde{\mathsf{m}}_n(l_2|y_2)} + \log_2 \frac{\tilde{\mathsf{m}}_n(l_2|y_2)}{\tilde{\mathsf{m}}_n(l_2|y_1)+\tilde{\mathsf{m}}_n(l_2|y_2)} \Big), \\
&= \mathrm{H}(Y) + \frac{1}{4}\left(\log_2 \frac{\alpha}{\alpha+\alpha} + \log_2 \frac{\alpha}{\alpha+\alpha} + \log_2 \frac{1-\alpha}{1-\alpha+1-\alpha} + \log_2 \frac{1-\alpha}{1-\alpha+1-\alpha} \right), \\
&= \mathrm{H}(Y) + \log_2 \frac{1}{2},
\end{aligned}
$$

irrespectively of α. The value obtained for any α is the same as for $\alpha = 1/2$ (i.e., the only value for which $\tilde{\mathsf{m}}_n(l|y) = \mathsf{p}(l|y)$). We therefore conclude that $\mathrm{PI}_n(Y;L) = \mathrm{MI}(Y;L)$ does not imply that the model accurately describes the distribution of leakage.

As a complement of this observation, we next characterize the conditions under which $\tilde{\mathsf{m}}_n(l|y) = \mathsf{p}(l|y)$ is the only maximum.

Proposition 1. *Let $\boldsymbol{P}$ be the matrix defined by $\boldsymbol{P}_{l,y} = \mathsf{p}(l|y)$. If $\boldsymbol{P}$ is full row rank, then $\mathrm{PI}_n(Y;L) = \mathrm{MI}(Y;L)$ if and only if $\tilde{\mathsf{m}}_n(l|y) = \mathsf{p}(l|y)$. If $\boldsymbol{P}$ is not full row rank then one can build alternative models leading to $\mathrm{PI}_n(Y;L) = \mathrm{MI}(Y;L)$.*

Proof. Let $\tilde{\mathsf{m}}_n(l|y)$ be a conditional probability distribution. Keeping the notations of Theorem 3, $\mathrm{PI}_n(Y;L) = \mathrm{MI}(Y;L)$ holds if and only if equality holds in Equation 23, and therefore if and only if it holds in Equation 22 for every l. By Lemma 3, this is equivalent to the existence of a positive vector k such that $\tilde{\mathsf{m}}_n(l|y) = k_l \cdot \mathsf{p}(l|y)$ holds for every y, l. Clearly, $\tilde{\mathsf{m}}_n(l|y) = \mathsf{p}(l|y)$ for all y, l if and only if all k_l's are equal to 1 (i.e., $\boldsymbol{k} = \mathbf{1}$). Now, for an arbitrary positive vector $\boldsymbol{k}$, the quantities $\tilde{\mathsf{m}}_n(l|y) = k_l\ \mathsf{p}(l|y)$ define valid conditional probabilities if and only if (i) they all belong to $[0,1]$, and (ii) $\sum_l \tilde{\mathsf{m}}_n(l|y) = 1$ for every y. We show next that these conditions imply $\boldsymbol{k} = \mathbf{1}$ if and only if $\boldsymbol{P}$ is full row-rank, which will imply our result. Define the matrix $\boldsymbol{M}$ as $\boldsymbol{M}_{l,y} = \tilde{\mathsf{m}}_n(l|y)$ and the diagonal matrix $\boldsymbol{K}$ as $\boldsymbol{K}_{ll} = k_l$ (so that $\boldsymbol{k} = \boldsymbol{K}\mathbf{1}$)). Condition (ii) can be rewritten as $\mathbf{1}^T\boldsymbol{M} = \mathbf{1}^T = \mathbf{1}^T\boldsymbol{P}$, and $\tilde{\mathsf{m}}_n(l|y) = k_l\ \mathsf{p}(l|y)$ can be re-expressed as $\boldsymbol{M} = \boldsymbol{K}\boldsymbol{P}$. Therefore:

$$
(k-\mathbf{1})^T\boldsymbol{P} = (\mathbf{1}^T\boldsymbol{K}-\mathbf{1}^T)\boldsymbol{P} = \mathbf{1}^T\boldsymbol{K}\boldsymbol{P}-\mathbf{1}^T\boldsymbol{P} = \mathbf{1}^T\boldsymbol{M}-\mathbf{1}^T\boldsymbol{P} = \mathbf{1}^T-\mathbf{1}^T = 0.
$$

That is, the vector $(\boldsymbol{k}-\mathbf{1})^T$ is in the left-kernel of $\boldsymbol{P}$. Hence, if $\boldsymbol{P}$ has full-row rank, the only vector $\boldsymbol{k}$ for which (ii) is satisfied is $\boldsymbol{k} = \mathbf{1}$. Otherwise, any vector of the form $\boldsymbol{k} = \mathbf{1} + \alpha v$ for $\alpha \neq 0$ and $v \neq 0$ in the left-kernel of $\boldsymbol{P}$ would lead $\tilde{\mathsf{m}}_n(l|y)$ to satisfy condition (ii). To finish the proof, we show that we can also have condition (i) satisfied. By taking a sufficiently

small α, we can ensure that $\boldsymbol{k}$ is positive, and therefore that the $\tilde{\mathsf{m}}_n(l|y)$'s are non-negative. Because $\sum_l \tilde{\mathsf{m}}_n(l|y) = 1$ by condition (ii), this implies that $\tilde{\mathsf{m}}_n(l|y) \leq 1$ for every l, y and that condition (i) is satisfied. □

Note that this full row rank condition may not be achieved in so-called Simple Power Analysis (SPA) attacks with "compressive" leakage functions. For example, imagine an implementation leaking the noise-free Hamming weight of an n-bit key. Then, the number of leakages (i.e., $n+1$) is lower then the number of keys (i.e., 2^n) and $\boldsymbol{P}$ cannot have full row rank. By contrast, in the DPA setting, the amount of leakages that the adversary can observe is multiplied by the number of plaintexts (i.e., 2^n) and the matrix $\boldsymbol{P}_{(l,x),k} = \mathsf{p}(l, x|k)$ is expected to be of full row rank.

3.4 Discussion and application of the results

The previous theorems can be quite directly applied in a side-channel evaluation context. Yet the following clarifications are worth being pointed out before moving to experiments.

First and as previously mentioned, one technical difficulty that may arise is the presence of outliers (or simply rare events) leading to zero probabilities for the good key candidate, and therefore to a $\log(0)$ in the PI equation (for the HI equation, we assume $0 \cdot \log(0) = 0$). A simple heuristic to deal with these cases is to lower-bound such probabilities to $\frac{1}{n_t(k)}$ and to report the fraction of corrected probabilities (which vanishes as n increases) with the experimental results.

Second, the HI bound of Section 3.2 is stated for the empirical distribution that is straightforward to estimate in a discrete case with finite support thanks to histograms. In this respect, we observe that actual leakages are measured thanks to sampling devices (hence are inherently discrete and finite). We also refer to the fast leakage assessment methodology in [31] for a motivation why this may lead to performance gains for the evaluator. Yet, there is actually nothing specific to discrete distributions in the way we obtain this bound (up to the slightly different convergences discussed in the full version of the paper [3]). So it is applicable to continuous distributions and estimators. For example, we could replace the estimation of the discrete MI based on histograms that we use to compute the eHI by a Kernel-based one such as used in [7, 24]). In the next section, we also consider a simplified (Gaussian) model family and show how the HI bound can be useful in this context.

4 Empirical confirmation

4.1 Simulated experiments

In order to demonstrate the relevance of the previous tools, we start by investigating a standard simulation setting where the evaluator / adversary exploits the leakages corresponding to several executions of the AES S-box. Our first scenario corresponds to a univariate attack against an unprotected implementation of this S-box, where the leakage samples are of the form:

$$l_i^1 = \mathsf{HW}\Big(\mathsf{S}(x \oplus k)\Big) + r_i,$$

with HW the Hamming weight function, and r_i a Gaussian distributed noise sample with variance σ^2. The noise level is a parameter of our simulations. For convenience (and simpler interpretation) we report it as a Signal-to-Noise Ratio (SNR) which is defined as in [19] as the variance of the signal (which is worth 2 in the case of a random 8-bit Hamming weight value) divided by σ^2.

Our second simulated scenario corresponds to a bivariate attack against the same unprotected implementation of the AES S-box, where the leakage vectors are of the form:

$$\boldsymbol{l}_i^2 = \Big[\mathsf{HW}(x \oplus k) + r_i;\ \mathsf{HW}\Big(\mathsf{S}(x \oplus k)\Big) + r_i'\Big].$$

Finally, our third scenario corresponds to a univariate attack against a masked (i.e., secret shared [4]) implementation of this S-box, where the leakage samples are of the form:

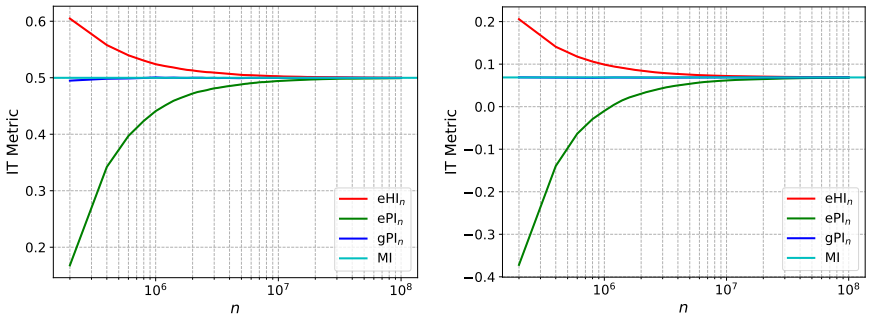
$$\boldsymbol{l}_i^3 = \Big[\mathsf{HW}\Big(\mathsf{S}(x \oplus k) \oplus q\Big) + \mathsf{HW}(q) + r_i\Big],$$

with q a secret mask picked up uniformly at random by the leaking device.

The results of our first scenario for high and medium SNRs are in Figure 1, where we plot the MI (that is known since we are in a simulated setting), the eHI, the ePI (considered in our bounds) and the Gaussian PI (gPI) which is the PI corresponding to a Gaussian leakage model. The IT metrics are plot in function of the number of traces in the profiling set n.[5] As expected, the eHI provides an average upper bound that converges monotonically towards the MI, and the ePI provides a lower bound.

[5] We use $n_t = n$, which leads to good estimations since the number of measurements needed to estimate a model is usually larger than the number of leakages needed to recover the key with a well-estimated model [32].

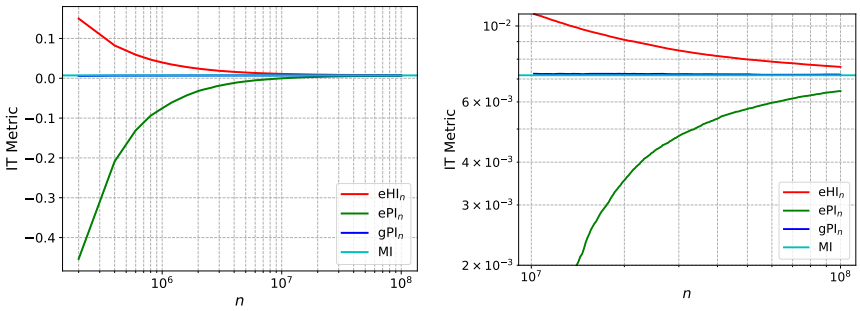
Besides, the gPI converges rapidly towards the true MI since in our simulations, the leakages are generated based on a Gaussian distribution. So making this additional assumption in such an ideal setting allows faster model convergence without information loss.



(a) SNR = 1 (lin. scale). (b) SNR = 0.1 (lin. scale).

Fig. 1: Simulations, unprotected S-box, high & medium SNRs, univariate.

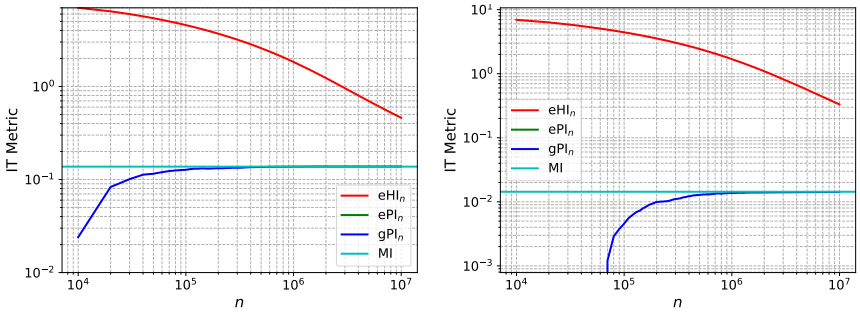
These results are confirmed with the similar plots given in Figure 2 for a lower SNR of 0.01. For readability, the right plot switches to a logarithmic scale for the Y axis. It illustrates a context where it is possible to formally bound the mutual information to values lower than 10^{-2}.



(a) SNR = 0.01 (lin. scale). (b) SNR = 0.01 (log. scale).

Fig. 2: Simulations, unprotected S-box, low SNR, univariate.

Figure 1 and 2 correspond to simple (unprotected, univariate) cases where the estimation of the empirical distribution (despite significantly more expensive than the one of a Gaussian distribution) leads to reasonably tight bounds for the MI. We complement this observation with experiments corresponding to our second (unprotected, bivariate) context. As illustrated in Figure 3 for medium and low SNRs, this more challenging context leads to considerably less tight bounds, which can be explained by the (much) slower convergence of multivariate histograms. Note that we could not reach a positive ePI with $n = 10^7$ in this case (and the gPI still does it rapidly).



(a) SNR = 0.1 (log. scale). (b) SNR = 0.01 (log. scale).

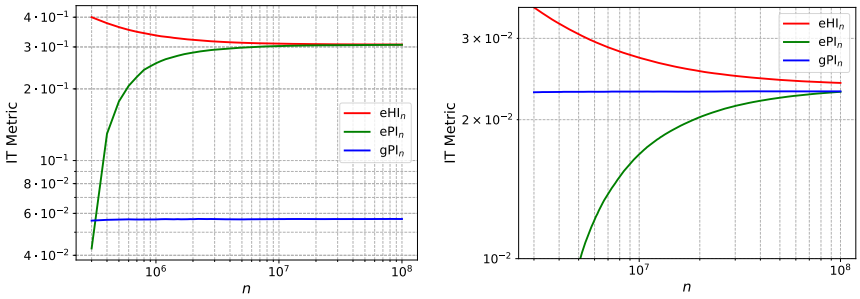
Fig. 3: Simulations, unprotected S-box, medium & low SNR, bivariate.

We finally report the results of the simulated masked implementation in Figure 4 for very high and high SNRs. The very high SNR case is intended to illustrate a context where the Gaussian assumption is not satisfied (since the masked leakage distribution is actually a Gaussian mixture), so that the gPI is considerably lower than the ePI. By contrast, and as observed (for example) in [14], Figure 1 (right), this Gaussian approximation becomes correct and the gPI gets close to the ePI as the noise increases, which we also see on the right part of Figure 4.

An open source code allowing to reproduce these results is given in [1].

4.2 Real measurements

We complement the previous simulated experiments with analyzes performed on actual measurements obtained from an FPGA implementation of the AES S-box. In order to instantiate a noise parameter as in our



(a) SNR = 10 (log. scale). (b) SNR = 1 (log. scale).

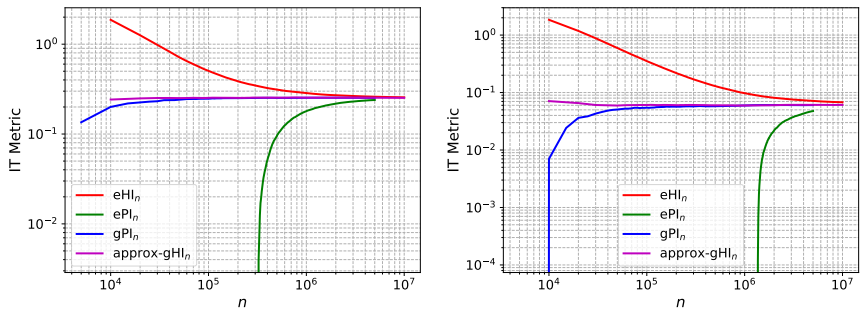
Fig. 4: Simulations, masked S-box, very high & high SNR, univariate.

simulations, we consider different architectures for this purpose: the target S-box is computed in parallel with $\pi \in \{0, 3, 7, 11)$ other S-boxes whose computations (for random inputs) generate "algorithmic noise". We implemented our design on a SAKURA-X board embedding a Xilinx Kintex-7 FPGA. The target device was running at 4 MHz and sampled at 500 Ms/s (i.e., 125 leakage points per cycle). We split our experiments in two parts. In a first part, we consider a univariate evaluation (similar to the first setting of our simulated setup) allowing reasonably tight worst-case bounds. In a second part, we consider a highly multivariate evaluation (i.e., an adversary exploiting all the 125 points of each clock cycle) and discuss how to connect this context with nearly worst-case security arguments for (e.g., masked) cryptographic implementations.

Univariate analyses & theoretical worst-case bounds. The eHI/ePI bounds computed for the most informative leakage points of our measurements for $\pi = 0$ and 7 are in Figure 5. The $\pi = 3$ and 11 cases are given in the full version of the paper [3]. We again observe that it is possible to obtain reasonably tight bounds (e.g., to bound the MI below 10^{-1} which is a sufficient noise for the masking countermeasure to be effective). Yet, as π increases and the MI decreases, we also see that tightening the bounds becomes increasingly data-intensive.

In view of the important amount of samples n needed to bound the MI, and of the popularity of the Gaussian assumption in SCAs [5], we additionally considered the Gaussian HI (gHI) which is the HI corresponding to a Gaussian model, and evaluated it based on the formula:

$$\mathsf{approx\text{-}gHI}_n(Y, L) = -\frac{1}{2} \cdot \log_2 \left(1 - \rho(M, L)^2\right), \tag{24}$$



(a) $\pi = 0$ (log. scale). (b) $\pi = 7$ (log. scale).

Fig. 5: Actual measurements, unprotected S-box, univariate.

where ρ is Pearson's correlation coefficient, L the leakage random variable (as previously) and M the model random variable. As discussed in [19], $\rho(Y, M)$ can be related to the leakages' SNR, which (in the case of Gaussian leakages) can be linked to the MI metric [11]. As observed in [21], the formula holds well for noisy Hamming weight leakages in case of "reasonably small" correlations values (i.e., typically $\rho < 0.1$). The latter is confirmed in our experiments of Figures 5. Namely, these figures first illustrate that the gHI is also an upper bound for the gPI and converges monotonically (as expected from the results in Section 3). They additionally show that the gHI and gPI are very close to the worst-case MI in our experimental setting. The latter is particularly interesting since the gHI converges very fast compared to the other metrics.

Multivariate analyzes and efficient evaluations. Ultimately, an evaluator would be interested in efficiently and tightly bounding the total amount of information provided by his leakage points. As clear from the Section 4.1 (and the bivariate analysis of Figure 3), obtaining tight MI bounds with two dimensions is already data-intensive. Hence, applying such a straightforward approach to our measurements where each clock cycle has 125 points is unlikely to provide any tight result. So here as well, we considered the multivariate gHI as a useful alternative (yet, this time without possibility to compare it to the eHI). For this purpose, we use the formula for the differential entropy of a multivariate Gaussian distribution:

$$\mathrm{gH}(\boldsymbol{Z}) = \frac{\frac{1}{2}\log\left(\mathsf{det}(2\pi e\boldsymbol{\Sigma})\right)}{\log(2)}, \tag{25}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the Gaussian-distributed random variable $\boldsymbol{Z}$, $\mathsf{det}(.)$ denotes the matrix determinant and the $\log(2)$ of the denominator is to obtain a value in bits. We then used this standard formula to approximate the multivariate gHI as:

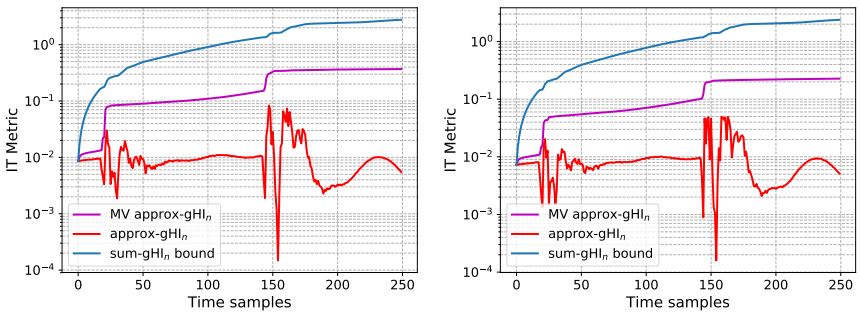
$$\mathsf{MV\ approx\text{-}gHI}_n(Y, \boldsymbol{L}) = \mathrm{gH}(\boldsymbol{M}) + \mathrm{gH}(\boldsymbol{L}) - \mathrm{gH}(\boldsymbol{M}; \boldsymbol{L}), \qquad (26)$$

which is the multivariate generalization of Equation 24. Note that as in Equation 24, this approximation is based on the (multivariate) model random variable, which captures the possibility that different leakage points can have different leakage behaviors despite depending on the same Y.

Note also that as the number of dimensions increases, using such an approximation is increasingly useful from the time complexity viewpoint. Indeed, while the univariate gHI can be computed directly by integration, computing the multivariate gHI in our experimental case study (where we exploit the measurements of two clock cycles corresponding to 250 leakage points) would require integrating a 250-dimension distribution. By contrast, evaluating Equation 26 only requires estimating the covariances matrices of the model, leakages and their joint distribution.

The approximations of the multivariate gHI for the cases $\pi = 3$ and 11 are in Figure 6. The $\pi = 0$ and 7 cases are given in the full version of the paper [3]. For completeness, the plots first report the univariate gHI for each time sample (in red). The multivariate Gaussian approximations of Equation 26 are then reported in purple in a cumulative manner: the value for time sample x corresponds to the x-variate estimation for dimensions 1 to x. Eventually, we added a conservative bound in blue, based on the assumption that each leakage point provides independent information and is summed. Those results are practically-relevant for two main reasons:

– First, they allow estimating the information of a very powerful yet realistic, close to worst-case adversary (since the univariate gHI is close to the eHI) in a more accurate (and less conservative) manner than bounds obtained based on an independence assumption. For example, the most informative point of Figure 6(b) has a (univariate) gHI of $4 \cdot 10^{-2}$ while our approximation of the multivariate gHI is worth $2 \cdot 10^{-1}$ (i.e., a factor 5 more) and the bound would suggest a gHI larger than one (i.e., no security). So it illustrates a case where our approximation provides a useful intermediate between a too optimistic univariate analysis and a too conservative bound based on an independence assumption. We note that as for the univariate case, the approximation of Equation 26 only holds for small HI values (i.e., typically below 0.1). For example, the approximation for the $\pi = 0$ case



(a) $\pi = 3$ (log. scale). (b) $\pi = 11$ (log. scale).

Fig. 6: Actual measurements, unprotected S-box, multivariate.

(given in the full version of the paper [3]) overestimates the information leakages. Yet, the quantitative analysis of those cases is anyway not very interesting (since they correspond to a too weak security).

– Second, these close to worst-case evaluations of the information leakages are obtained very efficiently (from the data complexity viewpoint). Taking again the $\pi = 11$ case for illustration, the Gaussian approximation of the 250-variate gHI already reaches a good convergence after approximately $n = 10^6$ samples (while the gPI is still negative with this amount of measurements). For completeness, we report the convergence plots of the multivariate gPI and gHI in the full version of the paper [3], where we can observe this faster convergence for lower number of dimensions (for which the gPI is positive).

5 Conclusions

This paper provides first quantitative tools to bound the information leakages exploited in SCAs, taking into account the risk of a "false sense of security" due to incorrect assumptions about the leakage distributions. In case of low-dimensional leakages, we are able to formally bound the amount of information obtained on a target random variable. In case of high-dimensional leakages (which typically happen in case of strong adversaries trying to exploit all the information in power or electromagnetic measurements), tightening these bounds usually requires an unrealistic amount of data. Yet, even in these cases, our tools can be used to approximate the information provided by more specialized (close to worst-case) adversaries, by exploiting simplifying (e.g., Gaussian) assumptions. As a

result, a natural approach to leakage certification is to mix (i) a low-dimension analysis estimating both the empirical and (for example) the Gaussian HI and PI metrics, in order to gauge the quality of the simplifying (e.g., Gaussian) assumption and (ii) a high-dimension analysis based on the simplifying assumption(s) only. Such an approach can considerably speed up security evaluations. First, estimating an HI bound is significantly less expensive than estimating the PI, both in terms of data complexity (as clear from the convergence plots of the previous section) and in terms of time complexity. For example, the multivariate gHI estimations of Section 4.2 are obtained within minutes of computations on a desktop computer whereas the gPI estimations take several hours (due to their expensive cross-validation step). Next, such information theoretic metrics can be used to bound the success rate of actual side-channel attacks much faster than by directly mounting attacks. These bounds can be used both in the context of standard divide-and-conquer adversaries as usually considered in current security evaluations (e.g., using the formulas in [11]), and for analyzing more advanced adversaries trying to combine the information leakages beyond the operations that can be easily guessed by a divide-and-conquer adversary (e.g., using the Local Random Probing Model in [16]). We believe these tools are important ingredients to strengthen the understanding of side-channel security evaluations and the design of countermeasures with strong security guarantees. We also believe they are of general interest and could find applications in other contexts such as timing attacks or privacy-related applications [23].

Acknowledgments. The authors thank Philippe Delsarte for stimulating discussions and Carolyn Whitnall for useful feedback on the HI/PI definitions and comments on early versions of this manuscript. Julien Hendrickx holds a WBI.World excellence fellowship. François-Xavier Standaert is a Senior Research Associate of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the EU through the ERC project SWORD (Consolidator Grant 724725) and the H2020 project REASSURE, and by a Concerted Research Action of the "Communauté Française de Belgique".

References

1. *https://github.com/obronchain/Leakage_Certification_Revisited*.
2. E. Brier, C. Clavier, and F. Olivier, *Correlation power analysis with a leakage model*, in Cryptographic Hardware and Embedded Systems - CHES 2004: 6th International Workshop Cambridge, MA, USA, August 11-13, 2004. Proceedings,

M. Joye and J. Quisquater, eds., vol. 3156 of Lecture Notes in Computer Science, Springer, 2004, pp. 16–29.
3. O. Bronchain, J. M. Hendrickx, C. Massart, A. Olshevsky, and F. Standaert, *Leakage certification revisited: Bounding model errors in side-channel security evaluations*, IACR Cryptology ePrint Archive, 2019 (2019), p. 132.
4. S. Chari, C. S. Jutla, J. R. Rao, and P. Rohatgi, *Towards sound approaches to counteract power-analysis attacks*, in Advances in Cryptology - CRYPTO '99, 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999, Proceedings, M. J. Wiener, ed., vol. 1666 of Lecture Notes in Computer Science, Springer, 1999, pp. 398–412.
5. S. Chari, J. R. Rao, and P. Rohatgi, *Template attacks*, in Cryptographic Hardware and Embedded Systems - CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002, Revised Papers, B. S. K. Jr., Ç. K. Koç, and C. Paar, eds., vol. 2523 of Lecture Notes in Computer Science, Springer, 2002, pp. 13–28.
6. K. Chatzikokolakis, T. Chothia, and A. Guha, *Statistical measurement of information leakage*, in Tools and Algorithms for the Construction and Analysis of Systems, 16th International Conference, TACAS 2010, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2010, Paphos, Cyprus, March 20-28, 2010. Proceedings, J. Esparza and R. Majumdar, eds., vol. 6015 of Lecture Notes in Computer Science, Springer, 2010, pp. 390–404.
7. T. Chothia and A. Guha, *A statistical test for information leaks using continuous mutual information*, in Proceedings of the 24th IEEE Computer Security Foundations Symposium, CSF 2011, Cernay-la-Ville, France, 27-29 June, 2011, IEEE Computer Society, 2011, pp. 177–190.
8. T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*, Wiley, 2006.
9. P. M. Domingos, *A unified bias-variance decomposition and its applications*, in Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000, P. Langley, ed., Morgan Kaufmann, 2000, pp. 231–238.
10. A. Duc, S. Dziembowski, and S. Faust, *Unifying leakage models: From probing attacks to noisy leakage*, in Nguyen and Oswald [25], pp. 423–440.
11. A. Duc, S. Faust, and F. Standaert, *Making masking security proofs concrete - or how to evaluate the security of any leaking device*, in Advances in Cryptology - EUROCRYPT 2015 - 34th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Sofia, Bulgaria, April 26-30, 2015, Proceedings, Part I, E. Oswald and M. Fischlin, eds., vol. 9056 of Lecture Notes in Computer Science, Springer, 2015, pp. 401–429.
12. F. Durvaux, F. Standaert, and S. M. D. Pozo, *Towards easy leakage certification: extended version*, J. Cryptographic Engineering, 7 (2017), pp. 129–147.
13. F. Durvaux, F. Standaert, and N. Veyrat-Charvillon, *How to certify the leakage of a chip?*, in Nguyen and Oswald [25], pp. 459–476.
14. V. Grosso, F. Standaert, and E. Prouff, *Low entropy masking schemes, revisited*, in Smart Card Research and Advanced Applications - 12th International Conference, CARDIS 2013, Berlin, Germany, November 27-29, 2013. Revised Selected Papers, A. Francillon and P. Rohatgi, eds., vol. 8419 of Lecture Notes in Computer Science, Springer, 2013, pp. 33–43.
15. S. Guilley, A. Heuser, O. Rioul, and F. Standaert, *Template attacks, optimal distinguishers and the perceived information metric*, Cryptarchi, (2015). https://perso.uclouvain.be/fstandae/PUBLIS/162.pdf.

16. Q. Guo, V. Grosso, and F. Standaert, *Modeling soft analytical side-channel attacks from a coding theory viewpoint*, IACR Cryptology ePrint Archive, 2018 (2018), p. 498.
17. J. Lange, C. Massart, A. Mouraux, and F. Standaert, *Side-channel attacks against the human brain: The PIN code case study (extended version)*, in Brain Informatics, vol. 5, Oct 2018, p. 12.
18. L. Lerman, N. Veshchikov, O. Markowitch, and F. Standaert, *Start simple and then refine: Bias-variance decomposition as a diagnosis tool for leakage profiling*, IEEE Trans. Computers, 67 (2018), pp. 268–283.
19. S. Mangard, *Hardware countermeasures against DPA ? A statistical analysis of their effectiveness*, in Topics in Cryptology - CT-RSA 2004, The Cryptographers' Track at the RSA Conference 2004, San Francisco, CA, USA, February 23-27, 2004, Proceedings, T. Okamoto, ed., vol. 2964 of Lecture Notes in Computer Science, Springer, 2004, pp. 222–235.
20. S. Mangard, E. Oswald, and T. Popp, *Power analysis attacks - revealing the secrets of smart cards*, Springer, 2007.
21. S. Mangard, E. Oswald, and F. Standaert, *One for all - all for one: unifying standard differential power analysis attacks*, IET Information Security, 5 (2011), pp. 100–110.
22. D. P. Martin, J. F. O'Connell, E. Oswald, and M. Stam, *Counting keys in parallel after a side channel attack*, in Advances in Cryptology - ASIACRYPT 2015 - 21st International Conference on the Theory and Application of Cryptology and Information Security, Auckland, New Zealand, November 29 - December 3, 2015, Proceedings, Part II, T. Iwata and J. H. Cheon, eds., vol. 9453 of Lecture Notes in Computer Science, Springer, 2015, pp. 313–337.
23. C. Massart and F. Standaert, *Revisiting location privacy from a side-channel analysis viewpoint (extended version)*, IACR Cryptology ePrint Archive, 2019 (2019), p. 467.
24. L. Mather, E. Oswald, J. Bandenburg, and M. Wójcik, *Does my device leak information? an a priori statistical power analysis of leakage detection tests*, in Advances in Cryptology - ASIACRYPT 2013 - 19th International Conference on the Theory and Application of Cryptology and Information Security, Bengaluru, India, December 1-5, 2013, Proceedings, Part I, K. Sako and P. Sarkar, eds., vol. 8269 of Lecture Notes in Computer Science, Springer, 2013, pp. 486–505.
25. P. Q. Nguyen and E. Oswald, eds., *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, vol. 8441 of Lecture Notes in Computer Science, Springer, 2014.
26. L. Paninski, *Estimation of entropy and mutual information*, Neural Computation, 15 (2003), pp. 1191–1253.
27. R. Poussier, F. Standaert, and V. Grosso, *Simple key enumeration (and rank estimation) using histograms: An integrated approach*, in Cryptographic Hardware and Embedded Systems - CHES 2016 - 18th International Conference, Santa Barbara, CA, USA, August 17-19, 2016, Proceedings, B. Gierlichs and A. Y. Poschmann, eds., vol. 9813 of Lecture Notes in Computer Science, Springer, 2016, pp. 61–81.
28. E. Prouff and M. Rivain, *Masking against side-channel attacks: A formal security proof*, in Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings, T. Johansson and P. Q. Nguyen, eds., vol. 7881 of Lecture Notes in Computer Science, Springer, 2013, pp. 142–159.

29. M. Renauld, F. Standaert, and N. Veyrat-Charvillon, *Algebraic side-channel attacks on the AES: why time also matters in DPA*, in Cryptographic Hardware and Embedded Systems - CHES 2009, 11th International Workshop, Lausanne, Switzerland, September 6-9, 2009, Proceedings, C. Clavier and K. Gaj, eds., vol. 5747 of Lecture Notes in Computer Science, Springer, 2009, pp. 97–111.
30. M. Renauld, F. Standaert, N. Veyrat-Charvillon, D. Kamel, and D. Flandre, *A formal study of power variability issues and side-channel attacks for nanoscale devices*, in Advances in Cryptology - EUROCRYPT 2011 - 30th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Tallinn, Estonia, May 15-19, 2011. Proceedings, K. G. Paterson, ed., vol. 6632 of Lecture Notes in Computer Science, Springer, 2011, pp. 109–128.
31. O. Reparaz, B. Gierlichs, and I. Verbauwhede, *Fast leakage assessment*, in Cryptographic Hardware and Embedded Systems - CHES 2017 - 19th International Conference, Taipei, Taiwan, September 25-28, 2017, Proceedings, W. Fischer and N. Homma, eds., vol. 10529 of Lecture Notes in Computer Science, Springer, 2017, pp. 387–399.
32. F. Standaert, F. Koeune, and W. Schindler, *How to compare profiled side-channel attacks?*, in Applied Cryptography and Network Security, 7th International Conference, ACNS 2009, Paris-Rocquencourt, France, June 2-5, 2009. Proceedings, M. Abdalla, D. Pointcheval, P. Fouque, and D. Vergnaud, eds., vol. 5536 of Lecture Notes in Computer Science, 2009, pp. 485–498.
33. F. Standaert, T. Malkin, and M. Yung, *A unified framework for the analysis of side-channel key recovery attacks*, in Advances in Cryptology - EUROCRYPT 2009, 28th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cologne, Germany, April 26-30, 2009. Proceedings, A. Joux, ed., vol. 5479 of Lecture Notes in Computer Science, Springer, 2009, pp. 443–461.
34. N. Veyrat-Charvillon, B. Gérard, M. Renauld, and F. Standaert, *An optimal key enumeration algorithm and its application to side-channel attacks*, in Selected Areas in Cryptography, 19th International Conference, SAC 2012, Windsor, ON, Canada, August 15-16, 2012, Revised Selected Papers, L. R. Knudsen and H. Wu, eds., vol. 7707 of Lecture Notes in Computer Science, Springer, 2012, pp. 390–406.
35. N. Veyrat-Charvillon, B. Gérard, and F. Standaert, *Soft analytical side-channel attacks*, in Advances in Cryptology - ASIACRYPT 2014 - 20th International Conference on the Theory and Application of Cryptology and Information Security, Kaoshiung, Taiwan, R.O.C., December 7-11, 2014. Proceedings, Part I, P. Sarkar and T. Iwata, eds., vol. 8873 of Lecture Notes in Computer Science, Springer, 2014, pp. 282–296.
36. N. Veyrat-Charvillon and F. Standaert, *Adaptive chosen-message side-channel attacks*, in Applied Cryptography and Network Security, 8th International Conference, ACNS 2010, Beijing, China, June 22-25, 2010. Proceedings, J. Zhou and M. Yung, eds., vol. 6123 of Lecture Notes in Computer Science, 2010, pp. 186–199.