An Information-Theoretic Model for Side-Channel Attacks in Embedded Hardware

Éloi de Chérisey*, Sylvain Guilley†*, Olivier Rioul*, and Pablo Piantanida‡
*LTCI, Télécom Paris, Institut Polytechnique de Paris, 75013 Paris, France
†Secure-IC S.A.S., 39 rue Dareau, 75014 Paris, France
‡LSS, Centrale-Supélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

***Abstract*—Using information-theoretic tools, this paper establishes a mathematical link between the probability of success of a side-channel attack and the minimum number of queries to reach a given success rate, valid for *any* possible distinguishing rule and with the best possible knowledge on the attacker's side. This link is a lower bound on the number of queries, which depends on the mutual information between the traces and the secret key. This leads us to derive upper bounds on the mutual information that are as tight as possible and can be easily calculated. It turns out that, in the case of additive white Gaussian noise, the bound on the probability of success of any attack is directly related to the signal-to-noise ratio (SNR). This leads to easy computations and predictions of the success rate for any leakage model.**

I. Introduction

Since 1999 [8] side-channel analysis is considered as a serious threat against the security of ciphering chips. It exploits physical leakages (such as power consumption, electromagnetic radiation, execution time of the ciphering algorithm) to break the secret key byte per byte. This topic has grown widely over the last decades and theoretical approaches were proposed to understand the threat and optimize the attacks [6], [15].

We consider the following framework which was already leveraged at ISIT'2016 [12] for comparing different side-channel distinguishers. During the execution of the cryptographic algorithm (such as AES), some given random text byte vector $\mathbf{T} = (T_1, \ldots, T_q)$ is combined with the n-bit secret key byte K ($n = 8$ for AES) through a XOR operation noted $\oplus$. The secret key leaks through a so-called *sensitive variable* $\mathbf{Y} = (Y_1, \ldots, Y_q)$, given by the formula $Y_i = f(T_i \oplus K)$ where f is a deterministic function. In order to recover the actual key $k \in \mathcal{K}$, the attacker performs q measurements $\mathbf{X} = (X_1, \ldots, X_q) = \mathbf{Y} + \mathbf{N}$ under additive i.i.d. noise $\mathbf{N} = (N_1, \ldots, N_q)$. We assume the attacker also knows the corresponding plaintexts $\mathbf{T}$. She then computes a mathematical function called a *distinguisher* $\mathcal{D}$ that returns an estimation $\hat{K}$ of the secret key. As recalled in [6], the best distinguisher maximizes likelihood for uniformly distributed K:

$$\hat{K} = \mathcal{D}(\mathbf{X}, \mathbf{T}) = \arg\max_{k \in \mathcal{K}} \mathbb{P}(\mathbf{X}|\mathbf{T}, k). \tag{1}$$

This above framework is tailored for key extraction in symmetric cryptography from embedded systems (such as AES). We limit ourselves to such ideal setup, where physical side-channel measurements are perfectly synchronized and i.i.d. (no insertions nor deletions). Such model is classically used for differential power analysis [8], a major topic in hardware security which serves for real product evaluations (e.g., the 1000+ products listed in the Common Criteria portal[1]).

Recently, some more general frameworks have been proposed, such as the quantitative information flow [14] or the general operational approach of [7]. In contrast, we focus on the derivation of closed-form bounds on the probability of success, which cannot be estimated as a simple expression. These bounds should capture structural properties of the side channel (e.g., how it varies with the SNR), be easily computed, and be valid for any practical attack in order to be useful for real-world cryptographic chip designers.

Previous works such as [10] have already tackled similar issues, however either with approximations (assuming e.g., a law of large numbers), or using specific cryptographic assumptions [3] where constant terms are neglected in the proofs. In this paper, we aim at bridging the gap between Shannon's information theory and side-channel analysis in the context of cryptographic hardware security. The goal is to consolidate the state-of-the-art information-theoretic techniques for cryptographic evaluation, which may serve as a basis for further rigorous information-theoretic developments.

Arimoto [1] proved a lower bound of the error rate (hence an upper-bound of the success rate) in terms of a Gallager coefficient. However, not only its evaluation requires intensive computations, but also the model assumes a freely chosen input distribution. In our case, that input distribution is set by the leakage model and therefore, cannot be freely chosen. Arimoto's main result [1, Eq.(24)] remains true because it represents the best possible case for an attacker for all possible input distributions; but the resulting bound is very loose in our side-channel context. [1, Eq. (9)] could be used instead but depends on a parameter β which is difficult to optimize. Arimoto's bound requires multivariate integrations of degree q which is almost as complicated as a direct simulation of the success probability (q is very large in practice). Expressions of the probability of success can also be obtained in terms of the min-entropy, but this bears similar problem of estimation complexity.

In this work, we apply the converse coding theorem[2] and Fano's inequality to obtain upper bounds on the probability

[1] See website: https://commoncriteriaportal.org/products/stats/

[2] Shannon's (direct) coding theorem cannot apply here because no particular "coding" scheme is used: we consider the worst possible case for the designer (i.e., the best case for the attacker).

of success for *any* attack that depends on mutual information (MI) between secret and measurements. The obtained bounds on the success rate and number of measurements are *universal* in the sense that they are valid for any possible attack.

This remainder of this paper is organized as follows. Section II provides our main result and three different ways to exploit it. An application to additive Gaussian noise is carried out in Section III, where we show that knowledge of the SNR is enough to predict the security of a device.

II. Apply the Converse Shannon Theorem

The side-channel can be modeled as the "communication channel" [6], [15] shown in Fig. 1. From this figure we have

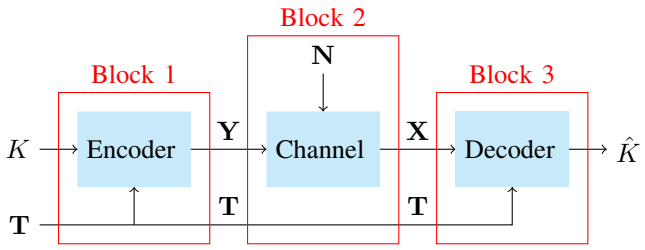

Fig. 1: Side-channel leakage seen as a communication channel

the Markov chain

$$(K, \mathbf{T}) \longrightarrow (\mathbf{Y}, \mathbf{T}) \longrightarrow (\mathbf{X}, \mathbf{T}) \longrightarrow \hat{K}. \tag{2}$$

Let us emphasize that we make no particular assumption on the actual distribution of texts $\mathbf{T}$. For instance, $\mathbf{T}$ can be pieces of ciphertext obtained from a block cipher under a mode operation running on unknown plaintext. Also, the model equally applies to situations where there are *countermeasures*, such as random masking or shuffling [13]. The countermeasure would be an unknown random variable explaining $\mathbf{Y}$ in block 1 of Fig. 1, along with $\mathbf{T}$.

A. Fundamental Lower Bound on $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

From (2) we have the data processing inequality [2]

$$I((K, \mathbf{T}) ; (\mathbf{X}, \mathbf{T})) \leq I((\mathbf{Y}, \mathbf{T}) ; (\mathbf{X}, \mathbf{T})). \tag{3}$$

We now expand both sides. Since the channel is memoryless and K and $\mathbf{T}$ are independent, the l.h.s. is

$$\begin{aligned} I((K, \mathbf{T}); (\mathbf{X}, \mathbf{T})) &= H(K, \mathbf{T}) - H((K, \mathbf{T}) \mid (\mathbf{X}, \mathbf{T})) \\ &= H(K) + H(\mathbf{T}) - H(K \mid \mathbf{T}, \mathbf{X}). \end{aligned}$$

As $\hat{K}$ is a deterministic function of $\mathbf{T}$ and $\mathbf{X}$, adding knowledge of $\hat{K}$ does not change the entropy:

$$\begin{aligned} I((K, \mathbf{T}); (\mathbf{X}, \mathbf{T})) &= H(K) + H(\mathbf{T}) - H(K \mid \mathbf{T}, \mathbf{X}, \hat{K}) \\ &\geq H(K) + H(\mathbf{T}) - H(K \mid \hat{K}). \end{aligned}$$

The latter inequality holds because conditioning can only reduce entropy [2]. Now by Fano's inequality [2],

$$H(K \mid \hat{K}) \leq H_2(\mathrm{P}_e) + \mathrm{P}_e \log_2(|\mathcal{K}| - 1)$$

where $\mathrm{P}_e = \mathbb{P}(K \neq \hat{K})$ is the probability of error. The probability of success is $\mathrm{P}_s = 1 - \mathrm{P}_e$ so that $H_2(\mathrm{P}_e) = H_2(\mathrm{P}_s) = -\mathrm{P}_e \log_2 \mathrm{P}_e - \mathrm{P}_s \log_2 \mathrm{P}_s$. Thus $H(K|\hat{K}) \leq H_2(\mathrm{P}_s) + (1 - \mathrm{P}_s) \log_2(2^n - 1)$, which gives

$$\begin{aligned} &I((K, \mathbf{T}); (\hat{K}, \mathbf{T})) \geq \\ &\quad H(K) + qH(T) - H_2(\mathrm{P}_s) - (1 - \mathrm{P}_s) \log_2(2^n - 1). \end{aligned} \tag{4}$$

On the other hand, the r.h.s. of (3) is:

$$\begin{aligned} I((\mathbf{Y}, \mathbf{T}); (\mathbf{X}, \mathbf{T})) &= H(\mathbf{X}, \mathbf{T}) - H(\mathbf{X}, \mathbf{T} \mid \mathbf{Y}, \mathbf{T}) \\ &= I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) + H(\mathbf{T}). \end{aligned} \tag{5}$$

Combining (4) and (5), we obtain the fundamental inequality:

$$H(K) - H_2(\mathrm{P}_s) - (1 - \mathrm{P}_s) \log_2(2^n - 1) \leq I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}). \tag{6}$$

which shows that there is a direct link between the probability of success of an attack P_s and mutual information (MI) $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$.

Remark 1: The expression $n + (\mathrm{P}_s - 1) \log_2(2^n - 1) - H_2(\mathrm{P}_s)$ is always non-negative for any P_s in the range $(0, 1)$ and vanishes if and only if $\mathrm{P}_s = 1/2^n$. Therefore, when there are no traces, $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) = 0$, the only probability that can respect inequality (6) is $\mathrm{P}_s = 1/2^n$, meaning that without information, the blind attacker cannot have a better success rate than $1/2^n$ obtained with an equiprobable random guess, as expected. Every trace will bring additional information and therefore increases the probability of success.

Remark 2: In the context of cryptanalysis, we are interested in *high* values of P_s. Indeed, for successful key extraction in practice, one needs to extract not only *one*, but, say, *sixteen* bytes of keys (e.g., in the case of AES-128). Attacks are thus based on a "divide-and-conquer" approach, and 128-bit key extraction results from the conjunction of 16 attacks, which must thus all be successful. Therefore, probability of success for each individual byte must be quite high. In this regime, it happens that Fano's inequality is fairly tight.

B. First Upper Bound on $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

Since $\mathbf{X}, \mathbf{Y}$ and $\mathbf{T}$ are vectors of length q, with the above i.i.d. assumption on the noise, it is easy to check that

$$I(\mathbf{X}, \mathbf{Y} \mid \mathbf{T}) \leq q \cdot I(X; Y \mid T) \tag{7}$$

with equality if and only if all random vectors are i.i.d. This trivial upper bound on MI is linear in q, and, therefore, will diverge as the number q of measurements increases.

However, in our side-channel model, the MI is always bounded by the number n of key bits since $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) \leq H(Y \mid T) \leq H(K) \leq n$. Thus (7) will eventually be loose as q increases. Therefore, it is important to derive another bound that will converge to n as $q \to \infty$. This is done next.

C. Second Upper Bound on $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$

We need the following lemma, proved in Appendix A.

Lemma 1: For any random variables X and Y and real-valued function $(x, y) \mapsto f(x, y)$,

$$-\mathbb{E}_Y \log_2 \mathbb{E}_X[\exp(f(X, Y))] \leq -\log_2 \mathbb{E}_X[\exp(\mathbb{E}_Y f(X, Y))].$$

The following corollary is proved in Appendix B.

Corollary 1: For any random variables X and Y and positive function $(x,y) \mapsto g(x,y)$,

$$\exp \mathbb{E}_Y \log_2 \mathbb{E}_X[g(X,Y)] \geq \mathbb{E}_X[\exp(\mathbb{E}_Y \log g(X,Y))]$$

Equipped with these inequalities, we compute MI as follows:

$$\begin{aligned} I(\mathbf{X}; K \mid \mathbf{T}) &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_{\mathbf{X},K|\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K\mathbf{T})}{\mathbb{P}(\mathbf{X} \mid \mathbf{T})} \\ &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K\mathbf{T})}{\mathbb{P}(\mathbf{X} \mid \mathbf{T})}. \end{aligned}$$

Let K' be an independent copy of K.

$$\begin{aligned} I(\mathbf{X}; K \mid \mathbf{T}) &= \mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})}{\mathbb{E}_{K'} \mathbb{P}(\mathbf{X} \mid K', \mathbf{T})} \\ &= -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \mathbb{E}_{K'} \frac{\mathbb{P}(\mathbf{X} \mid K', \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})} \end{aligned}$$

Thus

$$\begin{aligned} I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) = -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \\ \log_2 \mathbb{E}_{K'} \exp \left[\log_2 \frac{\mathbb{P}(\mathbf{X} \mid K', \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})} \right] \end{aligned}$$

By Lemma 1 we obtain

$$\begin{aligned} & I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T}) \\ & \leq -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \log_2 \mathbb{E}_{K'} \exp \left[\mathbb{E}_{\mathbf{X}|K,\mathbf{T}} \log_2 \frac{\mathbb{P}(\mathbf{X} \mid K', \mathbf{T})}{\mathbb{P}(\mathbf{X} \mid K, \mathbf{T})} \right] \\ & = -\mathbb{E}_{\mathbf{T}} \mathbb{E}_K \log_2 \mathbb{E}_{K'} \exp \left[-D(\mathbb{P}_{\mathbf{X}|K,\mathbf{T}} \,||\, \mathbb{P}_{\mathbf{X}|K',\mathbf{T}}) \right]. \quad (8) \end{aligned}$$

which is our second upper-bound of $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{T})$.

Remark 3: This second bound improves on top of (6) derived from Fano's inequality, and actually resorts to relative entropy $D(\mathbb{P}_{\mathbf{X}|K,\mathbf{T}} \,||\, \mathbb{P}_{\mathbf{X}|K',\mathbf{T}})$. While Shannon's entropy is perhaps not always the best metric, in our situation, it appears to be tractable and will be seen to perform very well (in contrast to Arimoto's bound).

D. Graphical Comparison

In order to visualize the difference between the two upper bounds (7), (8) given above, we have plotted in Figure 2 the mutual information $I(\mathbf{X}; \mathbf{Y} \,|\, \mathbf{T} = \mathbf{t})$, where $\mathbf{t}$ is a fixed balanced vector. The leakage model chosen is given by the equation

$$y(k, t_i) = \mathrm{H_w}(\mathrm{S_{box}}(t_i \oplus k)) \qquad (i = 1, 2, \ldots, q)$$

where $\mathrm{H_w}(\cdot)$ is the Hamming weight (of the value written in binary), and $\mathrm{S_{box}}(\cdot)$ is the AES substitution box [11]. We assumed a zero-mean additive white Gaussian noise with standard deviation $\sigma = 4$, hence a signal-noise ratio $\mathrm{SNR} = 1/8$. The results of Figure 2 were obtained by Monte-Carlo simulation. Notice that as expected in Subsection II-B, the first upper bound is linear in q; and as expected in Subsection II-C, the second upper bound converges to $H(K) = n = 8$.

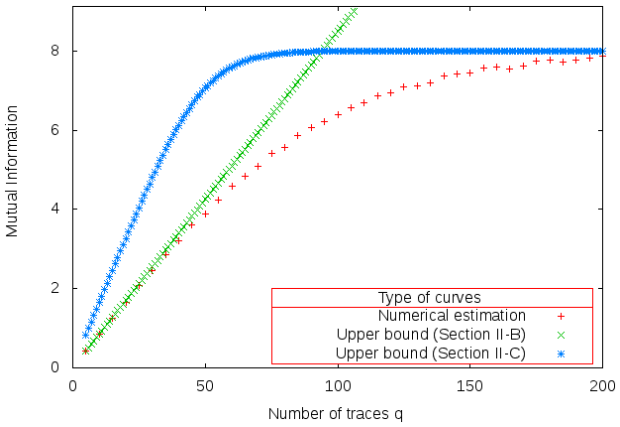

Fig. 2: Comparison of the two upper bounds of subsections II-B and II-C, for $n = 8$.

III. Application to AWGN

In this section, we develop the results of Section II for leakages with additional white Gaussian noise (AWGN). This is the most common case for attacks such as DPA, where the noise comes from other algorithmic computations running during the execution of the ciphering algorithm and is usually modeled as Gaussian.

With this model, we can link the success rate to Shannon's capacity $C = \frac{1}{2}\log(1 + \mathrm{SNR})$, and therefore, to the $\mathrm{SNR} = \mathrm{Var}(Y)/\sigma^2$.

A. Shannon's Channel Capacity

Under the AWGN assumption, it is easily seen that the scalar mutual information $I(X; Y \mid T)$ does not exceed Shannon's capacity:

$$\begin{aligned} I(X; Y \mid T) &= \mathbb{E}_T I(X; Y \mid T = t) \\ &= H(f(K) + N) - H(X \mid Y) \\ &\leq \frac{1}{2} \log_2(2\pi e(\mathrm{Var}_K(f(K)) + \mathrm{Var}(N))) - H(X \mid Y) \\ &= \frac{1}{2} \log_2(1 + \mathrm{SNR}). \end{aligned}$$

Combining this with inequality obtained in Subsection II-B yields a lower bound on the number of traces to reach a given probability of success:

$$q \geq \frac{n + (\mathrm{P}_s - 1)\log_2(2^n - 1) - H_2(\mathrm{P}_s)}{\frac{1}{2}\log_2(1 + \mathrm{SNR})}. \quad (9)$$

Remark 4: The number of traces q needed to recover the key reliably is lower-bounded by:

$$\lim_{\mathrm{P}_s \to 1} q \geq \frac{n}{\frac{1}{2}\log_2(1 + \mathrm{SNR})}. \quad (10)$$

However, because as we have seen the MI can never be higher than $H(K)$, the above constant bound is not accurate for real attacks. The next subsection provides a much more accurate estimation.

B. Evaluation of the Relative Entropy

The upper bound (8) has a relative entropy (divergence) term that depends on $\mathbb{P}_{\mathbf{X}|K,\mathbf{T}}$. In the AWGN model, $\mathbb{P}_{\mathbf{X}|K,\mathbf{T}}$ follows a multivariate normal distribution $\mathcal{N}(\mathbf{y}(K,\mathbf{T}),\sigma^2 I_q)$. For such distributions, the relative entropy is very easy to compute as the covariance matrix is diagonal. It is easily found that

$$D(\mathbb{P}_{\mathbf{X}|K,\mathbf{T}}\|\mathbb{P}_{\mathbf{X}|K',\mathbf{T}}) = \frac{\|\mathbf{y}(K,\mathbf{T})-\mathbf{y}(K',\mathbf{T})\|_2^2}{2\sigma^2}.$$

The upper-bound (8) becomes

$$\begin{aligned} &n+(\mathrm{P}_s-1)\log_2(2^n-1)-H_2(\mathrm{P}_s) \\ &\quad\le -\mathbb{E}_{\mathbf{T}}\mathbb{E}_K \log_2 \mathbb{E}_{K'} \exp\Big(-\frac{\|\mathbf{y}(K,\mathbf{T})-\mathbf{y}(K',\mathbf{T})\|_2^2}{2\sigma^2}\Big). \end{aligned} \tag{11}$$

This is easy to evaluate with numerical computations. We note that it is also reminiscent of so-called *generalized confusion coefficients* [5, Def. 8]:

$$\kappa(k,k') = \mathbb{E}_T\left(\Big[\frac{y(k,T)-y(k',T)}{2}\Big]^2\right). \tag{12}$$

C. Example for Monobit Leakage

In order to compare these bounds with practical cases we consider a *monobit* leakage model:

$$f(t_i\oplus k) = \mathrm{LSB}(\mathrm{S}_{\mathrm{box}}(t_i\oplus k)) \qquad (i=1,2,\ldots,q)$$

where $\mathrm{S}_{\mathrm{box}}$ is the AES substitution box and LSB is the least significant bit of a bit vector. Figure 3 represents the success rate of a monobit leakage under AWGN with $\sigma = 4$. The distinguisher used is the maximum likelihood distinguisher which is optimal [6]. The graphical plot of the success rate

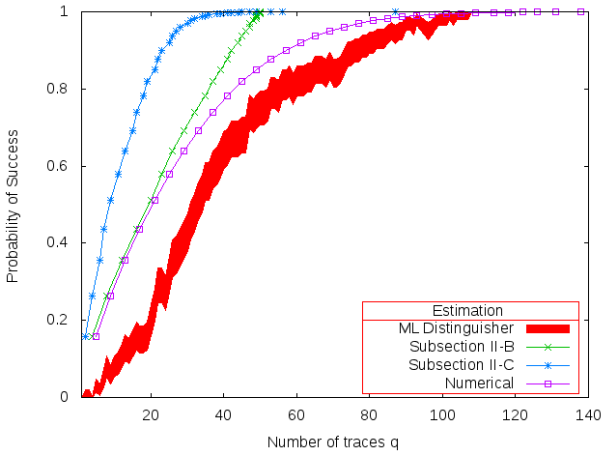

Fig. 3: Success rates with monobit leakage for $\sigma = 1$

(in red) follows [9] where error bars are taken into account. The other curves are the bounds obtained with:

- a numerical estimation of $I(\mathbf{X};\mathbf{Y}|\mathbf{T})$ (by a Monte-Carlo simulation);
- MI's upper bound (7);
- MI's upper bound (8).

The three bounds curves lie above the success rate curve as expected, the one obtained with a numerical estimation of $I(\mathbf{X};\mathbf{Y}\,|\,\mathbf{T})$ being the tightest (since it gives the closest approximation of the MI). The two other curves obtained from (7) and (8) are not as tight as the bound obtained with a Monte-Carlo estimation but very easy to compute.

D. Example for Hamming Weight Leakage

In practice, the AES algorithm computes SubBytes based on bytes (i.e., with 8 bits), which impacts the leakage function:

$$y_i = f(t_i\oplus k) = \mathrm{H_w}(\mathrm{S}_{\mathrm{box}}(t_i\oplus k)) \qquad (i=1,2,\ldots,q)$$

where $\mathrm{S}_{\mathrm{box}}$ is the AES substitution box and $\mathrm{H_w}$ is the Hamming weight function. Figure 4 shows the success rate compared with the three other types of estimation under AWGN with $\sigma = 1$. Once again, we observe that all bounds

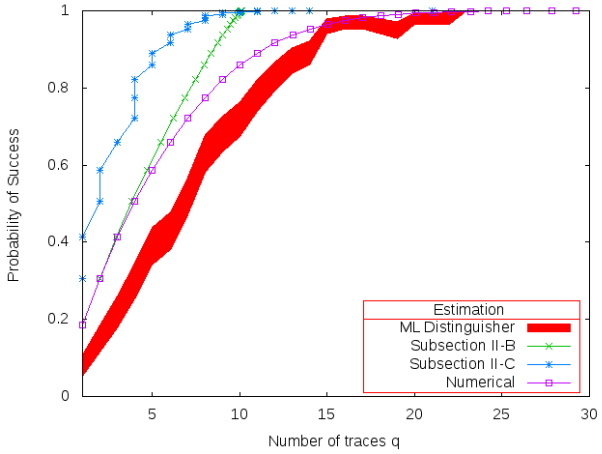

Fig. 4: Success rate for a Hamming weight leakage for $\sigma = 1$

are above the success rate of the optimal distinguisher and that the closest estimation of the MI gives the tightest bound.

IV. Conclusion

In this paper, we have linked two metrics used in the field of side-channel analysis: the probability of success of an attack (computed as the success rate) and the mutual information between the leaked traces and the secret key. Our results are of interest to better understand the different factors that impact the success rate of an attack.

We obtained *universal* bounds to the success rate, in the sense that they are independent of what the attacker may exploit with the measurements. This can be seen as an advantage for designers since in practice, they are never able to know how their devices will be attacked in the future. This work is a first step to allow them ensure a minimal security of their device in *any* adversarial context.

While our results presented in the paper lie within the specific framework of "power-line attacks" (e.g., monobit leakage or Hamming weight leakage), we emphasize that our theoretical development is fairly general since it requires only the assumption of the Markov chain (2). In fact, our bounds

are somewhat generic in that they do not depend on the leakage model itself, but rather on an aggregated quantity of it (measuring the link between Y and (K,T)), namely $I(\mathbf{X};\mathbf{Y} \mid \mathbf{T})$ or $D(\mathbb{P}_{\mathbf{X}|K1,\mathbf{T}} \| \mathbb{P}_{\mathbf{X}|K',\mathbf{T}})$, similar to works such as [4], [16] which rely on confusion coefficients (12).

The bounds obtained in this paper also happen to be empirically tight, hence relevant for the business of Common Criteria evaluation, and more generally for designers who seek more precise tools to secure their cryptographic chips.

Appendix A
Proof of Lemma 1

Let X' be an independent copy of X. Consider the difference

$$\begin{aligned}
\Delta &= -\mathbb{E}_Y \log \mathbb{E}_X [\exp f(X,Y)] + \log \mathbb{E}_X \exp \mathbb{E}_Y f(X,Y) \\
&= -\log \exp \mathbb{E}_Y \log \mathbb{E}_{X'} \exp f(X',Y) \\
&\quad + \log \mathbb{E}_X \exp \mathbb{E}_Y \log \exp f(X,Y) \\
&= \log \mathbb{E}_X \frac{\exp \mathbb{E}_Y \log \exp f(X,Y)}{\exp \mathbb{E}_Y \log \mathbb{E}_{X'} \exp f(X',Y)} \\
&= \log \mathbb{E}_X \exp \mathbb{E}_Y [\log \exp f(X,Y) - \log \mathbb{E}_{X'} \exp f(X',Y)] \\
&= \log \mathbb{E}_X \exp \mathbb{E}_Y \left[\log \frac{\exp f(X,Y)}{\mathbb{E}_{X'} \exp f(X',Y)} \right].
\end{aligned}$$

By concavity of the logarithm,

$$\begin{aligned}
\Delta &\leq \log \mathbb{E}_X \exp \log \mathbb{E}_Y \left[\frac{\exp f(X,Y)}{\mathbb{E}_{X'} \exp f(X',Y)} \right] \\
&= \log \mathbb{E}_X \mathbb{E}_Y \left[\frac{\exp f(X,Y)}{\mathbb{E}_{X'} \exp f(X',Y)} \right] \\
&= \log \mathbb{E}_Y \left[\frac{\mathbb{E}_X \exp f(X,Y)}{\mathbb{E}_{X'} \exp f(X',Y)} \right] = \log \mathbb{E}_Y [1] = 0.
\end{aligned}$$

Appendix B
Proof of Corollary 1

Lemma 1 reads

$$\mathbb{E}_Y \log \mathbb{E}_X [\exp(f(X,Y))] \geq \log \mathbb{E}_X [\exp(\mathbb{E}_Y f(X,Y))]$$

where the r.h.s. rewrites as $\log \mathbb{E}_X [\exp(\mathbb{E}_Y \log \exp f(X,Y))]$. Setting $g(x,y) = \exp(f(x,y))$, we obtain

$$\mathbb{E}_Y \log \mathbb{E}_X [g(X,Y)] \geq \log \mathbb{E}_X [\exp(\mathbb{E}_Y \log g(X,Y))].$$

which taking the exponential on both sides gives the result.

References

[1] S. Arimoto, "On the converse to the coding theorem for discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 19, no. 3, pp. 357–359, May 1973.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[3] A. Duc, S. Faust, and F. Standaert, "Making masking security proofs concrete—or how to evaluate the security of any leaking device," in Proc. *Advances in Cryptology - EUROCRYPT* 2015, Sofia, Bulgaria, Apr. 26-30, 2015, Lecture Notes in Computer Science, vol. 9056, Springer, 2015, pp. 401–429.

[4] Y. Fei, A. A. Ding, J. Lao, and L. Zhang, "A statistics-based success rate model for DPA and CPA," *J. Cryptographic Engineering*, vol. 5, no. 4, pp. 227–243, 2015.

[5] S. Guilley, A. Heuser, and O. Rioul, "A key to success: Success exponents for side-channel distinguishers," in Proc. *Progress in Cryptology - INDOCRYPT 2015, Bangalore, India, Dec. 6-9, 2015*, Lecture Notes in Computer Science, vol. 9462, Springer, 2015, pp. 270–290.

[6] A. Heuser, O. Rioul, and S. Guilley, "Good is not good enough: Deriving optimal distinguishers from communication theory," in Proc. *CHES*, Busan, South Korea, Sept. 23-26, 2014, Lecture Notes in Computer Science, vol. 8731, Springer, 2014, pp. 55–74.

[7] I. Issa, A. B. Wagner, and S. Kamath, "An Operational Approach to Information Leakage," preprint, 2018. [Online]. Available: http://arxiv.org/abs/1807.07878

[8] P. C. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *CRYPTO*, Lecture Notes in Computer Science, vol. 1666, Springer, 1999, pp. 388–397.

[9] H. Maghrebi, O. Rioul, S. Guilley, and J.-L. Danger, "Comparison between side-channel analysis distinguishers," in *ICICS*, Lecture Notes in Computer Science, vol. 7618, Springer, 2012, pp. 331–340.

[10] S. Mangard, E. Oswald, and F. Standaert, "One for all—all for one: Unifying standard differential power analysis attacks," *IET Information Security*, vol. 5, no. 2, pp. 100–110, 2011.

[11] NIST/ITL/CSD, "Advanced Encryption Standard (AES). FIPS PUB 197," Nov 2001, http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf (also ISO/IEC 18033-3:2010).

[12] O. Rioul, A. Heuser, S. Guilley, and J. Danger, "Inter-class vs. mutual information as side-channel distinguishers," in Proc. *ISIT* 2016, Barcelona, Spain, July 10-15, 2016, pp. 805–809.

[13] M. Rivain, E. Prouff, and J. Doget, "Higher-order masking and shuffling for software implementations of block ciphers," in *CHES 2009*, Lausanne, Switzerland, Lecture Notes in Computer Science, vol. 5747, Springer, Sept. 6-9, 2009, pp. 171–188.

[14] G. Smith, "On the foundations of quantitative information flow," in Proc. *FOSSACS 2009, York, UK, Mar. 22-29, 2009*, Lecture Notes in Computer Science, vol. 5504, Springer, 2009, pp. 288–302.

[15] F.-X. Standaert, T. Malkin, and M. Yung, "A unified framework for the analysis of side-channel key recovery attacks," in *EUROCRYPT*, Köln, Germany, Apr. 26-30, 2009, Lecture Notes in Computer Science, vol. 5479, Springer, pp. 443–461.

[16] A. Thillard, E. Prouff, and T. Roche, "Success through confidence: Evaluating the effectiveness of a side-channel attack," in *CHES*, Lecture Notes in Computer Science, vol. 8086, Springer, 2013, pp. 21–36.