Confident Leakage Assessment - A Side-Channel Evaluation Framework based on Confidence Intervals

Florian Bache[1], Christina Plump[2], and Tim Güneysu[1,3]
[1]Horst Görtz Institute for IT Security, Ruhr-University Bochum, 44780 Bochum, Germany
[2]University of Bremen, 28359 Bremen, Germany
[3]Cyber Physical Systems, DFKI GmbH, 28359 Bremen, Germany
Email:{florian.bache, tim.gueneysu}@rub.de christina.plump@uni-bremen.de

***Abstract*—Cryptographic devices that potentially operate in hostile physical environments need to be secured against side-channel attacks. In order to ensure the effectiveness of the required countermeasures, scientists, developers, and evaluators need efficient methods to test the security level of a device. In this paper we propose a new framework based on confidence intervals that extends established t-test based approaches for test-vector leakage assessment (TVLA). In comparison to previous TVLA approaches the new methodology does not only enable the detection of leakage but can also assert its absence. The framework is robust against noise in the evaluation system and thereby avoids false negatives. These improvements can be achieved without overhead in measurement complexity and with a minimum of additional computational costs compared to previous approaches.**

We evaluate our method under realistic conditions by applying it to a protected implementation of AES.

I. Introduction

Physical attacks impose serious threats towards all exposed hardware containing cryptographic functions such as smart cards, hardware security modules or IoT devices. In this paper we focus on the subgroup of passive side-channel attacks that exploit power consumption or electro-magnetic emanation generated by a device in order to reveal its secrets [1], [2]. A designer needs to incorporate effective countermeasures, such as hiding [3] or masking [4]–[6], in order to counter these types of attacks.

Test Vector Leakage Assessment (TVLA) methods are frequently used to validate the effectiveness of countermeasures in cryptographic devices. These methods avoid the very costly application of many known attacks and rather use generic approaches to provide statements about physical security. In (moments-based) TVLA, the relevant side-channel, e.g., power consumption, is measured under different inputs, yielding side-channel traces. Then the evaluation procedure tries to decide whether the statistical moments of these traces are distinguishable. A widely adopted scheme (e.g. [7], [8]) is based on Welch's t-test that decides if the mean values of two random variables are different using the t-statistic. It should be noted that TVLA can not assert a device's security in all cases. It may fail if the noise present in a device is too low [9].

a) Motivation: A main feature of leakage detection schemes is their ability to assure the presence of leakage in a cryptographic computation with a given confidence $1 - \alpha$. This is possible because hypothesis tests, such as the often-used Welsh's t-test, are designed to limit the error-probability for false-positive results (α-error) – in the case of side-channel analysis this would correspond to detected leakage – to an arbitrary level α. The downside of this approach is missing assurance about the error-probability for false-negative results (β-error). Therefore, these methods cannot support the statement: ”No leakage is present”. In consequence, a t-test based SCA evaluation is not guaranteed to find existing leakage in a device. A negative result merely proves the inability of the test to find leakage, independent of its actual presence.

This issue is strongly related to the dependence of the t-test's result on the sample size. As the (Welch's) t-statistic is computed as

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{s_x}^2}{n_x} + \frac{{s_y}^2}{n_y}}} \propto \sqrt{n}\,,$$

given a positive absolute difference in the sample means, the test statistic will increase as the sample size increases and the sample means converge to the means of the underlying distributions. If the difference in means is small and/or the the variances are high, a very large n is required to detect leakage. Therefore, an insecure implementation might be evaluated as secure if the sample size is insufficient.

If a fixed threshold (e.g. [7], [8]) is used, current TVLA methods fail to account for the number of sample points in the measurements. When measuring the leakage of a device, a trace with more sample points will generally have a higher maximal t-value than a trace with less horizontal resolution. This problem was identified by the authors of [10].

b) Contribution: We develop a new framework for TVLA based on confidence intervals. This solves two related problems with hypothesis test approaches:

- It allows the establishment of an upper bound for leakage that is robust against noisy evaluation systems. This allows statements about an implementations security even if a t-test can not identity leakage.
- It provides natural cut-off values for the number of measurements required in order to assert or reject security claims about cryptographic implementations. By choosing an allowable leakage level, measurements can be stopped when either the maximum lower limit or the maximum upper limit crosses that threshold.

978-3-9819263-0-9/DATE18/©2018 EDAA

We also solve the problem of sample-point dependence by applying the Šidák correction to confidence intervals. Finally, we demonstrate the effectiveness of our framework in a case study.

II. Confidence-Interval based Leakage Detection

In the following section we will introduce a new method to detect the presence of leakage based on confidence intervals as well as discuss some improvements of our basic methodology. To clarify formalisms the following paragraph introduces the notation used throughout the paper.

Notation: Uppercase letters (X) denote random variables, whereas lowercase letters (x) denote their realization through a random sample. Sample sizes are denoted by n, where a subscript denotes the affiliated random variable when necessary. Arithmetic means (as random variable as well as realization) are denoted with $\bar{X}$, $\bar{x}$ resp. . The sample variance (again in both notations) is denoted with S_X^2, the sample standard deviation with S_X. Parameters of a normally distributed random variable are μ and σ, subscripts denoting the affiliated random variable. As for the normal distribution we have $\mu_X = E(X)$ and $\sigma_X^2 = E((X - E(X))^2) = Var(X)$ and use this denotations interchangeably as well as μ and σ^2 for higher moments (normal and centralized with even order) with the respective subscript. A $1-\alpha$ quantile of the t-distribution is denoted by $t_{1-\alpha}$. If the degree of freedom is relevant or unclear from context, it is added as a subscript.

A. From Statistical Tests to Confidence Intervals

In t-test TVLA, the comparison with a fixed number (e.g. the common 4.5) relates to the comparison of the computed t-value from the evaluation to a quantile of the t-distribution (hence the name t-test) for a significance level α. The significance level α states the probability that although the null-hypothesis (*there is no leakage*) is true, the evaluator accepts the alternative hypothesis (*there is leakage*) - and thus makes a mistake. This mistake can be controlled through choice of α and accordingly the corresponding quantile of the t-distribution. Therefore, a smaller α corresponds to a higher confidence when accepting the alternative hypothesis (*there is leakage*). It does not, however, yield any confidence for accepting the null-hypothesis (*there is no leakage*). The probability of making a mistake when deciding for the null-hypothesis, i.e. when instead the alternative hypothesis is true, the so called β-error, can not be determined a priori.

A standard method known in statistics to cope with the problems that arise when using hypothesis tests is the introduction of confidence intervals. They are constructed in a similar way but their significance level can be determined without assuming the null-hypothesis to hold. Confidence intervals are usually constructed symmetrically around a good estimator of the investigated parameter. Given a random sample and a desired confidence, the following statement holds: With a probability of said confidence the process to construct a confidence interval yields an interval that contains the unknown parameter.

We are now interested in a confidence interval for the absolute difference of means of two random samples of power consumption stemming from an implementation of a cryptographic primitive under different inputs. To that end, given samples of the random variables X and Y, two confidence intervals for the difference of means $\bar{X} - \bar{Y}$ and $\bar{Y} - \bar{X}$ are constructed for a given confidence. These are then combined to the required interval for $\left|\bar{X} - \bar{Y}\right|$. As we are now combining two statistical and thus probabilistic statements, the confidence for the final interval differs from its predecessors, namely the actual error α_t varies between $[\alpha; 2\alpha]$, where 2α is the original error. Turning this procedure around, we devised the following framework to compute a confidence interval with a given confidence $1 - \alpha_t$ for the absolute difference of means on the basis of given random samples:

1) Choose a confidence level $1 - \alpha_t$.
2) Draw samples from both random variables X and Y.
3) Compute $\widetilde{t}_{X,Y} = \frac{\bar{X}-\bar{Y}}{\widetilde{s}_n}$ with $\widetilde{s}_n = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$
4) Interpolate $\alpha \in \left[\frac{\alpha_t}{2}, \alpha_t\right]$ and compute a lower and upper bound $\Delta_{min}, \Delta_{max}$:
   a) $|\widetilde{t}_{X,Y}| \leqslant t_{1-\alpha_t}$:
   $$\begin{aligned} \alpha_t &= \alpha + T(-2|\widetilde{t}_{X,Y}| - t_{1-\alpha}) \\ \Delta_{min} &= 0 \\ \Delta_{max} &= \widetilde{s}_n(|\widetilde{t}_{X,Y}| + t_{1-\alpha}) \end{aligned}$$
   b) $|\widetilde{t}_{X,Y}| > t_{1-\alpha_t}$:
   $$\begin{aligned} \alpha_t &= 2\alpha - T(2\widetilde{t}_{X,Y} + t_{1-\alpha}) + T(2\widetilde{t}_{X,Y} - t_{1-\alpha}) \\ \Delta_{min} &= \widetilde{s}_n\left(|\widetilde{t}_{X,Y}| - t_{1-\alpha}\right) \\ \Delta_{max} &= \widetilde{s}_n\left(|\widetilde{t}_{X,Y}| + t_{1-\alpha}\right) \end{aligned}$$

If the random samples are drawn identically and independently distributed from X (Y resp.), the construction of $I = [\Delta_{min}, \Delta_{max}]$ yields with probability $1 - \alpha_t$ an interval that contains $|\mu_X - \mu_Y|$.

Theorem 1. *Let $X \sim \mathcal{N}(\mu_X, \sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ be normally distributed random variables with all their parameters unknown. Then $|\mu_X - \mu_Y| \in [\Delta_{min}, \Delta_{max}]$ with confidence of $1 - \alpha_t$, where Δ_{min} and Δ_{max} are computed as above.*

Proof. For brevity we will only sketch the proof here.
The proof works in three steps. First, the intervals containing the absolute difference are determined. Second, the probability for those intervals is determined. Third, the turnaround to yield a framework needs to be verified.
Let X,Y be distributed as in the theorem and $X_i \overset{iid}{\sim} X$ and $Y_i \overset{iid}{\sim} Y$ two random samples of size n_X and n_Y of X and Y. From basic statistics we then know that

$$T_{X,Y} = \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \overset{approx.}{\sim} t_\nu \tag{1}$$

with

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X-1}\left[\frac{s_X^2}{n_X}\right]^2 + \frac{1}{n_Y-1}\left[\frac{s_Y^2}{n_Y}\right]^2} \overset{n_X \approx n_Y}{\sim} n_X - 1 \tag{2}$$

the estimated degree of freedom from the sample. Standard algebraic techniques then yield a confidence interval for $\mu_X - \mu_Y$ for a given confidence $1 - 2\alpha$:

$$\begin{aligned}\mathbb{P}\Big(\mu_X - \mu_Y \in \Big[\quad & (\bar{X} - \bar{Y}) - t_{\nu,1-\alpha} \cdot \widetilde{S}_n; \\ & (\bar{X} - \bar{Y}) + t_{\nu,1-\alpha} \cdot \widetilde{S}_n \Big]\Big) \\ & = 1 - 2\alpha\end{aligned}$$

where $\widetilde{S}_n = \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$ for brevity. Repeating that calculation for $\mu_Y - \mu_X$, we receive the mirrored confidence interval with the same confidence. As we are only interested in the distance between means and not its direction, we can now combine these intervals to one, which yields after some algebra and case-by-case analysis:

$$\begin{aligned}I &= \left[0; \widetilde{S}_n \left(\left|\widetilde{t}_{X,Y}\right| + t_{1-\alpha}\right)\right] \; if \; \left|\widetilde{t}_{X,Y}\right| \leqslant t_{1-\alpha} \\ I &= \left[\widetilde{S}_n \left(\left|\widetilde{t}_{X,Y}\right| - t_{1-\alpha}\right); \widetilde{S}_n \left(\left|\widetilde{t}_{X,Y}\right| + t_{1-\alpha}\right)\right] \; else\end{aligned}$$

This concludes step one.
The second step investigates what confidence those intervals carry. It is clear, that this confidence should vary between $1 - 2\alpha$ and $1 - \alpha$, as it should at least have the confidence of one single directed confidence interval and might have, at best, half that error probability (this only exactly occurs when $\left|\widetilde{t}_{X,Y}\right| = t_{1-\alpha}$). Using probability theory, one obtains

$$\begin{aligned}\alpha_t &= \alpha + T(-2|\widetilde{t}_{X,Y}| - t_{1-\alpha}) \; if \; \left|\widetilde{t}_{X,Y}\right| \leqslant t_{1-\alpha} \\ \alpha_t &= 2\alpha - T(2\widetilde{t}_{X,Y} + t_{1-\alpha}) + T(2\widetilde{t}_{X,Y} - t_{1-\alpha}) \; else\end{aligned}$$

Finally, to be able to distinct cases only with the knowledge of $t_{1-\alpha_t}$ and not $t_{1-\alpha}$, we show that $\left|\widetilde{t}_{X,Y}\right| > t_{1-\alpha}$ if $\alpha \in \left[\frac{\alpha_t}{2}; \alpha_t\right]$, completing the proof. □

a) Relaxation of assumptions: As introduced above, the statistics hold for normally distributed random variables. However, in the case of power consumption the distribution is not clear. Nevertheless, it is not per se necessary for the distribution of X and Y to follow the normal distribution but for their arithmetic means $\bar{X}$ and $\bar{Y}$. Fortunately, due to the central limit theorem, as soon as the sample size n increases, the arithmetic mean of random variables with arbitrary (but still iid.) distribution follows a normal distribution [11]. Furthermore, the condition of independence is not a strict one. A small level of dependence in drawing the random samples (as is to be expected when using the same device for the measurement process) is tolerable (if n is big enough) as it does not influence the distribution of the respective arithmetic means. This results in the applicability of our approach for the given task at hand.

b) Comparison with t-Test Method: The resulting confidence interval subsumes the t-test's result. If the confidence interval has a positive lower bound (and is computed with the same significance level as the t-tests), this will be equivalent to the t-test asserting that there is some leakage. At the same time, the confidence interval does also contain information about the magnitude of the leakage. The closer the lower bound is to zero, the tighter the decision towards leakage detection has been. If the lower bound is zero, this will be equivalent to the t-test failing to demonstrate any leakage. In addition to making both decisions statistically *valid*, the confidence interval method bounds the maximal difference between the means of both power consumptions (with a significance level). That is, it is possible to make a statement of the form: With a confidence of $1 - \alpha_t$, both means do not differ more than the value of the upper bound.

c) Influence of parameters: There are two ways of looking at this method's parameters. First, to interpret the results (e.g. length of the confidence interval) and second, for the evaluator to construct the sample in a way that some properties are satisfied in the end.

- For a given confidence level, the length of a confidence interval is initially influenced by two factors: Deviation and sample size. Obviously, they work in opposite directions. A higher sample size will lead to a smaller confidence interval and a higher deviation will lead to a larger confidence interval. Additionally, the interval's length is also influenced by the position of $\widetilde{t}_{X,Y}$ in relation to $t_{1-\alpha_t}$. This stems from the fact, that the computed α from the above framework is closer to α_t if $\widetilde{t}_{X,Y}$ is close to $t_{1-\alpha_t}$ - thus resulting in a smaller confidence interval. If $\widetilde{t}_{X,Y}$ is closer to zero or its absolute value is large, α will converge to $\frac{\alpha_t}{2}$ and thus, the confidence interval will turn out to be relatively larger.
- For a given sample (and thus fixed sample size and deviation) the length of a confidence interval is bigger, if the confidence level $1 - \alpha_t$ is higher (or the probability of error smaller).

If the evaluator wants to keep the length of the confidence interval in a certain range and also wants his confidence to fulfill some requirements, he needs to specify his sample size accordingly and, additionally, needs to keep his deviation as small as possible, i.e., reduce measurement noise. This is in contrast to t-test based evaluation, where an unsuitable measurement system may indicate a false sense of security by producing low t-test results or requiring a high sample size.

B. Family-wise Error Rate Correction

Up until now, we only considered one timepoint j at a time. For each of those timepoints the introduced procedure yields an intervall $[\Delta_{min,j}, \Delta_{max,j}]$ that contains the expected difference of means with significance level of α.

However, to evaluate an implementation in terms of leakage, we are not only interested in every single point but in a statement of the form: With significance level α the difference of means (expected values) is contained in its respective interval at every time point.

Generally speaking, if we have m timepoints of interest to us each with their respective interval $I_j = [\Delta_{min,j}, \Delta_{max,j}]$ and a significance level $\alpha_{t,j}$, the statement

$$\forall j \in \{1, ..., m\} : |\mu_X - \mu_Y| \in [\Delta_{min,j}, \Delta_{max,j}] \qquad (3)$$

holds (under the assumption of independence) with significance level $\alpha_{total} = 1 - \prod_{j=1}^{m} (1 - \alpha_{t,j})$.

But this results in three problems: First, we can not simply assume independence. Second, α_{total} is obviously highly dependent on m which should not be the case (that would build an incentive for an evaluator to test less timepoints to get a better confidence level). Third, if we assume sensible significance levels, α_{total} converges to 1 very fast.

To adress the second and third problem, we propose the usage of Šidák correction [12].

1) Choose total confidence level $1 - \alpha_{total}$
2) Compute $1 - \alpha_t = \sqrt[m]{1 - \alpha_{total}}$
3) Execute the above procedure with $1 - \alpha_t$ as confidence level for every time point $1 \leqslant j \leqslant m$.

This method yields confidence intervals I_j for every time point, such that statement 3 holds with confidence $1 - \alpha_{total}$. This solves problems two and three to the extent that it shifts the problem towards the length of the confidence intervals. In this setting it is possible for the evaluator to set a confidence at which he wants to evaluate the given implementation which is then by design neither close to 1 nor dependent on m. Obviously, a large m constrains α_t to be rather small and thus the confidence intervals to be larger. Still, the resulting confidence intervals contain useful information (instead of an α_{total} that converges to 1) and as the size of $t_{1-\alpha_t}$, which is relevant to the confidence intervals' size, is highly affected by the sample size n, i.e. how many traces an evaluator uses, it is possible to counter large values of m with large values of n.

Problem 1, independence, is both more critical and less at the same time. On the one hand, it is unclear what kind of dependence exits and how it was to be measured if one wanted to include it in an evaluation. On the other hand, the Šidák correction gives a pessimistic computation of the Family-Wise Error Rate (FWER), assuming positive dependencies (which we can assume in an evaluation setting). That is, confidence intervals will only get smaller, when dependencies are accounted for. Furthermore, the potential multivariate dependence is in itself a vivid and complex area of research in statistics and as such extends the scope of this paper.

C. Measurement Noise

In t-test based evaluation the noise in the measurement system can play a significant role in detection capability. As $t \underset{\sim}{\propto} \sqrt{n}/s$, the required number of traces increases r^2-fold if the Signal-to-Noise Ratio (SNR) is lowered by a factor of r. Therefore, when using a t-test, a noisy measurement system may convince an evaluator of the security of an insecure implementation. This can not happen with our framework. While a noise measurement may fail to provide non-zero lower bounds for the leakage, the upper bound will increase accordingly. Consequently, in order to ensure a (pre-defined) limit is not exceeded, more measurements are required.

D. Signal to Noise Ratio

An evaluator might not be interested in the absolute difference of means of two random variables. This could be for two reasons: The result is scale dependent and it does not account for the difficulty for an attacker to measure and exploit this difference which is hidden in noise. A well-known solution for both problems is the introduction of an SNR. Given large sample sizes, small second order leakage ($\sigma_X \approx \sigma_Y$), and sufficient noise $\left(\sqrt{\sigma_X} \gg |\mu_X - \mu_Y|\right)$ the estimated variance s_X^2 from Sect. II-A corresponds to the noise an arbitrary attacker will face trying to distinguish the means. Therefore, a meaningful SNR for the confidence interval can be constructed as $I_{SNR} = \left[\frac{\Delta_{min}}{\sqrt{s_X^2}}; \frac{\Delta_{max}}{\sqrt{s_X^2}}\right]$. However, note that when using SNR, independence of measurement noise is lost. Therefore, SNR-intervals are only valid for a given acquisition system.

E. Higher-Order Moments

Our confidence interval based approach for leakage assessment can be extended to provide limits for higher-order leakage using methods analogous to the ones described in previous TVLA-schemes [7], [8]. There, the authors reduce the task of detecting higher-order leakage to the problem of detecting first-order leakage in pre-processed traces. First, the acquired traces are pre-processed, combining samples with an appropriate function. The result is then evaluated with the same metric as in the first-order case. For example, second-order univariate leakage is analyzed using mean-free, squared traces, yielding bounds for the difference of the traces' variance. In general, while the first-order confidence interval yields bounds for the difference of means of the signals, the higher order test can provide bounds for the difference of the higher-order moments. For statistical moments μ_d for $d > 2$, the authors of [8] construct a t-test based on standardized moments. We deviate from this approach by using centralized moments instead, allowing the analysis of the SNR if required. For $d > 1$, the means μ_d used to calculate the dth-order confidence interval are given as the dth central moments of the signals:

$$\mu_d = \frac{1}{n} \sum (x - \mu)^d. \tag{4}$$

The variance used to estimate the dth-order confidence interval can then be computed based on the central moments:

$$\begin{aligned} s_d^2 &= \frac{1}{n} \sum \left((x - \mu)^d - \mu_d\right)^2 \\ &= \frac{1}{n} \sum \left((x - \mu)^{2d} - 2(x - \mu)^d \mu_d + \mu_d^2\right) \\ &= \mu_{2d} - \mu_d. \end{aligned} \tag{5}$$

The confidence intervals can then be constructed according to Sect. II-A.

If the evaluator has access to the masks, the traces can be pre-processed by averaging before calculating the higher-order moments, in order to reduce the measurement noise as suggested in [10]. This will in general result in reduced measurement complexity. However, for SNR calculation, the variance over all randomly chosen masks should be used as this is the noise an attacker has to face.

F. Efficient Implementation

In order to compute confidence intervals for moments up to order d an evaluator needs to estimate the mean values and the central moments μ_i for $1 < i < 2d$ of the signals. This can be achieved by fast iterative methods described in [8] or using histograms as demonstrated in [13]. The actual source of the traces is independent of our framework: it is applicable to non-specific fixed-versus-random or fixed-vs-fixed measurements, as well as analysis regarding specific intermediate values. The confidence interval can then be computed in regular intervals, e.g., after each 1000 measurements. As the calculations are independent for each sample point, they can easily be parallelized.

III. Case Studies

In this section we study the effectiveness of our evaluation framework by applying it to a first-order secure implementation of the AES, protected by a domain-oriented masking scheme. We practically show the capability of our analysis framework and compare its results to a t-test based evaluation.

We acquired current consumption traces of the design for a non-specific leakage evaluation according to [8]. To this end, we fixed a key and randomly sent fixed or random masked plaintexts to the device under test, recording the current consumption traces, thus sampling X_j and Y_j for multiple time points j. We then analyze the absolute difference in the statistical moments of the power consumptions using our proposed framework and compare the results to a Welch's t-test. More precisely, we use our framework to answer the question: Which band contains the absolute difference between the means of the moments of X and Y for all sample points. If not stated otherwise, a confidence level of $1 - \alpha = 0.99$ is assumed throughout this section.

A. Measurement Setup

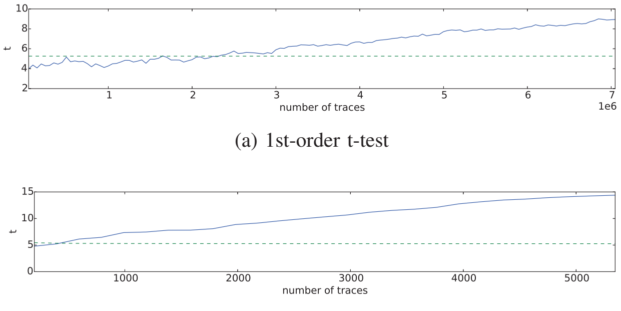
The algorithm was synthesized for a Xilinx Spartan6 FPGA on a SAKURA-G side-channel evaluation board [14]. The design was supplied with a $4\,\mathrm{MHz}$ clock. The current consumption of the target device was amplified by approximately $20\,\mathrm{dB}$ and then digitized at a sample rate of $1.25\,\mathrm{GS/s}$ using an $8\,\mathrm{bit}$ digital oscilloscope. Note that y-axis labels of the figures in Sect. III-C reflect the voltage after the amplifier. As an encryption on the target device took $55.25\,\mu\mathrm{s}$, the resulting traces consist of 69062 samples each. Therefore, the corrected confidence level is $1 - \alpha_t = 0.99^{\frac{1}{69062}} = 1 - 1.46 * 10^{-7}$ as by Sect. II-B.

B. Target Implementation

Our evaluation target was an 128-bit AES-core protected with a domain-oriented masking scheme presented in [15] and available online at [16]. The implementation allows arbitrary protection order by setting a VHDL-generic accordingly and uses $d + 1$ shares in order to achieve d-order security. For our evaluation, we chose the first-order secure variant of the design. The design was synthesized in the interleaved variant with 5-stage S-boxes. The bitstream was generated using Xilinx ISE 14.7 with the options ”keep hierarchy” on and ”register duplication” off, in order to prevent interference with the masking countermeasure.

C. Results

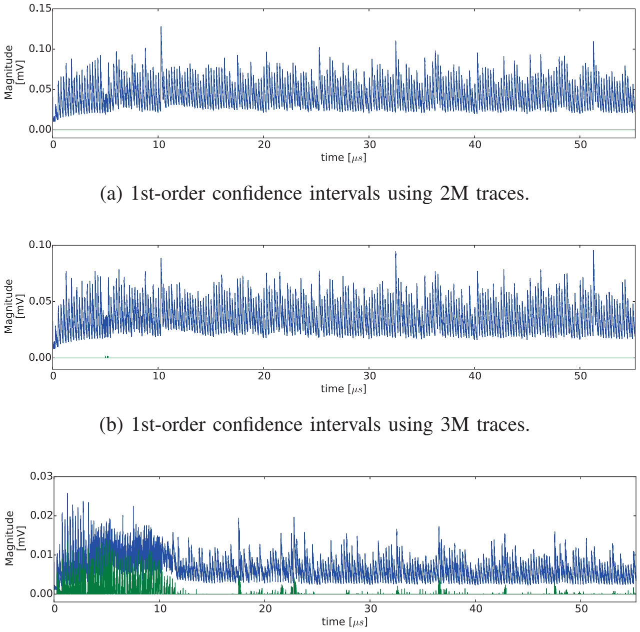
Figure 1 depicts the result of a t-test trying to prove the hypothesis H_1: $\exists j : \mu_{j,X} \neq \mu_{j,Y}$ which states that at some sample point(s) there is a difference in the mean power consumption. The results indicate that there is some first and second order leakage, as the maximum t-value exceeds the corresponding threshold after a sufficient number of measurements. However, the t-test can not provide a confident assurance of the magnitude of leakage. Even worse, if we only captured two million traces the t-test would not allow any confident statement regarding the first order leakage, as the maximum t-value would not have crossed the threshold. There could be two reasons for failing to detect leakage: either there is no leakage, or the number of measurements is too low to discover it. In contrast, Fig. 2a shows the confidence intervals, as constructed by our method, using two million traces. While the interval does not prove the presence of leakage (the lower limit is zero), it limits (with confidence 0.99) the maximum possible leakage at each point by the corresponding upper interval limit.



(a) 1st-order t-test

(b) 2nd-order t-test

Fig. 1. Maximum of absolute t-statistics and detection threshold for moments 1 and 2.



(a) 1st-order confidence intervals using 2M traces.

(b) 1st-order confidence intervals using 3M traces.

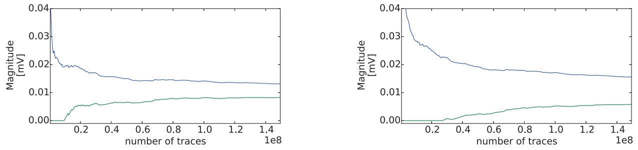
(c) 1st-order confidence intervals using 150M traces.

Fig. 2. Confidence intervals depicting lower (green) and upper (blue) bounds for the absolute difference in means. Please note different scales.

If more traces are available, the interval limits become tighter, allowing a closer estimation of the leakage. Figure 2b shows the intervals computed using three million traces. As the threshold of 5.26 (corresponding to $\alpha_t = 1.46 * 10^{-7}$)

is exceeded in the t-test, by definition of the interval, the lower limit becomes non-zero for some points, while the upper limits simultaneously converge against the true leakage. If even closer estimates are required, more traces can be recorded. Figure 2c shows the tight leakage estimation using 150 million traces.

Specific sample points can also be analyzed independently. Figure 3a depicts the development of the interval at sample point 8715, corresponding to a peak in leakage at 6.97 µs, over the number of traces. If statements about (single) arbitrarily chosen points are made, as in the multi-point case, it is important to use the FWER-corrected confidence level as described in Sect. II-B. If the point was chosen at random or using prior knowledge (such as prior measurements) and only a single point is considered, the unadjusted confidence level can be used.



(a) With measured SNR.

(b) With 50% of measured SNR.

Fig. 3. First-order confidence intervals at sample point 8715 over the number of measured traces.

Figure 3b depicts the influence of noise in the evaluation system. As expected, more traces are needed in order to achieve similar bounds.

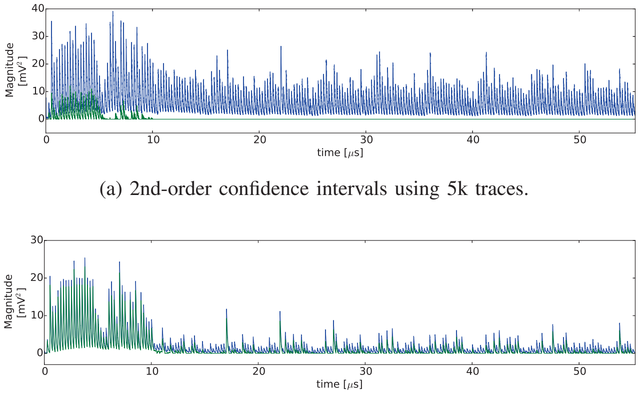
These figures illustrate a main advantage of our confidence interval approach: if a leakage level is defined as separating a secure and an insecure implementation, an evaluator can take measurements until either the maximum upper limit or the maximum lower limit of the confidence intervals crosses that level. Then, with a specified confidence, the implementation can be clearly classified, independent of the evaluators measurement system. However, as with all TVLA methods the evaluation results are only valid for the selected inputs. We therefore recommend careful plaintext selection and repetition of the evaluation with different inputs, if possible.

As discussed in Sect. II-E, the framework can also be used to assess higher-order leakage. Figure 2 shows the obtained intervals for second order. As the difference in the second moment is much higher than in the first, considerably less traces are required to tightly bound it.

It is important to note that our results do not necessarily support the conclusion that the countermeasures in the analyzed implementations fail to provide protection against side-channel attacks. We can merely state that the *concrete instantiation* using the bitstream created by us and programmed into our target device exhibits the leakage described above.

IV. Conclusion

We describe a new confidence interval-based framework for TVLA and demonstrate its advantages in comparison to established methods. We suggest our new framework to be applied as a drop-in replacement for t-test TVLA-methods currently in use in the industry and the scientific community to provide a clearer picture of the side-channel resistance of protected cryptographic implementations. To this end, instead of calculating a maximum t-value that is reached after a certain amount of measurements, maximum and minimum bounds for the leakage should be measured. In future work, an extension of our framework to multivariate leakage assessment might be worthwhile.



(a) 2nd-order confidence intervals using 5k traces.

(b) 2nd-order confidence intervals using 500k traces.

Fig. 4. Confidence intervals for the absolute difference in variance (2nd-order analysis).

References

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

[2] J. Quisquater and D. Samyde, “Electromagnetic analysis (EMA): measures and counter-measures for smart cards,” in *E-smart*, vol. 2140 of *Lecture Notes in Computer Science*, pp. 200–210, Springer, 2001.

[3] N. Veyrat-Charvillon, M. Medwed, S. Kerckhof, and F. Standaert, “Shuffling against side-channel attacks: A comprehensive study with cautionary note,” in *ASIACRYPT*, vol. 7658 of *Lecture Notes in Computer Science*, pp. 740–757, Springer, 2012.

[4] E. Prouff and M. Rivain, “Masking against side-channel attacks: A formal security proof,” in *EUROCRYPT*, vol. 7881 of *Lecture Notes in Computer Science*, pp. 142–159, Springer, 2013.

[5] B. Bilgin, S. Nikova, V. Nikov, V. Rijmen, N. N. Tokareva, and V. Vitkup, “Threshold implementations of small s-boxes,” *Cryptography and Communications*, vol. 7, no. 1, pp. 3–33, 2015.

[6] O. Reparaz, B. Bilgin, S. Nikova, B. Gierlichs, and I. Verbauwhede, “Consolidating masking schemes,” in *CRYPTO (1)*, vol. 9215 of *Lecture Notes in Computer Science*, pp. 764–783, Springer, 2015.

[7] G. Goodwill, B. Jun, J. Jaffe, and P. Rohatgi, “A testing methodology for side channel resistance validation,” in *NIST non-invasive attack testing workshop*, 2011.

[8] T. Schneider and A. Moradi, “Leakage assessment methodology - A clear roadmap for side-channel evaluations,” 2015.

[9] F. Standaert, “How (not) to use welch’s t-test in side-channel security evaluations,” *IACR Cryptology ePrint Archive*, vol. 2017, p. 138, 2017.

[10] L. Zhang, A. A. Ding, F. Durvaux, F. Standaert, and Y. Fei, “Towards sound and optimal leakage detection procedure,” *IACR Cryptology ePrint Archive*, vol. 2017, p. 287, 2017.

[11] B. Rüger, *Test- und Schaätztheorie Vol. 2: Statistische Tests*. Lehr- und Handbucher der Statistik, Munchen [u.a.]: Oldenbourg, 2002. XII, 570 S. : graph. Darst.

[12] Z. Sidak, “Rectangular confidence regions for the means of multivariate normal distributions,” *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–633, 1967.

[13] O. Reparaz, B. Gierlichs, and I. Verbauwhede, “Fast leakage assessment,” *IACR Cryptology ePrint Archive*, vol. 2017, p. 624, 2017.

[14] “Side-channel AttacK User Reference Architecture.” http://satoh.cs.uec.ac.jp/SAKURA/index.html.

[15] H. Groß, S. Mangard, and T. Korak, “Domain-oriented masking: Compact masked hardware implementations with arbitrary protection order,” *IACR Cryptology ePrint Archive*, vol. 2016, p. 486, 2016.

[16] H. Groß, “DOM protected hardware implementation of AES.” Available at https://github.com/hgrosz/aes-dom as of November 30, 2017, 2016.