Towards Optimal Pre-processing in Leakage Detection

Changhai Ou, Degang Sun, Zhu Wang and Xinping Zhou

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
ouchanghai@iie.ac.cn

Abstract. An attacker or evaluator can detect more information leakages if he improves the Signal-to-Noise Ratio (SNR) of power traces in his tests. For this purpose, pre-processings such as de-noise, distribution-based traces biasing are used. However, the existing traces biasing schemes can't accurately express the characteristics of power traces with high SNR, making them not ideal for leakage detections. Moreover, if the SNR of power traces is very low, it is very difficult to use the existing de-noise schemes and traces biasing schemes to enhance leakage detection. In this paper, a known key based pre-processing tool named Traces Linear Optimal Biasing (TLOB) is proposed, which performs very well even on power traces with very low SNR. It can accurately evaluate the noise of time samples and give reliable traces optimal biasing. Experimental results show that TLOB significantly reduces number of traces used for detection; correlation coefficients in ρ-tests using TLOB approach 1.00, thus the confidence of tests is significantly improved. As far as we know, there is no pre-processing tool more efficient than TLOB. TLOB is very simple, and only brings very limited time and memory consumption. We strongly recommend to use it to pre-process traces in side channel evaluations.

Keywords: traces optimal biasing · TOB · TLOB · leakage detection · biasing power traces · SNR · CPA · side channel attack

Introduction

Secret information may leak from devices through side channels such as electromagnetic [2], acoustic [12] and power consumption [18] during the implementation of cryptographic algorithms. These leakages are usually unconscious and difficult to be discovered. By taking advantage of statistical correlation between assumed power consumption of intermediate values and side channel leakages, an attacker can recover sensitive information (e.g. encryption key) in the target devices. Side channel attacks, such as Differential Power Analysis (DPA) [18], Correlation Power Analysis (CPA)[5], Template Attacks (TA) [6] and Collision Attacks (CA) [29, 27], pose serious threats to the security of cryptographic implementation. In order to improve the attack efficiency, an attacker always tries to make full use of leakage informations, constructs optimal distinguishers and leakage models. Pre-processings can be also used to enhance attacks. Current countermeasures, such as flawed masking implementations [28, 9, 8], failed to defend against the corresponding higher-order attacks as it claimed. In order to improve security, a defender always tries to reduce or eliminate the leakage of implementations, for whom side channel leakage detection and evaluation are very meaningful.

Leakage detections, such as Welch's t-test [7, 13] in CRI's TVLA proposal and extensions in [10, 26], Normalized Inter-Class Variance(NICV) in [4], Mutual Information Analysis (MIA) in [21], which relate to the concrete security level of an implementation,

are very important tools for side channel evaluations. They evaluate the security level according to whether the device leaks information or how many side channel measurements are required to detect the leakage. They can also simply detect whether leakage exists, independent of whether the leakage can be exploited. Correlation-based leakage detection ρ-test was proposed in [11], which could significantly enhance the most widely used Welch's t-test, approximately 5 times less measurements were used as stated by the authors. However, the above mentioned works only considered the univariate test on the selected single POI only. In fact, as stated by Zhang et al. in [33], the overall significance level increased as the total number of time samples on the traces increased. For long traces, the overall significance level could be quite large so that non-leaky devices could not pass TVLA t-test with critical value of 4.5. They optimally combined the detection tests of univariate leakage at all time samples along the power traces, thus improving the TVLA procedure.

Here we need to note that power traces pre-processing is independent of specific leakage detection. It can be used to enhance not only side channel attacks but also the above leakage detections. Since the presence of noise, if leakage detection tests are directly performed on power traces without pre-processing, the results may be not ideal. Take ρ-test for example, correlation coefficient ρ obtained by the evaluator is usually not very large. In other words, the noise level determines ρ. The number of power traces used in tests determines the stability of ρ. If the SNR is low, the detection is time-consuming and unreliable, a lot of power traces are needed. So, pre-processing schemes such as power traces alignment [32], averaging [30], de-noise [14], fourth-order cumulant [19], are usually the first step of leakage detection tests.

Biasing power traces, one of the common pre-processing tools to improve SNR, was firstly proposed by Kris et al [31]. The power consumption of the Point-of-Interest (POI) most relevant to S-box outputs could be approximated by a normal distribution [17]. They obtained the distribution through calculating the Probability Density Function (PDF) values of power consumption of this POI. The smaller the PDF value, the lower the noise of the power trace. The PDF values are sorted and power traces with high SNR are biased from two tails of the distribution. Noura et al. used a new power model to bias power traces to improve CPA in [23]. However, they did not make improvements to the strategy of biasing power traces proposed in [17]. Hu et al. proposed an Adaptive Chosen Plaintext Correlation Power Analysis (ACPCPA) [15]. They solved the problem of discarding too many power traces in the scheme proposed by Yongade et al [17]. They analyzed the correlation between S-box output Hamming weights and power consumption of POIs, and got a conclusion that Hamming weights 0, 1, 7, and 8 corresponded to power traces with high SNR. They acquired the corresponding measurements to perform CPA. This scheme was improved by Ou et al. in [24].

However, as we detailed in Section 1.3, the above traces biasing schemes improve SNR through enlarging the variance of exploitable power consumption component (see Section 4.1 in [20]) instead of de-noise. The SNR improved by these schemes is very limited. Recently, other schemes such as Principal Component Analysis (PCA), were used to bias power traces [16]. The above schemes can be able to accurately bias power traces with high SNR if the noise level is low. However, if the noise level is high, the power consumption and power leakage model (e.g. Hamming weight model) of intermediate values are no longer linear, accuracy of biasing power traces is greatly reduced. This also indicates that directly using the distribution of power consumption to bias power traces with high SNR is not ideal. How to accurately bias traces to improve the efficiency and confidence of detection is still a challenging problem.

The attacker or evaluator sorts the power traces according to PDF values and biases the 'optimal' ones to perform attacks. When a fixed number of power traces are used, the SNR of the biased ones is usually the highest. That means, if the attacker or evaluator

uses these biased traces to perform CPA, he gets the optimal correlation coefficients. This characteristic is very important, but none of the above papers have found it. Benefit from this, a known key based pre-processing tool named Traces Linear Optimal Biasing (TLOB) is proposed to bias power traces with high SNR in this paper. TLOB can accurately evaluate the noise level of each power trace according to the outputs of side channel distinguishers, and give the reliable traces optimal sorting. The number of power traces used in leakage detection tests using our TLOB can be significantly reduced to dozens or hundreds, the correlation coefficient of ρ test can also be significantly improved (approaching 1.00). Thus, the time-consuming tests can be performed very fast and the confidence of tests is significantly enhanced. What is most mysterious, TLOB works very well at different noise levels, especially when biasing a small number of traces from a large and noisy power traces set. However, if the SNR of power traces is very low, it is very difficult to use the existing de-noise schemes and traces biasing schemes to enhance leakage detection. So, we strongly recommend to use TLOB in side channel evaluations.

The rest of this paper is organized as follows. Leakage characteristics of POIs, cross validation and distribution-based traces biasing proposed by Yongdae et al. are introduced in Section 1. Leakage detection t-test and ρ-test are introduced in Section 2. Our TLOB is detailed in Section 3. In order to provide good references for evaluator to decide the threshold of the number of biased traces, ρ tests are performed on simulated traces pre-processed by TLOB under different SNRs and different numbers of power traces in Section 4. Then, we perform real experiments on the measurements of our AT89S52 microcontroller and DPA *contest* $v1$ in Section 5. Finally, Section 6 draws general conclusions.

1 Backgrounds

1.1 Leakage Characteristics of POIs

Let us denote the encrypted plaintext as $x = x_0||x_1||\cdots||x_{15}$, the key used in cryptographic device as $\kappa = \kappa_0||\kappa_1||\cdots||\kappa_{15}$, the execution of the S-box Sbox as $z_i = \mathsf{Sbox}(x_i \oplus \kappa_i)$, the corresponding leakage at time τ as $l(\tau)$. Here x_i denotes the i-th plaintext byte, κ_i denotes the i-th key byte, z_i denotes the corresponding intermediate value. The evaluator encrypts a set of plaintexts $\mathcal{P}$ and acquires a set of traces, which is denoted as $\mathcal{L}$.

According to [20], the power consumption of a single time sample $l(\tau)$ can be modeled as the sum of an operation dependent component $l_o(\tau)$, a data-dependent component $l_d(\tau)$, electronic noise $l_{el.n}(\tau)$, switching noise $l_{sw.n}(\tau)$, and the constant component $l_c(\tau)$. That is,

$$l(\tau) = l_o(\tau) + l_d(\tau) + l_{el.n}(\tau) + l_{sw.n}(\tau) + l_c(\tau). \tag{1}$$

These 5 components are independent of each other. For a time sample, the variance $\hat{\mathsf{var}}(\mathcal{L}_o(\tau)) = \hat{\mathsf{var}}(\mathcal{L}_c(\tau)) = 0$. For a classical DPA attack, the attackers only consider one of the 8 bits intermediate values (e.g. the outputs of Sbox), the power consumption of other 7 bits is switching noise (i.e. algorithm noise). The variance of switching noise here is larger than 0. For a classical CPA attack considering all bits of intermediate values, the variance of switching noise $\hat{\mathsf{var}}(\mathcal{L}_{sw.n}(\tau)) = 0$. The electronic noise is normal distributed with mean 0 and variance σ^2. Let $l_n(\tau)$ denote the noise component including $l_{el.n}(\tau)$ and $l_{sw.n}(\tau)$.

For all components of power consumption, $l_d(\tau)$ is the only component correlating to the leakage model (e.g. Hamming weight model). Let $\hat{\mathsf{model}}(\tau)$ denote the profiled mean power consumption model of the intermediate values, the correlation coefficient between $\hat{\mathsf{model}}(\tau)$ and the total power consumption $\mathcal{L}(\tau)$ is

$$\hat{\rho}(\hat{\mathsf{model}}(\tau), \mathcal{L}(\tau)) = \frac{\mathsf{cov}(\hat{\mathsf{model}}(\tau), \mathcal{L}(\tau))}{\sqrt{\mathsf{var}(\hat{\mathsf{model}}(\tau))\,\mathsf{var}(\mathcal{L}(\tau))}}. \tag{2}$$

cov here is the covariance matrix operator. Mangard et al. further analyzed the correlation between power consumption and $\hat{\mathsf{model}}(\tau)$ in [20]. The important formula in their paper can be expressed as:

$$\hat{\rho}\left(\hat{\mathsf{model}}(\tau), \mathcal{L}(\tau)\right) = \frac{\hat{\rho}\left(\hat{\mathsf{model}}(\tau), \mathcal{L}_d(\tau)\right)}{\sqrt{1 + \frac{1}{\hat{\mathsf{SNR}}(\tau)}}}, \tag{3}$$

which lays the theoretical foundation for biasing power traces in CPA. For a classical CPA attack, $\hat{\rho}\left(\hat{\mathsf{model}}(\tau), \mathcal{L}_d(\tau)\right)$ is a constant for a time sample, SNR determines the correlation coefficient $\hat{\rho}\left(\hat{\mathsf{model}}(\tau), \mathcal{L}(\tau)\right)$. To improve it, the attacker or evaluator should improve SNR. The SNR of a time sample is the variance ratio of exploitable power consumption components $(\mathcal{L}_o(\tau) + \mathcal{L}_d(\tau))$ and the noise components. It can be modeled as

$$\hat{\mathsf{SNR}}(\tau) = \frac{\hat{\mathsf{var}}\left(\mathcal{L}_o(\tau)\right) + \hat{\mathsf{var}}\left(\mathcal{L}_d(\tau)\right)}{\hat{\mathsf{var}}\left(\mathcal{L}_n(\tau)\right)}. \tag{4}$$

For a time sample, $\hat{\mathsf{SNR}}(\tau) = \frac{\hat{\mathsf{var}}(\mathcal{L}_d(\tau))}{\hat{\mathsf{var}}(\mathcal{L}_n(\tau))}$. The formula indicates that there are two ways to improve SNR, one is noise reduction, and the other is to enlarge $\hat{\mathsf{var}}\left(\mathcal{L}_d(\tau)\right)$.

1.2 Cross Validation

Cross-validation is a kind of statistical analysis used to verify the performance of classifier, such as distinguishers in side channel attacks. For a k-fold cross-validation, the evaluator splits the acquired traces $\mathcal{L}$ into k non-overlapping sets $\mathcal{L}^{(i)}$ $(1 \leq i \leq k)$ of approximately the same size. Then, profiling sets $\mathcal{L}_p^{(j)} = \bigcup_{i \neq j} \mathcal{L}^{(i)}$ and test sets $\mathcal{L}_t^{(j)} = \mathcal{L} \setminus \mathcal{L}_p^{(i)}$ are defined. The profiling sets are used to profile the leakage model, such as the $\hat{\mathsf{model}}$ we used in Equation 2. The test sets are used to test and evaluate the performance of the trained model.

In order to obtain a reasonable model, the $\hat{\mathsf{model}}$ is calculated k times and then averaged. As a result, the number of power traces used increases and the computational cost is higher. The evaluator obtains a more accurate power consumption model, which promotes the evaluation. Moreover, for correlation ρ tests, the mean of ρ-s obtained in 10-fold cross validation is more stable, which makes the detection results more reliable and confidence.

1.3 Distribution-Based Traces Optimal Biasing

The power consumption of a time sample can be modeled by a normal distribution. The noise component $l_n(\tau)$ is uncontrollable. Preprocessing can be used to de-noise. In addition, the evaluator or attacker can improve the SNR by enlarging $\hat{\mathsf{var}}\left(l_d(\tau)\right)$, thereby improving the correlation coefficient $\hat{\rho}\left(\hat{\mathsf{model}}(\tau), \mathcal{L}(\tau)\right)$. Suppose that the leakage of device follows Hamming weight model, then $\hat{\mathsf{var}}\left(l_d(\tau)\right) \propto \hat{\mathsf{var}}\left(\mathsf{HW}\left(Z(\tau)\right)\right)$, where HW is the Hamming weight function. For an unknown key implementation, the attacker or evaluator sorts the PDF values of time samples and biases power traces at both tails of the normal distribution [17, 24]. We name this kind of traces biasing schemes as Distribution-Based Traces Optimal Biasing (DTOB).

The principle of the adaptive chosen plaintext correlation power analysis proposed by Hu et al [15] is similar to DTOB. However, it doesn't sort the noise of power traces. This is not surprising, since the current published papers have not found the advantages of monotonically decreasing correlation coefficient ρ in TOB. They only consider biasing power traces as a way to improve CPA. Specifically, as detailed in [15], Hamming weights 0, 1, 7, and 8 corresponded to power traces with high SNR, which were biased to enhance

CPA. In this way, the $\hat{\mathsf{var}}(\mathsf{HW}(Z))$ improves from 2.0078 to 10.3529. However, these schemes improve SNR by enlarging $\hat{\mathsf{var}}(\mathsf{HW}(Z))$ instead of de-noise.

Let R denote the Gaussian distributed noise component $\mathcal{L}_n(\tau)$ with mean 0 and variance σ^2, C denote the sum of constant component $\mathcal{L}_c(\tau)$ and operation dependent component $\mathcal{L}_o(\tau)$. The leakage model of the S-box output implementation can be modeled as

$$\mathcal{L} = \mathsf{HW}(Z) + C + R. \tag{5}$$

We simulate 25600 power traces and bias 12800 out of them. Here C in Equation 5 is set to 100, σ^2 is set to 0.25. The biased power traces at both tails of the overall normal distribution are shown in Fig.1.

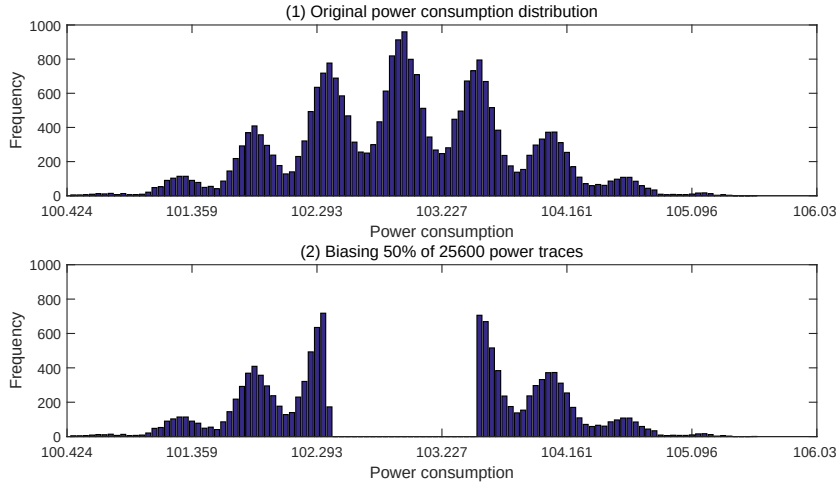

Figure 1: Using DTOB to bias 50% of 25600 power traces.

As shown in Fig.1, 50% of 25600 power traces are biased from two tails of the distribution. The smaller the PDF values, the higher the SNR of power traces. As we mentioned before, this scheme doesn't reduce the variance of noise $\hat{\mathsf{var}}(\mathcal{L}_n(\tau))$, but enlarges $\hat{\mathsf{var}}(\mathcal{L}_d(\tau))$ instead. If the noise level is low, it can bias power traces with high SNR accurately. However, if the noise level is high, noisy time samples corresponding to Hamming weights close to 4 also appear at both two tails of the distribution. Similarly, time samples corresponding to small or large Hamming weights may also appear in the middle of the distribution. Therefore, it is not ideal. In this case, the evaluator can get better results if he directly uses ACPCPA in [15] to detect leakage. However, ACPCPA is also not ideal if the noise on traces is very large.

2 Leakage Detection Tests

2.1 Student's t-test

Student's t-test is the most popular leakage detection test. It considers one bit of intermediate value, which is denoted as X here. The evaluator or attacker collects n_0 and n_1 power traces corresponding to $X = 0$ and $X = 1$ respectively, and stores them in vectors $\mathcal{T}_0$ and $\mathcal{T}_1$. Then, Student's t-test is computed as follows:

$$\Delta = \frac{\hat{\mathsf{E}}(\mathcal{T}_0(\tau)) - \hat{\mathsf{E}}(\mathcal{T}_1(\tau))}{\sqrt{\frac{\hat{\mathsf{var}}(\mathcal{T}_0(\tau))}{n_0} + \frac{\hat{\mathsf{var}}(\mathcal{T}_1(\tau))}{n_1}}}, \tag{6}$$

where $\hat{\mathsf{E}}$ denotes the sample mean operator and $\hat{\mathsf{var}}$ denotes the sample variance operator. The probability of null hypothesis that $\Delta = 0$ can be computed as follows:

$$p = 2 \times (1 - \mathsf{CDF}_{\tau}(|\Delta|, \nu)), \tag{7}$$

where CDF is the cumulative function of a Student's t distribution, and ν is its number of freedom degrees. If n_0 and n_1 are large enough, Student's t distribution is close to normal distribution $\mathcal{N}(0, 1)$.

2.2 Correlation ρ-test

A correlation-based leakage detection ρ-test was proposed by Durvaux and Standaert [11]. Unlike Student's t-test, leakage models such as Hamming weight model[3], Hamming distance model [5], switch distance model [25], are needed when profiling the leakage of the devices in ρ-test. According to [11], a k-fold cross-validation is used. The evaluator splits the full set of traces $\mathcal{L}$ into k non-overlapping sets $\mathcal{L}^{(i)}$ $(1 \leq j \leq k)$ of approximately the same size, and gets profiling sets $\mathcal{L}_p^{(j)} = \bigcup_{i \neq j} \mathcal{L}^{(i)}$ and test sets $\mathcal{L}_t^{(j)} = \mathcal{L} \setminus \mathcal{L}_p^{(i)}$. For each cross-validation set $\mathcal{L}_t^{(j)}$ with $1 \leq j \leq k$,

$$\hat{r}^{(j)}(\tau) = \hat{\rho}(\mathcal{L}_t^{(j)}(\tau), \hat{\mathsf{model}}^{(j)}(\tau)), \tag{8}$$

where $\hat{\mathsf{model}}$ denotes the profiled mean power consumption model, and τ denotes the time sample on power traces, $\hat{r}^{(j)}$ is the corresponding estimated correlation coefficient. Then, Fisher's z-transformation is applied and the evaluator obtains:

$$\hat{r}_z(\tau) = \frac{1}{2} \times \mathsf{ln}\left(\frac{1 + \hat{r}(\tau)}{1 - \hat{r}(\tau)}\right), \tag{9}$$

where ln is the natural logarithm function. Let CDF denote the Student's t cumulative distribution function. If $\hat{r}_z(\tau)$ is normalized with standard deviation $\frac{1}{\sqrt{N-3}}$, where N is the size of time samples in the set $\mathcal{L}$. Then, the probability for a null hypothesis assuming no correlation:

$$p = 2 \times (1 - \mathsf{CDF}_{\mathcal{N}(0,1)}(|\hat{r}_z(\tau)|)). \tag{10}$$

We still use $\hat{\rho}$ to denote the averaged correlation coefficient in the next sections. Durvaux et al. stated in [11] that correlation ρ-test could significantly improve Welch's t-test with significantly faster detection speed (with approximately 5 times less measurements in their experiments). Just as different distinguishers have different distinguishing ability, different leakage detection tests have different capabilities to detect leakages. The Student's t-test is the most widely used leakage detection test. However, we think ρ-test is much better than it. In addition to its high efficiency, the correlation coefficient output by ρ-test directly reflects the linearity between the power consumption and the profiled leakage model.

3 Traces Linear Optimal Biasing

The first step of side channel attacks is usually power traces pre-processing, of which one of the main goals is noise reduction. Biasing power traces can help the attacker or evaluator achieve better noise reduction purposes. Yongdae et al. proposed DTOB in [17], in which the idea of traces sorting was given for the first time. We name all traces biasing schemes using sorting as Traces Optimal Biasing (TOB). As we mentioned in Section 1, DTOB can obtain the highest correlation coefficient when biasing a fixed number of power traces and the noise level is low. However, this advantage has not been found by Yongdae et al. and further researched in other current published papers. If the SNR is low, DTOB is not ideal. In this section, we will give an optimal strategy to bias traces, which performs well under different noise levels, especially very large ones.

3.1 New Definition of Noise Level

There are a lot of noise reduction methods used for pre-processing such as power traces averaging [22, 20, 30] and PCA [16]. Taking power traces averaging for an example, we assume that the leakage corresponding to the first plaintext byte having value 15 (the corresponding Hamming weight is 4) in the first round of AES follows normal distribution $\mathcal{N}(\mu, \sigma^2)$. If the evaluator captures N power traces and averages them to get a new trace. The new power consumption follows new distribution $\mathcal{N}(\mu, \sigma^2/N)$ (see Section 4.6 in [20]). Taking $\mu = 104$, $\sigma^2 = 1$ and $N = 10$ for an example, the two normal distributions are shown in Fig.2. $\mathcal{N}(104, 1/10)$ is higher and thinner than $\mathcal{N}(104, 1)$. Traces averaging does not change the μ, but reduces the noise variance. Compared to $\mathcal{N}(104, 1)$, most of time samples in distribution $\mathcal{N}(104, 1/10)$ concentrate close to μ. The correlation coefficient $\hat{\rho}$ improves if correlation test is performed on the pre-processed traces. The larger the $\hat{\rho}$-s, the better the linearity of the pre-processed traces and the $\mathsf{m\hat{o}del}$. In other words, the closer the time samples to the $\mathsf{m\hat{o}del}$, the better the linearity of them.

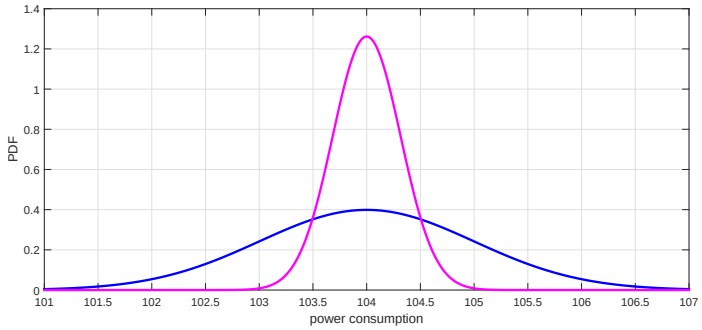

Figure 2: Probability density function of $\mathcal{N}(104, 1)$ (blue) and $\mathcal{N}(104, 0.10)$ (magenta).

For evaluators, leakage detection is to determine whether a device leaks information and evaluate the corresponding security level. Compared to attackers, they have more advantages, such as the ability to accurately capture the measurements, and the knowledge of the key in the cryptographic device. They can be able to obtain a power consumption distribution for each plaintext. The power traces with small noise component are the ones close to the mean power consumption. The linearity of these traces and mean power consumption leakage model $\mathsf{m\hat{o}del}$ is good, and the SNR of these traces is high. On the contrary, the farther the power consumption of a trace from the profiled model, the larger the noise component is. This definition of 'traces with high SNR' is more thorough and accurate than the one of Yongdae et al [17]. In order to achieve the purpose of noise reduction, the evaluator only needs to use a pre-processing tool to make most time samples distribute near their corresponding mean power consumption model $\mathsf{m\hat{o}del}$.

3.2 Linear Measurement

Let $l_i(\tau)$ denote the τ-th time sample on the i-th power trace in set $\mathcal{T}$, $\mathsf{m\hat{o}del}(\tau)$ denote the mean power consumption leakage model. For each trace, the evaluator gets:

$$\hat{d}_i^{(j)}(\tau) = \hat{\mathsf{D}}\left(\mathsf{m\hat{o}del}^{(j)}(\tau), l_i^{(j)}(\tau), \mathcal{L}_t^{(j)}(\tau)\right). \tag{11}$$

Here D is an evaluation tool such as a side channel distinguisher, or a linear metric that outputs linear evaluation values according to the power consumption. For example, the correlation coefficients output by CPA. The higher the correlation, the better the linearity.

In this case, we need to make full use of power trace set $\mathcal{L}_t^{(j)}(\tau)$. Actually, the correlation coefficients $\hat{\rho}$-s in areas without leakage are very high and undistinguishable with the ones in leakage areas. Moreover, it is very time-consuming if CPA is performed for each trace. This shortcoming is very outstanding when the power trace set is very large. Finally, the distinguisher used here can be regarded as a simplified Template Attack (TA). Each trace $l_i(\tau)$ is compared with its corresponding mean power consumption leakage model. The traces closest to the model are biased. We named this pre-processing scheme as Traces Linear Optimal Biasing (TLOB).

3.3 Traces Optimal Sorting and Biasing

For TLOB, the evaluator sorts the power traces according to $\hat{d}^{(j)}$ using an evaluation function S. Here we define it as a distance sorting function. The distance vector $\hat{d}^{(j)}$ is sorted in ascending order for each time sample $\mathcal{L}_t^{(j)}(\tau)$,

$$\mathcal{L}^{',(j)}(\tau) = \mathsf{S}\left(\mathcal{L}_t^{(j)}(\tau), \hat{d}^{(j)}(\tau)\right). \tag{12}$$

Then, n_t traces with smallest distances are biased, and the new leakage detection ρ-test is performed:

$$\hat{r}^{(j)}(\tau) = \hat{\rho}(\mathcal{L}^{'^{,(j)}}_{[1\cdots n_t]}(\tau), \hat{\mathsf{model}}^{(j)}(\tau)). \tag{13}$$

They are the optimal ones with highest SNR he can get. Similar distinguishers can also be introduced into TLOB. Here we can see that TLOB is very simple. It only includes a linear measurement step and a traces optimal sorting and biasing step, both of which only bring very limited time and memory consumption. Moreover, TLOB only biases a small number of power traces to perform leakage detection, which significantly improves the efficiency.

Suppose that the power consumption of a device can be profiled using Hamming weight model. The evaluator profiles the mean power consumption for each Hamming weight. Let's give an intuitive example using the 25600 samples in our simulation. We use TLOB to bias 12800 power traces with high SNR from a total number of 25600 measurements. The experimental result is shown in Fig.3. Time samples biased by TLOB are very close to the mean power consumption leakage model $\hat{\mathsf{model}}$ of each Hamming weight. When the variance of noise approaches 0, the $\hat{\mathsf{model}}$ linearly correlates to the power consumption. If the evaluator performs correlation based leakage detection tests on these traces, the correlation coefficient will be very high. This also validates our power consumption assumption in Formula 5.

DTOB, such as [17, 15, 24] introduced in Section 1.3, improves SNR through enlarging $\hat{\mathsf{var}}(\mathcal{L}_d(\tau))$. The $\hat{\mathsf{var}}(\mathcal{L}_n(\tau))$ of noise is unchanged. Unlike DTOB, TLOB improves SNR through indirectly reducing $\hat{\mathsf{var}}(\mathcal{L}_n(\tau))$, since the power traces with small noise are biased and others are discarded. It is worth noting that the purpose of noise reduction is to minimize the noise on power traces. When biasing a fixed number of traces using TLOB, the noise of them are the smallest, since they are the ones closest to their $\hat{\mathsf{model}}$. In this case, the correlation coefficients output by correlation tests approach 1.00. As far as we know, there is no noise reduction tool that can improve the correlation coefficient $\hat{\rho}$ of CPA to such a height. However, how to choose a good threshold n_t (see Equation 13) is worth our further discussion, which will be detailed in Section 4.

3.4 Chosen Plaintexts based TLOB

Similar to DTOB, Hamming weight based biasing proposed by Hu et al. can improve $\hat{\mathsf{var}}(\mathcal{L}_d(\tau))$. Since it only biases traces according Hamming weights of intermediate values and does not sort the corresponding noise level, we do not compare it with DTOB and

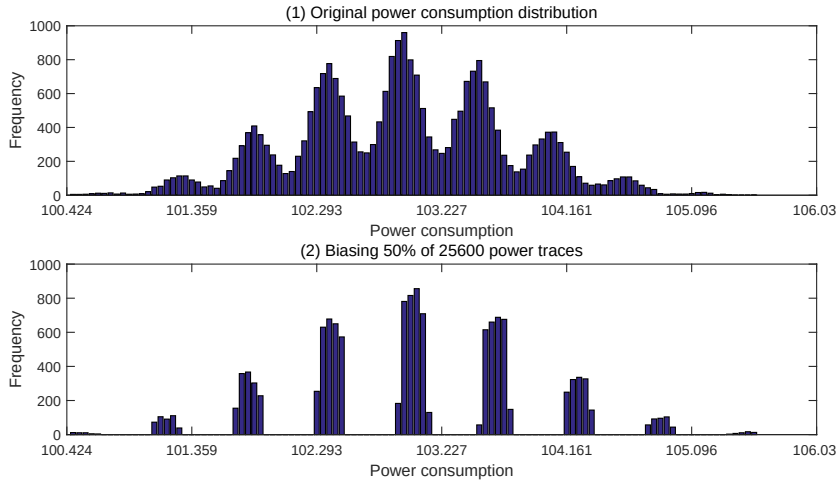

Figure 3: Biasing power traces using our TLOB.

TLOB in Fig.4, Fig.5 and Fig.6. The evaluator can combine the advantages of it and TLOB to detect leakage, the scheme of which is named Chosen Plaintexts based TLOB (CPTLOB) in this paper. It firstly encrypts plaintexts with large or small Hamming weights of intermediate values (e.g. Hamming weights 0, 1, 7 and 8 of AES S-box outputs). Then, TLOB is used to further optimize the power traces biasing. In this way, $\hat{\mathsf{var}}\left(\mathcal{L}_d\left(\tau\right)\right)$ is improved and $\hat{\mathsf{var}}\left(\mathcal{L}_n\left(\tau\right)\right)$ is reduced, the total number of power traces used in leakage detection can be greatly reduced.

Although TLOB can significantly improve the $\hat{\rho}$-s in leakage areas, it also improves the $\hat{\rho}$-s in areas without leakage. CPTLOB can't improve the $\hat{\rho}$-s in leakage areas to the same height as TLOB, it reduce the $\hat{\rho}$-s in areas without leakage (see Section 5). The evaluator can choose TLOB or CPTLOB according to the actual situation.

4 Threshold of the Number of Biased Traces

We find in our experiments that when the number of biased traces n_t increases, $\hat{\rho}$ decreases gradually. When n_t reaches N, it drops to the smallest. This indicates that TLOB accurately sorts power traces according to their noise level, and accurately biases traces with high SNR. However, It's not possible to intuitively determine power traces biased by which scheme (in Fig.1 and Fig.3) has higher SNR and greater $\hat{\rho}$. So, we compare DTOB and TLOB using the simulated traces, of which the parameters (e.g. mean power consumption leakage model $\hat{\mathsf{model}}$ and noise variance σ^2) are easy to control.

4.1 Consistency of Correlation

Consistency of correlation, means that the $\hat{\rho}$-s of all TOB schemes are the same when $n_t = N$, regardless of their performance. For a good TOB scheme, $\hat{\rho}$ should be high at the beginning, then monotonically decreases. The performance of it determines the height and the decline speed of $\hat{\rho}$.

The σ^2 of noise in our experiments is set to 4, 20, 100, 500 respectively. 20000 power traces and 10-fold cross validation are used, with 2000 traces in the test sets and other 18000 in the training sets for each repetition. We then perform correlation-based leakage detection tests on the biased traces. The experimental results are shown in Fig.4.

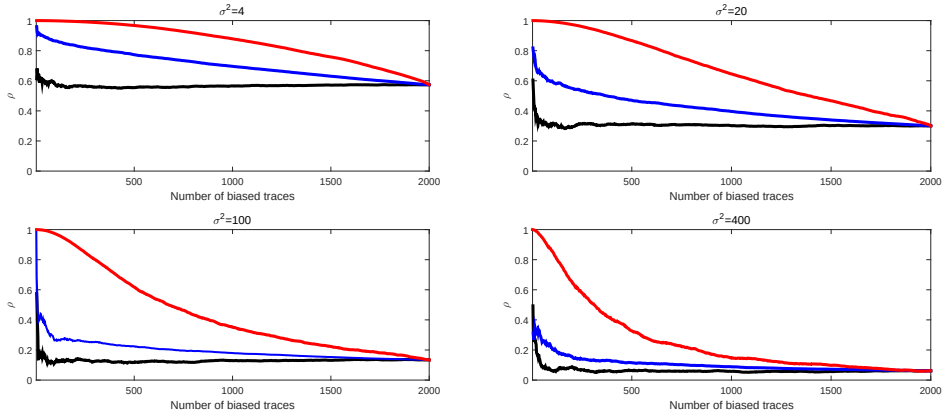

Figure 4: ρ-tests under different SNRs using traces without pre-processing (black), pre-processed by DTOB (green) and TLOB (red).

In the case of small noise, the evaluator can obtain a high correlation coefficient if he directly uses the power traces without pre-processing to perform leakage detection. With the increase of noise, the correlation coefficient drops rapidly. Larger noise makes $\hat{\rho}$ require more traces to become stable. It is about 0.583, 0.307, 0.139 and 0.063 when the σ^2 is 4, 20, 100 and 400 respectively. If the same number of power traces are biased, the $\hat{\rho}$-s corresponding to TLOB are significantly higher than the ones corresponding to CPA and DTOB. This also indicates that TLOB can more accurately bias power traces with high SNR.

The $\hat{\rho}$-s corresponding to TLOB and DTOB decrease, too. When a few traces are biased, the $\hat{\rho}$-s fluctuate. However, unlike directly performing tests on power traces without pre-processing, $\hat{\rho}$-s corresponding to TLOB and DTOB decrease monotonically with the increase of n_t. When noise enlarges, $\hat{\rho}$-s corresponding to DTOB and classical CPA without preprocessing get closer. This indicates that noise level plays a very important role in DTOB. Large noise makes the biasing inaccurate. However, ρ tests using our TLOB are still very robust. This is the experimental result of our simulation.

It's worth noting that $\hat{\rho}$ corresponding to DTOB is larger than the one of classical CPA without pre-processing, and $\hat{\rho}$ corresponding to TLOB is the largest when the same number of traces are biased. However, the higher $\hat{\rho}$ doesn't mean the better performance of leakage detection. A good leakage detection strategy should be able to make $\hat{\rho}$-s in leakage areas significantly higher than these in areas without leakage. Here we use simulation experiments to compare the performance of CPA, DTOB and TLOB at different noise levels and different numbers of traces, so as to provide a reference for evaluator to decide the threshold n_t.

4.2 Threshold under Different SNRs

We simulate 40000 power traces with σ^2 equaling 100, 400, 1600 and 6400, of which the SNR is 0.02, 0.005, 0.0013 and 0.000315 respectively. To compare the ability of ρ-test on traces without pre-processing, pre-processed by DTOB and TLOB to distinguish the POIs from other time samples without leakage, we additionally simulate 10 time samples without information leakage under the same noise level. The experimental results are shown in Fig.5.

With the same results getting from Section 4.1, with the decrease SNR of power traces, $\hat{\rho}$-s corresponding to CPA, DTOB and TLOB decrease. They are finally equal when

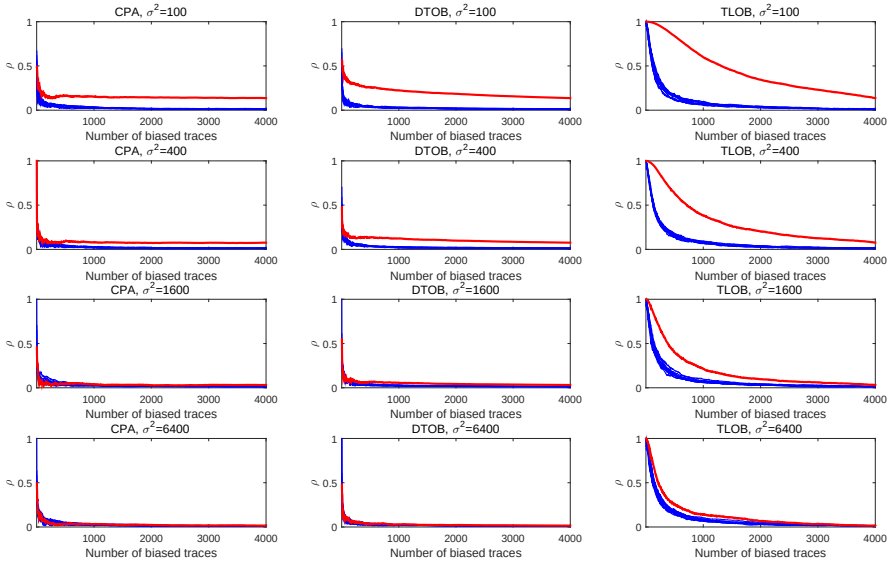

Figure 5: ρ-tests under different SNRs using traces without pre-processing (black), pre-processed by DTOB (green) and TLOB (red).

$n_t = N$. $\hat{\rho}$ corresponding to classical CPA without pre-processing and DTOB appear a small peak respectively when σ^2 equals 100 and 400. $\hat{\rho}$ corresponding to DTOB drops from about 0.50 (0.16) to 0.1365 (0.079) when σ^2 equals 100 (400). However, large noise makes the $\hat{\rho}$-s in leakage points and non-leakage points quickly drop to 0 and become indistinguishable. The benefits of biasing traces in DTOB also reflects when σ^2 equals 100. However, this is not reflected when σ^2 equals 400.

Compared to CPA and DTOB, TLOB performs better under different noise levels. It can accurately bias traces with high SNR when σ^2 equals 100, 400, 1600 and 6400 respectively. This indicates that TLOB has great ability to overcome noise, makes the $\hat{\rho}$-s in leakage areas and non-leakage areas distinguishable. When σ^2 equals 6400 and 4000 test traces are used, a small peak appears in TLOB. However, this does not happen in CPA and DTOB. The number of power traces with high SNR reduces when noise enlarges, which makes $\hat{\rho}$ smaller. If the noise enlarges, ρ tests using TLOB may not be able to detect leakage. In order to detect leakage, more traces are required. As shown in Fig.5, less traces n_t used by the evaluator, the higher SNR of them biased using TLOB. However, if n_t is very small, the $\hat{\rho}$-s in areas without leakage are also very high, which affects the results of leakage detection. Moreover, if n_t is large, noise of the biased power traces is large, the advantages of biasing power traces can't be reflected. $\frac{N}{10} \leq n_t \leq \frac{N}{4}$ may be a good choice as shown in Fig.5.

4.3 Threshold under Different Numbers of Traces

In order to compare the division of $\hat{\rho}$-s in leakage areas and non-leakage areas, we simulate 10000, 20000, 40000 and 80000 traces under the same noise level ($\sigma^2 = 100$). Similar to Section 4.2, 10 time samples without leakage are additionally simulated. With the increase number of biased power traces, the $\hat{\rho}$ of POI tends to be stable at about 0.20. $\hat{\rho}$-s corresponding to CPA and DTOB in leakage areas and non-leakage areas are indistinguishable when n_t is small. With increase of n_t, $\hat{\rho}$-s in the leakage areas increase and

decrease to 0 in areas without leakage. A small peak appears on the POI.

DTOB and TLOB are TOB schemes, which sort power traces and bias ones with high SNR. The $\hat{\rho}$ corresponding to them gradually declines and finally equals the one corresponding to ρ test on traces without preprocessing. The increase of noise reduces the number of traces with high SNR, accelerates the download trend of $\hat{\rho}$. The performance of DTOB and TLOB determines the height of $\hat{\rho}$ when n_t is small. $\hat{\rho}$ corresponding to DTOB fluctuates sharply when n_t is small (as shown in Fig.6). A similar phenomenon occurs in CPA, too. The larger the noise, the smaller the initial value of $\hat{\rho}$. It decreases from about 0.50 to about 0.140 in Fig.6. This also indicates that DTOB can no longer accurately bias traces. Without pre-processing, the increase number of power traces has no significant effect on DTOB.

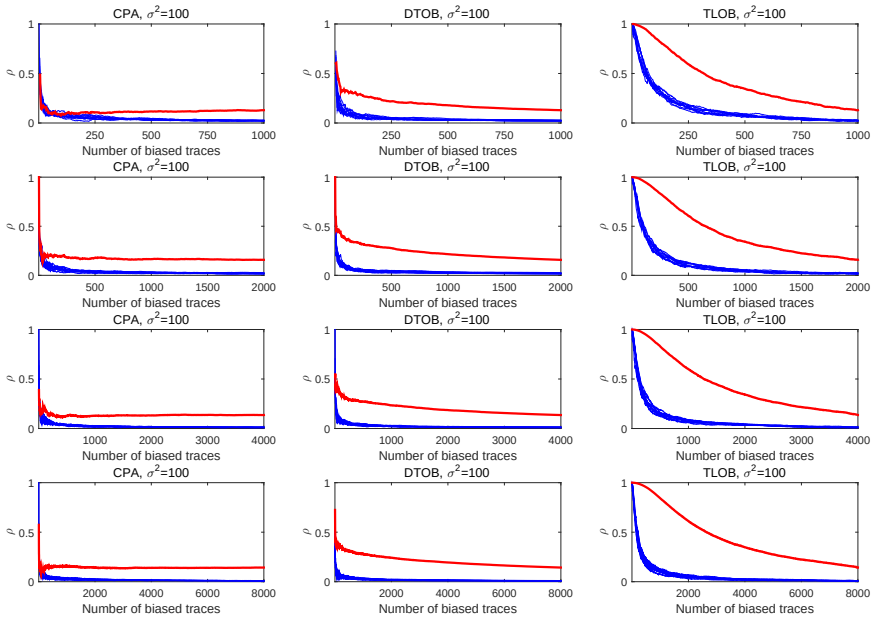

Figure 6: ρ-tests using different numbers of traces without pre-processing (black), preprocessed by DTOB (green) and TLOB (red).

Compared to DTOB, TLOB performs better. When a few traces are biased, $\hat{\rho}$-s of POI are close to 1.00 and gradually declines to about 0.14. They are high when $n_t < \frac{N}{2}$. As shown in the 4 sub graphs of TLOB, $\hat{\rho}$-s of time samples without leakage decline faster than the ones of POI. They are close to 0 when $n_t > \frac{N}{4}$. The difference of $\hat{\rho}$-s is most obvious when $\frac{N}{10} < n_t < \frac{N}{2}$. The experimental result also validates the conclusions obtained in Fig.4 and Fig.5. So, $\frac{N}{10} < n_t < \frac{N}{2}$ is a good threshold.

Cautionary Note. There are several factors affecting the experimental results in the simulation. However, the optimal threshold n_t that making $\hat{\rho}$ largest is hard to obtain. Since the traces in the real experiments are more complex. So, we only try to find a good threshold in our experiments. We suggest the evaluator choose a threshold that the correlation is high and the distinction between the interesting points and the uninteresting points is clear. The total number of power traces used can be huge, but the number of biased traces had better be small. Sometimes, dozens or hundreds of them are enough. $\frac{N}{10} < n_t < \frac{N}{2}$ can also be used here.

5 Software and Hardware Experimental Results

We use the simulation experiments in the previous sections to compare the performance of DTOB, TLOB and classical CPA without pre-processing under different numbers of power traces and different SNRs, since the parameters of simulated traces are easy to control. We perform our experiments on real leakages to compare them again in this section. It is worth noting that CPA using biased power traces is usually performed on POIs, since these points leak more information, the attacker gets better attack efficiency. Although DTOB in [17] analyzed all time samples, the whole traces were sorted only according to the POI that leaked most information. Our purpose here is to detect leakage, so we perform these 4 schemes on each time sample independently.

5.1 Experiments on AT89S52 Micro-controller

Our first experiment is performed on an AT89S52 micro-controller, the clock frequency of which is 12MHz. The shortest instructions take 12 clock cycles for execution. We use a Tektronix DPO 7254 oscilloscope to capture measurements of the look-up table instruction "MOVC A,@A+DPTR". We acquire 10000 power traces, of which the length is 4000 samples. We use the $2000^{th} \sim 2300^{th}$ samples to perform our leakage detection tests, and 10-fold cross-validation is used on MATLAB R2014b. The experimental result of ρ-test using classical CPA without pre-processing is shown in Fig.7. Leakage occurs between the 50^{th} and the 200^{th} time samples. The highest correlation coefficients of two most obvious leakages at the 58^{th} and 142^{th} time samples are about 0.376 and 0.453. The $\hat{\rho}$-s of two leak areas around these two time samples are between 0.1 and 0.3, compared to about 0 of areas without leakage.

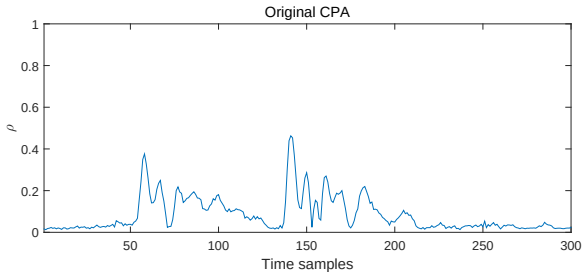

Figure 7: ρ tests using traces without pre-processing.

Let n denote the number of traces corresponding to Hamming weights 0, 1, 2, 4, 5, 6, 7 and 8 of S-box outputs, which is about 720 in our experiments. We bias traces to perform correlation based leakage detection tests in our next experiments (as shown in Fig.8, Fig.9 and Fig.10). It is worth noting that we use 10-fold cross-validation in our experiments. n in the test set $\mathcal{L}_t^{(j)}$ $(1 \leq j \leq 10)$ stated in Section 2 changes in a small range. However, this does not affect the mean of $\hat{\rho}$-s. Since DTOB, TLOB and CPTLOB can bias corresponding number of traces to calculate the $\hat{\rho}$ regardless of n.

We bias 500, 400, 300 and 200 power traces for leakage detection tests respectively and repeat this operation for each time sample. The choice of these parameters is not arbitrary, but has a relationship with the power consumption characteristics of our AT89S52. When we bias a small proportion of traces, the power consumption of some samples on traces are very close to their means, which makes the denominator in the formula of Pearson correlation coefficient approach 0. So, MATLAB outputs "NAN". We determine the parameters through many experiments.

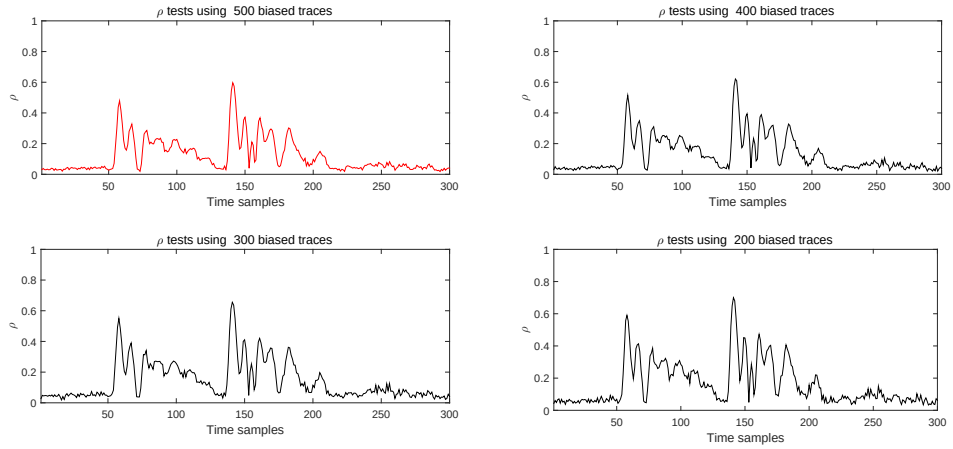

Figure 8: ρ tests using DTOB pre-processed traces.

TOB always outputs the traces closest to their corresponding mean power consumption model, which we regard as the current optimal ones. As shown in the figure, the correlation of the leakage areas is higher when leakage detection tests are performed with a smaller number of traces. This also validates the correctness of our TOB.

ρ-test using DTOB-biased traces are shown in Fig.8. The correlation-based tests detect more leakage when biasing power traces with large variance $\hat{\mathsf{var}}\left(\mathcal{L}_d\left(\tau\right)\right)$. The instruction "MOVC A, @A+DPTR" has several leakage areas. The 58^{th} and 142 are two most informative samples, near which there are two wide areas showing clear leakage. In fact, DTOB is one of TOB schemes, the evaluator gets higher correlation coefficient $\hat{\rho}$ if he biases less power traces to perform tests. The correlation of the leak areas are generally improved by biasing traces (as shown in Fig.8). However, the change is not very obvious. When biasing 500, 400, 300 and 200 power traces, the correlation coefficients of the two most informative samples are (0.479, 0.573), (0.514, 0.610), (0.550, 0.655) and (0.589, 0.700) respectively. The $\hat{\rho}$-s of two leak areas are about from 0.2 to 0.4, compared to smaller than 0.10 of areas without information leakage. This also indicates the limitation of DTOB. The SNR of the biased traces is improved, but limited.

Unlike DTOB, TLOB and CPTLOB perform better on biasing traces (as shown in Fig.9 and Fig.10). $\hat{\rho}$-s of leak areas are more than 0.60 when biasing 300 and 200 power traces. Moreover, with the decrease number of biased traces, $\hat{\rho}$ increases gradually and finally approaches 1.00. The $\hat{\rho}$-s in the leak areas corresponding to CPTLOB are between 0.20 and 0.60. The performance of CPTLOB is stable, the $\hat{\rho}$-s are below 0.20 in the areas without leakage. It increases with the decrease number of biased traces and the corresponding increase of SNR.

TLOB performs better than CPTLOB when the same number of traces are moderately biased. For example, when 300 traces are biased, most $\hat{\rho}$-s corresponding to TLOB in leakage areas are large than 0.60. $\hat{\rho}$-s of the two 58^{th} and 142^{th} time samples even reach 0.85 and 0.92. However, they are only about 0.20 in areas without leakage. Compared to TLOB, $\hat{\rho}$-s corresponding to CPTLOB on leak areas are bout 0.16, and the two most obvious samples are about 0.777 and 0.862. The reason for this phenomenon is because the number of biased traces is too small, TLOB biases traces with highest linearity, resulting in the difference of $\hat{\rho}$-s between leak areas and no leakage areas are not obvious. In this case, the evaluator can simply bias more traces. This does not affect the conclusion that TLOB performs better than CPTLOB in most cases. Since TLOB sorts all traces,

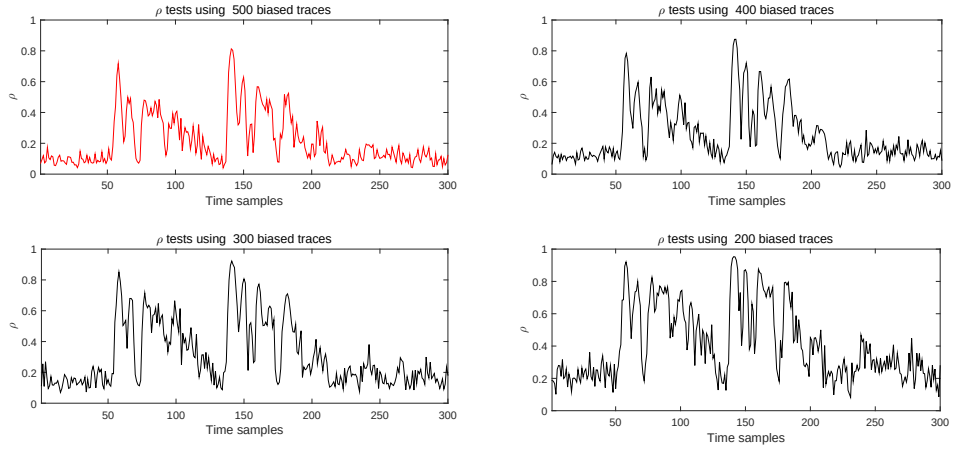

Figure 9: ρ tests using TLOB pre-processed traces.

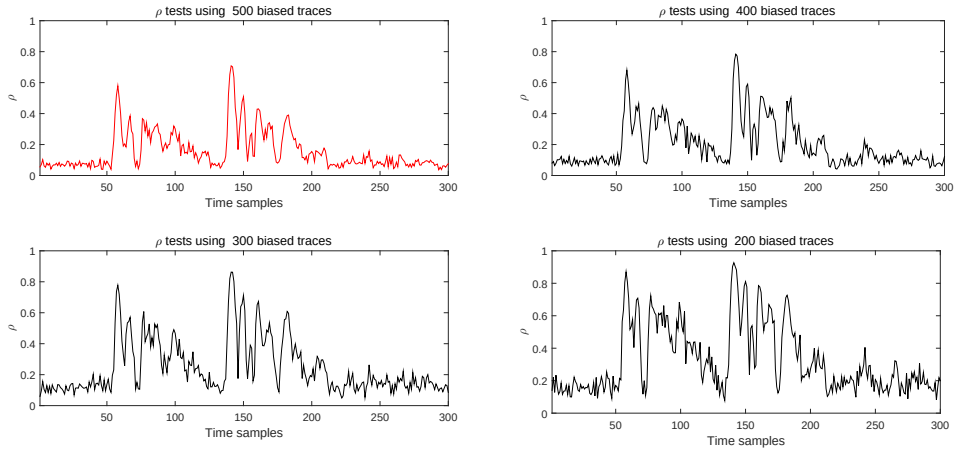

Figure 10: ρ tests using CPTLOB pre-processed traces.

CPTLOB only sorts traces corresponding to Hamming weights 0, 1, 2, 4, 5, 6, 7 and 8. If the same number of traces are biased, the ones biased by CPTLOB have larger noise since they are farther from their corresponding mean power consumption leakage model.

5.2 Experiments on DPA Contest v1

Our second experiment is perform on the measurements of the unprotected DES crypto-processor on the SecmatV1 SoC in ASIC provided by DPA *contest* $v1$ [1]. We attack the the output of the first S-box in the last round of DES. Time samples from 16000^{th} to 17000^{th} on the first 10000 power traces are used. Compared to power traces of our $AT89S52$ micro-controller, the power consumption characteristic of DPA *contest* $v1$ is better suited to bias a smaller number of power traces. DES has 4 bits S-box output, of which $\hat{\mathsf{var}}\left(\mathsf{L}_d\right) = 1.0667$. We bias power traces corresponding to Hamming distances 0, 1, 3, 4, which account for about 5/8. $\hat{\mathsf{var}}\left(\mathsf{L}_d\right) = 1.7778$ after biasing traces. We bias 600, 400, 200 and 100 from 1000 power traces in the test sets. Experimental results of ρ-tests

on traces without pre-processing are shown in Fig.11.

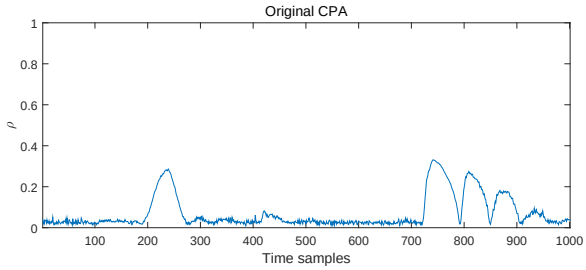

Figure 11: ρ tests using traces without pre-processing.

There are two leak areas $16200^{th} \sim 16270^{th}$ and $16720^{th} \sim 16910^{th}$ in the selected time samples. One and three peaks appear in these two areas, $\hat{\rho}$ of which is about 0.275, 0.331, 0.266 and 0.173 respectively. $\hat{\rho}$ is about 0.03 in the areas without leakage. For convenience, we use the symbol $\bar{\rho}$ to represent the correlation coefficients vector of these four peaks.

ρ tests using DTOB-biased traces are performed and the experimental results are shown in Fig.12. When biasing 600 traces with maximum variances of population mean, $\bar{\rho}$ is significantly improved to (0.347, 0.397, 0.328, 0.220). When biasing 400, 200 and 100 power traces, $\bar{\rho}$ changes to (0.395, 0.462, 0.373, 0.258), (0.480, 0.544, 0.428, 0.313), (0.532, 0.596, 0.478, 0.395) respectively. It is worth noting that TOB also improves the correlation in areas without leakage. In fact, TOB always biases the power traces that are most beneficial to the guess key, including correct one and wrong ones. This makes $\hat{\rho}$-s corresponding to both correct key and wrong guess keys improve. In other words, TOB always biases power traces that most linearly correlates to the profiled leakage model. This also explains the reason why correlation in areas without leakage improves. The $\hat{\rho}$-s in no leaky areas are about the same after 10-fold cross validation. When the number of biased traces changes from 600 to 100, $\hat{\rho}$-s in these areas are about 0.03, 0.03, 0.07 and 0.08 respectively. Ghost peaks do not appear in these areas, which reflects the robust of DTOB. Compared to classical CPA without pre-processing, DTOB performs better in correlation leakage detection tests. With the decrease of n_t, $\hat{\rho}$-s in leak areas improve, but not very obviously.

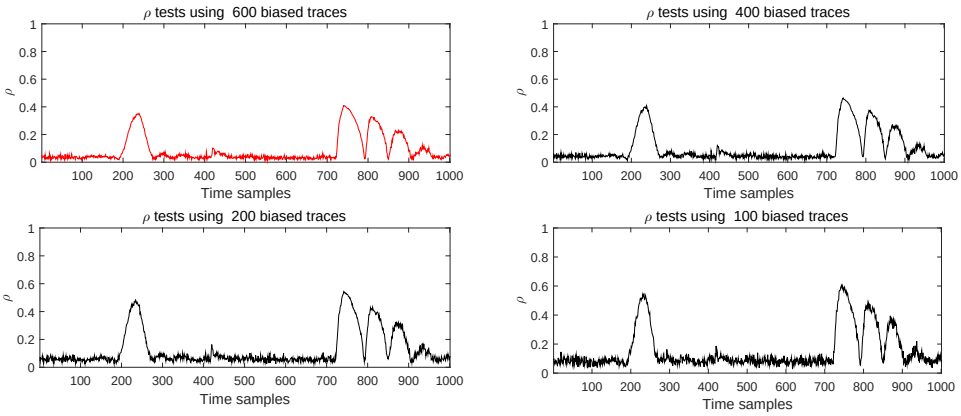

Figure 12: ρ tests using DTOB pre-processed traces.

Correlation coefficients increase significantly when DTOB, TLOB and CPTLOB are introduced into correlation based tests. The experimental results of ρ-tests using TLOB- and CPTLOB biased traces are shown in Fig.13 and Fig.14. When the same number of traces are biased from the same set, TLOB performs better than DTOB. Taking the correlation coefficients $\bar{\rho}$ of 4 peaks for example, when 600, 400, 200 and 100 traces are biased, $\bar{\rho}$ corresponding to CPTLOB changes to (0.322, 0.369, 0.293, 0.214), (0.510, 0.579, 0.483, 0.365), (0.770, 0.821, 0.756, 0.590) and (0.934, 0.956, 0.912, 0.842). The corresponding $\hat{\rho}$ in areas without leakage is average 0.03, 0.06, 0.10 and 0.20. $\bar{\rho}$ corresponding to TLOB changes to (0.521, 0.596, 0.522, 0, 360), (0.698, 0.761, 0.672, 0.526), (0.904, 0.924, 0.886, 0.784) and (0.970, 0.980, 0.966, 0.930). The corresponding $\hat{\rho}$-s in areas without leakage are average 0.04, 0.07, 0.17 and 0.30. $\hat{\rho}$-s of two leak areas are also higher than the ones in the corresponding locations of DTOB.

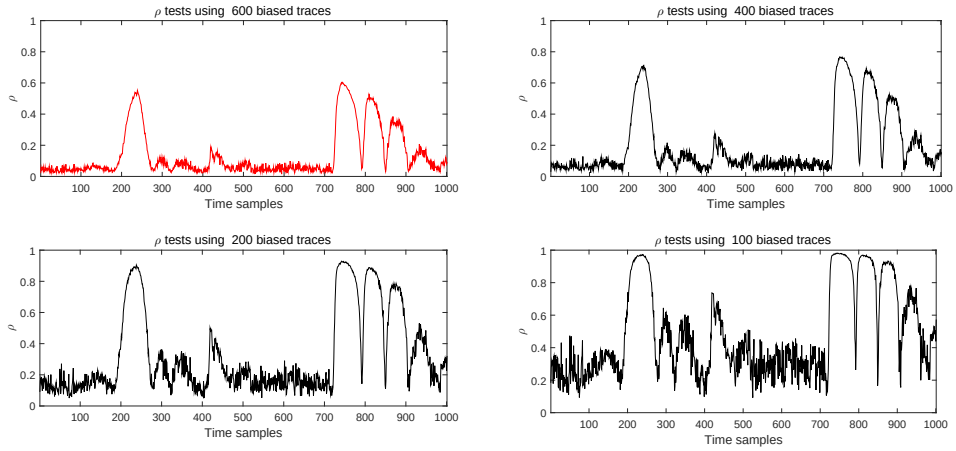

Figure 13: ρ tests using TLOB pre-processed traces.

This indicates that the biased traces have high SNR. Compared to DTOB, TLOB and CPTLOB performs better in 4 leakage areas. In addition, ρ-test using DTOB-biased traces appears 'lean and high' correlation in these 4 leakage areas. The correlation declines rapidly from the peaks to the edge areas. Unlike DTOB, ρ-test using TLOB- and CPTLOB-biased traces looks "fat and tall". This indicates that power traces with high SNR not only in the peaks and their vicinities but also in the edges of 4 leakage areas are biased, which makes the correlation in wide leakage areas improve. This also indicates that, compared to DTOB, ρ-tests using TLOB and CPTLOB can detect leakage more effectively.

As shown in Section 4, TLOB works very well when n_t is smaller than 50 percent of the total number of traces. The difference of $\hat{\rho}$-s in the leakage areas and areas without leakage are most obvious. If the evaluator enlarges n_t, $\hat{\rho}$-s in these areas decrease. If n_t is reduced, $\hat{\rho}$-s in these areas enlarges. $\hat{\rho}$-s in leakage areas keep growing, and become stable after close to the limit 1.00. If the evaluator keeps reducing n_t, the $\hat{\rho}$-s in areas without leakage grow rapidly (see the two sub graphs in Fig.13).

In fact, if the number of TLOB-biased traces is too small, the $\hat{\rho}$-s in leak areas and no leaky areas fluctuate very seriously and are indistinguishable. This experimental result is very similar to the one in Section 5.1. This also validates the conclusion that n_t can not be too small in Section 4. If n_t is properly chosen and the total number of power traces is fixed, TLOB is generally better than CPTLOB and DTOB. A good n_t makes the difference of $\hat{\rho}$-s in leak areas and areas without leakage more obvious.

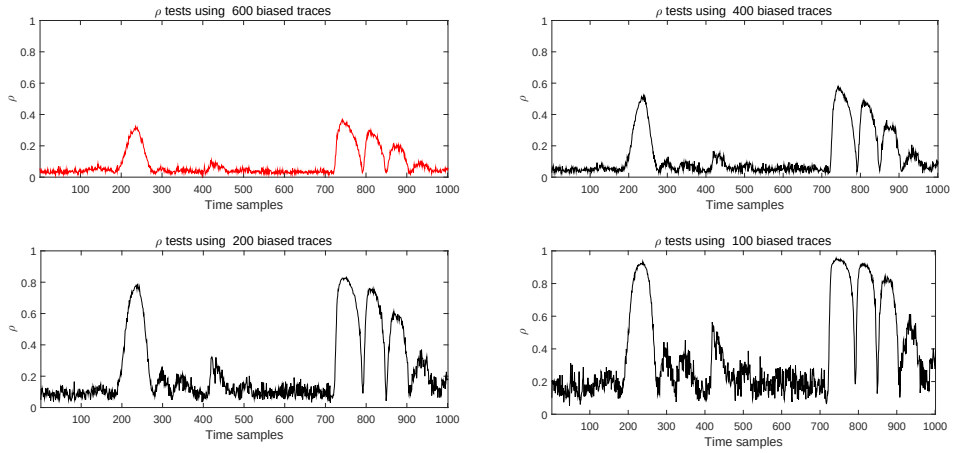

Figure 14: ρ tests using CPTLOB pre-processed traces.

6 Conclusions and Future Works

Biasing power traces is one of the most efficient de-noising pre-processing tools to improve SNR, which can significantly enhance leakage detection tests. In this paper, SNR of power traces is defined more accurately, and the concept of TOB is proposed for the first time. TLOB and CPTLOB, two specific TOB schemes are given. Compared to DTOB, TLOB and CPTLOB can bias traces with high SNR more accurately. TLOB based schemes including CPTLOB are very robust and perform very well even on very noisy traces. Taking advantage of known key and plaintexts, the evaluator can use a very small number of traces to perform leakage detection tests. The experimental results in Section 5 shows that the correlation coefficients of leakage detection tests using TLOB and CPTLOB approach 1.00. This indicates the high SNR of biased power traces, and high efficiency of TLOB based leakage detection tests. The ρ-test is only an example given here, TLOB can also be used in other kinds of tests, such as Student's t-test detailed in Section 2.1. Since the noise variance can be reduced very low.

Since CPA was proposed, a lot of optimizations have been done. However, as far as we know, there has no scheme that can improve the correlation coefficient to the same level of TLOB and CPTLOB. TLOB-based schemes are still worth further study. As we mentioned in Section 5, TOB always biases traces best correlating to the mean power consumption leakage model of intermediate values, if the number of biased traces is too small, the $\hat{\rho}$-s in areas without leakage are also very high. This makes the difference of $\hat{\rho}$-s in leakage areas and areas without leakage undistinguishable, which seriously affects the results of our detection tests. In this case, we can simply enlarge the number of biased traces. Other methods to optimize this are also worth studying. Moreover, we use a TA-like distinguisher to measure the noise level of traces. How to use other distinguishers to optimize TLOB and further improve the efficiency of leakage detection tests is also a very interesting problem.

Finally, as we mentioned in our paper, TLOB is based on known key and can only be used for evaluation purposes. The attackers can't use it to perform attacks. How to apply TLOB to unknown key based leakage detections and attacks is also very meaningful.

References

[1] Dpa contest. http://www.dpacontest.org/home/.

[2] D. Agrawal, B. Archambeault, J. R. Rao, and P. Rohatgi. The EM side-channel(s). In *Cryptographic Hardware and Embedded Systems - CHES 2002, 4th International Workshop, Redwood Shores, CA, USA, August 13-15, 2002, Revised Papers*, pages 29–45, 2002.

[3] M. Akkar, R. Bevan, P. Dischamp, and D. Moyart. Power analysis, what is now possible... In *Advances in Cryptology - ASIACRYPT 2000, 6th International Conference on the Theory and Application of Cryptology and Information Security, Kyoto, Japan, December 3-7, 2000, Proceedings*, pages 489–502, 2000.

[4] S. Bhasin, J. Danger, S. Guilley, and Z. Najm. NICV: normalized inter-class variance for detection of side-channel leakage. *IACR Cryptology ePrint Archive*, 2013:717, 2013.

[5] 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*, pages 16–29, 2004.

[6] 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*, pages 13–28, 2002.

[7] J. Cooper, E. DeMulder, G. Goodwill, J. Jaffe, G. Kenworthy, and P. Rohatgi. Test vector leakage assessment (tvla) methodology in practice, 2013. http://icmc-2013.org/wp/wp-content/uploads/2013/09/goodwillkenworthtestvector.pdf.

[8] J. Coron. Higher order masking of look-up tables. In *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, pages 441–458, 2014.

[9] J. Coron, E. Prouff, M. Rivain, and T. Roche. Higher-order side channel security and mask refreshing. In *Fast Software Encryption - 20th International Workshop, FSE 2013, Singapore, March 11-13, 2013. Revised Selected Papers*, pages 410–424, 2013.

[10] A. A. Ding, C. Chen, and T. Eisenbarth. Simpler, faster, and more robust t-test based leakage detection. In *Constructive Side-Channel Analysis and Secure Design - 7th International Workshop, COSADE 2016, Graz, Austria, April 14-15, 2016, Revised Selected Papers*, pages 163–183, 2016.

[11] F. Durvaux and F. Standaert. From improved leakage detection to the detection of points of interests in leakage traces. In *Advances in Cryptology - EUROCRYPT 2016 - 35th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Vienna, Austria, May 8-12, 2016, Proceedings, Part I*, pages 240–262, 2016.

[12] D. Genkin, A. Shamir, and E. Tromer. RSA key extraction via low-bandwidth acoustic cryptanalysis. In *Advances in Cryptology - CRYPTO 2014 - 34th Annual Cryptology Conference, Santa Barbara, CA, USA, August 17-21, 2014, Proceedings, Part I*, pages 444–461, 2014.

[13] G. Goodwill, B. Jun, J. Jaffe, and P. Rohatgi. A testing methodology for side-channel resistance validation, 2011. http://www.rambus.com/wp-content/uploads/2015/08/a-testing-methodology-for-side-channel-resistance-validation.pdf.

[14] S. Hajra and D. Mukhopadhyay. On the optimal pre-processing for non-profiling differential power analysis. In *Constructive Side-Channel Analysis and Secure Design - 5th International Workshop, COSADE 2014, Paris, France, April 13-15, 2014. Revised Selected Papers*, pages 161–178, 2014.

[15] W. Hu, L. Wu, A. Wang, X. Xie, Z. Zhu, and S. Luo. Adaptive chosen-plaintext correlation power analysis. In *Tenth International Conference on Computational Intelligence and Security, CIS 2014, Kunming, Yunnan, China, November 15-16, 2014*, pages 494–498, 2014.

[16] Y. Kim and H. Ko. Using principal component analysis for practical biasing of power traces to improve power analysis attacks. In *Information Security and Cryptology - ICISC 2013 - 16th International Conference, Seoul, Korea, November 27-29, 2013, Revised Selected Papers*, pages 109–120, 2013.

[17] Y. Kim, T. Sugawara, N. Homma, T. Aoki, and A. Satoh. Biasing power traces to improve correlation power analysis attacks. In *Constructive Side-Channel Analysis and Secure Design - COSADE 2010 - First International Workshop, Daemstadt, Germany, February 4-5, 2010*, pages 77–80, 2010.

[18] P. C. Kocher, J. Jaffe, and B. Jun. Differential power analysis. In *Advances in Cryptology - CRYPTO '99, 19th Annual International Cryptology Conference, Santa Barbara, California, USA, August 15-19, 1999, Proceedings*, pages 388–397, 1999.

[19] T. Le, J. Clédière, C. Servière, and J. Lacoume. Noise reduction in side channel attack using fourth-order cumulant. *IEEE Trans. Information Forensics and Security*, 2(4):710–720, 2007.

[20] S. Mangard, E. Oswald, and T. Popp. *Power analysis attacks - revealing the secrets of smart cards*. Springer, 2007.

[21] 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*, pages 486–505, 2013.

[22] T. S. Messerges, E. A. Dabbish, and R. H. Sloan. Examining smart-card security under the threat of power analysis attacks. *IEEE Trans. Computers*, 51(5):541–552, 2002.

[23] B. Noura, M. Mohsen, and T. Rached. Optimized power trace numbers in CPA attacks. In *The 8-th International Multi-Conference on Systems, Signals and Devices, Sousse, Tunisia, March 22-25, 2011. Proceedings*, pages 1–5, 2011.

[24] C. Ou, Z. Wang, D. Sun, X. Zhou, J. Ai, and N. Pang. Enhanced correlation power analysis by biasing power traces. In *Information Security - 19th International Conference, ISC 2016, Honolulu, HI, USA, September 3-6, 2016, Proceedings*, pages 59–72, 2016.

[25] E. Peeters. *Advanced DPA theory and practice*. Springer, 2013.

[26] O. Reparaz. Detecting flawed masking schemes with leakage detection tests. In *Fast Software Encryption - 23rd International Conference, FSE 2016, Bochum, Germany, March 20-23, 2016, Revised Selected Papers*, pages 204–222, 2016.

[27] K. Schramm, G. Leander, P. Felke, and C. Paar. A collision-attack on AES: combining side channel- and differential-attack. In *Cryptographic Hardware and Embedded Systems - CHES 2004: 6th International Workshop Cambridge, MA, USA, August 11-13, 2004. Proceedings*, pages 163–175, 2004.

[28] K. Schramm and C. Paar. Higher order masking of the AES. In *Topics in Cryptology - CT-RSA 2006, The Cryptographers' Track at the RSA Conference 2006, San Jose, CA, USA, February 13-17, 2006, Proceedings*, pages 208–225, 2006.

[29] K. Schramm, T. J. Wollinger, and C. Paar. A new class of collision attacks and its application to DES. In *Fast Software Encryption, 10th International Workshop, FSE 2003, Lund, Sweden, February 24-26, 2003, Revised Papers*, pages 206–222, 2003.

[30] F. Standaert. How (not) to use welch's t-test in side-channel security evaluations. *IACR Cryptology ePrint Archive*, 2017:138, 2017.

[31] K. Tiri and P. Schaumont. Changing the odds against masked logic. In *Selected Areas in Cryptography, 13th International Workshop, SAC 2006, Montreal, Canada, August 17-18, 2006 Revised Selected Papers*, pages 134–146, 2006.

[32] J. G. J. van Woudenberg, M. F. Witteman, and B. Bakker. Improving differential power analysis by elastic alignment. In *Topics in Cryptology - CT-RSA 2011 - The Cryptographers' Track at the RSA Conference 2011, San Francisco, CA, USA, February 14-18, 2011. Proceedings*, pages 104–119, 2011.

[33] L. Zhang, A. A. Ding, F. Durvaux, F. Standaert, and Y. Fei. Towards sound and optimal leakage detection procedure. *IACR Cryptology ePrint Archive*, 2017:287, 2017.