A Cautionary Note Regarding the Usage of Leakage Detection Tests in Security Evaluation

Carolyn Whitnall and Elisabeth Oswald

University of Bristol, Department of Computer Science,
Merchant Venturers Building, Woodland Road, BS8 1UB, Bristol, UK
{carolyn.whitnall, elisabeth.oswald}@bris.ac.uk

Abstract. An established ingredient in the security evaluation of cryptographic devices is *leakage detection*, whereby physically observable characteristics such as the power consumption are measured during operation and statistically analysed in search of sensitive data dependencies. However, depending on its precise execution, this approach potentially suffers several drawbacks: a risk of false positives, a difficulty interpreting negative outcomes, and the infeasibility of covering every possible eventuality. Moreover, efforts to mitigate for these drawbacks can be costly with respect to the data complexity of the testing procedures. In this work, we clarify the (varying) goals of leakage detection and assess how well-geared current practice is towards meeting each of those goals. We introduce some new innovations on existing methodologies and make recommendations for best practice. Ultimately, though, we find that many of the obstacles cannot be fully overcome according to existing statistical procedures, so that it remains to be highly cautious and to clearly state the limitations of the methods used when reporting outcomes.

1 Introduction

Ever since Kocher et al. [27] raised awareness of the vulnerability of 'provably secure' cryptography to attacks exploiting auxiliary information not accounted for in traditional security models – information such as the power consumption or other measureable characteristics of cryptographic devices – designers and certification bodies have been increasingly concerned with ensuring and evaluating the physical security of cryptographic implementations. Given the difficulty of appropriately modelling the full range of physical threats so as to build perfect theoretic security into the algorithms themselves [52], it typically remains to subject actual products to experimental testing in a laboratory setting. One option is to run 'all' the best known attacks developed in the side-channel literature to date, which becomes costly given the growing number of such strategies and the difficulty of determining *a priori* which are the most pertinent to a particular scenario (see e.g. [11,50]). An increasingly preferred option is to rely on *leakage detection* testing along the lines of the Test Vector Leakage Assessment (TVLA) framework first proposed by Cryptography Research, Inc. (now Rambus) [22].

Rather than aim at the successful extraction of sensitive information from side-channel measurements, as an attack-based evaluation would do, leakage detection simply seeks evidence (or convincing lack of evidence) of sensitive data dependencies in the measured traces. TVLA does this via a suite of Welch's t-tests targeting mean differences in carefully chosen partitions of trace measurements. For example, the fixed-versus-random test looks for a statistically significant difference between a trace set associated with a fixed plaintext input and another trace set associated with randomly varying inputs. Alternatively, the leakage associated with a specific intermediate value (such as an S-box output) can be targeted by comparing a trace set that has been partitioned into two according to the value of that bit or byte. Both the 'specific' and the 'non-specific' type tests are univariate and are performed on each point in a trace set separately in order to draw conclusions about the overall vulnerability of the implementation. So called 'higher order' tests exist to target

leakage of more complex functional form that does not present via differences in the mean but can be found in higher order (joint) statistical moments; these typically entail pre-processing the traces before performing the same univariate point-wise test procedures [45].

Detection methodologies outside of the TVLA framework use other quantities such as the mutual information [7,8,32] or the correlation [15] between measured traces and known intermediates, and the F-statistic for the classes imposed on a measured trace by the values of a known intermediate [4]. Crucially, though, all of these approaches are essentially statistical hypothesis tests, regardless of whether the statistical formalities have been understood and observed or not.

A short-falling of leakage detection in practice has been a lack of transparency about what the tests do or don't show, which is essential for responsible and meaningful interpretation of outcomes, and especially for ensuring that conclusions are 'fair' from one evaluation to another – a key priority for the purposes of certification. In particular, it is seldom clear what to think or how to proceed if a test 'fails' to find leakage, or how to ensure like-for-like rigour across different target devices and between different lab settings.

To an extent, this can be addressed by revisiting the often-overlooked formalities of the underlying methodologies. Classical statistics places a strong emphasis on informed test design and awareness and control of error rates. In this work, we seek to better understand the tools available and their application to the goal of leakage detection.

But first, it is necessary to clarify what that goal ultimately *is*. We note that it is different in different settings, and delineate four broad possibilities ranging in degree of ambition (see Section 2). The task is then to explore the statistical formalism of the various leakage detection tests available (overviewed in Section 3) in order to assess how well they are able to meet each of the four goals, and how (if at all) they can be adapted in order to meet them better (see Sections 4 and 5). We also examine the related issue of coverage – how to measure and improve the extent to which test strategies can be considered comprehensive relative to the full range of possible vulnerabilities (see Section 6).

Unfortunately, we find that existing procedures are extremely limited in their scope to fulfil even the most modest of detection goals. Since many of the obstacles arise from the impact of multiple testing on error rates (and on the easy analysis of error rates) we explore the possibility to bypass these particular issues via a multivariate approach. Drawing on recent work [34] advocating the use of the Hotelling's T^2 test to detect leaks in whole traces or (by way of compromise) trace segments, we experiment with the idea of *clustering* trace points so as to reduce the number of (multivariate, arguably independent) tests to perform (see Section 7). Such an approach is not suitable for all possible evaluation goals, and is found to be unreliable, or at least highly sensitive to configuration decisions.

We conclude from our investigations that the challenge faced by evaluators is really not trivial: considerable breakthroughs are yet to be made in the wider statistics literature that would enable the types of ideal solution needed to produce truly fair and like-for-like comparison across target devices and analysis scenarios. We attempt, in the light of this lack, to provide sound advice about the best known approaches towards each goal and how to interpret and present results with appropriate caution (see Section 8).

2 The Goal of Leakage Detection

Leakage detection is typically carried out as part of an exercise to evaluate the security of a cryptographic device. It might be performed by an evaluation laboratory with the aim of providing security certification for when the device goes on sale, or it might be an in-house effort during

the development process in order to highlight and fix potential problems prior to formal external evaluation.

Either way, it is helpful to recognise that the particular goal of a detection attempt can vary, and that the approach taken needs to be chosen with the desired end result in mind. We have identified four different possible intentions:

Certifying vulnerability: Find a leak in **at least one** trace point. In such a case it is especially important to avoid false positives (that is, concluding there is a leak where there isn't one).

Certifying security: Find no leaks having tested thoroughly. Here false negatives (failure to find leaks that are really there) become a concern. As we will see, the statistical methods used for leakage detection cannot 'prove' that there is no effect, they can at best conclude that there is evidence of a leak or that there is no evidence of a leak. Hence it is especially important to design tests with 'statistical power' in mind – that is, to make sure the sample size is large enough to detect a present effect of a certain size with reasonable probability (see Section 4.1). Then, in the event that no leak is discovered, these constructed features of the test form the basis of a reasoned interpretation. A further, considerable challenge implicit to this goal is the necessity to be convincingly exhaustive in the range of tests performed – that is, to target 'all possible' intermediates and all relevant higher-order combinations of points. (This suggests analogues with the idea of *coverage* in code testing, which we discuss in Section 6).

Demonstrating an attack: Map a leaking point (or tuple) to its associated intermediate state(s) and perform an attack. Typically it is of interest to report attack outcomes and/or projections derived from those outcomes, such as the number of traces required for key recovery, and/or the global key rank for a given sample size. False positives are undesirable as they represent wasted effort in the attack phase.

Highlighting vulnerabilities: Map all exploitable leakage points to their associated intermediate states in order to guide designers seeking to secure the device. This has something in common with certifying security, as both require an 'exhaustive' analysis, and something in common with demonstrating an attack, as both require being able to locate the source of the leakage. False negatives are of greater concern than false positives as they represent vulnerabilities that will remain unfixed.

3 Existing Methods for Leakage Detection

Most leakage evaluation procedures, regardless of their ultimate goal, are essentially statistical hypothesis tests or informal adaptations thereof.

3.1 Statistical Hypothesis Tests

A statistical hypothesis test begins with the formulation of a *null* hypothesis and an *alternative* hypothesis about some random process or population of interest. Testable hypotheses have been considered key to the process of learning from empirical evidence since (at least) the emergence of modern science in the 17th century, and the particular choices made dictate the information you are able to derive from the observed data.

It then typically entails the computation of a test statistic with a known or derivable (e.g. through randomised resampling) distribution under the null hypothesis. If the observed value is 'extreme' – i.e., the probability of such a value occurring under the null hypothesis is smaller than some fixed probability of false rejection α – then the null is rejected. Otherwise the null is 'not rejected'. Crucially, this is not the same as 'accepting' the null. For example, there may simply not

be enough data to provide conclusive evidence against it. This is why it quickly becomes important to understand the concept of *statistical power*: the probability of rejecting the null in the case that the alternative hypothesis is correct, which, for a given α, depends on the magnitude of the effect as well as the sample size. We expand on the notion of power, α, effect size and sample size in Section 4.1.

There are many different ways of constructing hypothesis tests, depending on the question that you seek to answer. For example, they be used to make informed judgements about the sameness of two populations, as characterised by particular parameters (e.g. in the case of the t-test [23], which forms the basic component of the TVLA framework [22]), by frequency tabulations (e.g., Pearson's χ^2 test for discrete distributions [36]), or by the empirical distribution function (e.g. two sample Kolmogorov–Smirnov [28,49]). Alternatively, they can be used to decide if a particular parameter or quantity is larger or smaller than (or different to, in the case of 'two-sided' tests) a hypothesised value (often zero). Hypothesis tests are needed because estimates computed on sampled data can only ever acquire the true underlying parameters up to a certain precision, so that whether or not there is a match with the null cannot be decided on the basis of exact equality but on how large the disparity between them is relative to certain distributional expectations.

In the case of leakage detection, the null hypothesis is often that two sets of trace measurements associated with different (known) intermediate values have the same distribution, with the alternative hypothesis being that the distributions are different. Another (less common) approach is to test the null hypothesis that some measure of correspondence between the data and the trace measurements (e.g. the mutual information [7,8,32] or the correlation [15]) is zero, versus the alternative hypothesis that it is non-zero.

3.2 Typical Detection Strategies

Any dependency of the measured side-channel on sensitive data presents a potential vulnerability to an attacker. An 'ideal' test would therefore be one that simply seeks to reject the null hypothesis that the side-channel and the sensitive data are independent, in favour of the alternative hypothesis that they are related in some arbitrary way.

On the other hand, when the goal goes beyond finding or not finding leakage, to mapping, understanding, exploiting and/or addressing leakage, arbitrary detection (to the extent it is possible) suffers drawbacks: it is typically difficult to translate a non-specific vulnerability into a successful exploitation, or even to tie it to a particular operation. For the third and fourth goals it might therefore be preferable to use *specific* tests, targeting particular intermediate values. This facilitates the mapping of the leakage – although, at the expense of an increased number of separate tests needing to be performed in order to be confident of covering all eventualities.

In the following two subsections we formalise and discuss a range of arbitrary and specific tests, most of which exist in some form in the SCA literature.

Detecting Arbitrary Leaks Two variables A and B are statistically independent if and only if their joint distribution $F_{A,B}$ is identical to the product of their marginal distributions F_A, F_B: $F_{A,B} = F_A \cdot F_B$. In side-channel terms, the question of whether or not trace measurements Y reveal information about (i.e. are dependent on) the key (or plaintext) X can be formalised as a hypothesis test via this definition:

$$H_0 : F_{X,Y} = F_X \cdot F_Y \text{ vs. } H_{alt} : F_{X,Y} \neq F_X \cdot F_Y \tag{1}$$

Unfortunately, mathematical theory does not always translate neatly into statistical practice. Performing a hypothesis test requires the formulation of a test statistic that can be computed from the data sample and that has a known distribution under the null hypothesis. (Meanwhile, as we shall see in Section 4, evaluating the statistical power of the test requires also knowing the distribution under the alternative hypothesis).

If both X and Y were discrete variables, a χ^2 test could be used to decide between the above two hypotheses. However, side-channel measurements are strictly non-categorical. Alternatives for quantitative data include the Kolmogorov–Smirnov test [28,49] (adopted in [57,59] for side-channel *attacks* but, as far as we know, as-yet unused to directly test for independence in a detection setting), or tests based on distance correlation [53] (as-yet unused in *any* side-channel setting, as far as we know), all of which are non-parametric in nature. Being non-parametric means that they do not rely on assumptions about the functional forms of the distributions of A, B, and (A, B). Often (but not necessarily) this also implies that they compare based on general distributional 'shapes', rather than summary measures such as distributional moments. Thus they are potentially capable of detecting arbitrary leaks, as opposed to just those that manifest in particular moments. However, genericness comes at a cost: the number of observations required to draw statistically significant conclusions tends to be much higher for non-parametric tests than for parametric ones.

In order to arrive at more efficient tests, two approaches can be taken: rephrasing the problem or making additional assumptions.

Rephrasing the Problem Rather than purely attempting to detect an information leak (i.e. a data-dependency), one can attempt to *quantify* an information leak and decide whether it is non-zero. This approach has led to the development of robust mutual information tests for continuous and discrete data (CMI and DMI), of which the former has been shown to be suitable for typical power leakage traces [32].

$$H_0 : \mathrm{I}(X;Y) = 0 \text{ vs. } H_{alt} : \mathrm{I}(X;Y) \neq 0 \tag{2}$$

The CMI test has been shown to be capable of detecting arbitrary leaks, with the drawback that it naturally requires more leakage traces to do so than a parametric test [32]. Thus in a setting where normality cannot be guaranteed, and it is not known in which moments we expect to see leakage (nor indeed whether the distribution can be fully characterised by its moments), CMI is, to the best of our knowledge, the only leakage test able to provide some guarantees of capturing all possible (univariate) dependencies. (Finding multivariate dependencies is considerably more computationally and data intensive, though some efforts have been made in this direction [42]). However, analysts should be reluctant to interpret the estimated magnitude of the MI beyond the reject/don't reject outcome of the test: the bias of MI estimators and their unknown convergence properties [35] make the actual values estimated unreliable as quantitative measures, particularly as the sample size needed for quality estimates will inevitably be much larger than that needed to reject the null.

Making Additional Assumptions Whilst the previous rephrasal preserves the fully general scope of the detection test, any introduction of additional assumptions inevitably compromises this. However, if the assumptions are reasonable then they can lead towards tests which, though less robust, are more efficient with respect to the sample size required to draw statistically significant conclusions. It is commonplace in the literature to suppose that the non-deterministic part of the trace measurements (the noise) is Gaussian distributed. In the case of 'raw' traces this is typically true.

The further assumption that it is of the same magnitude for all inputs[1] gives rise to the method of ANalysis Of VAriance (ANOVA) for detecting data-dependency. This was proposed by Bhasin *et al.* under the name Normalised Inter-Class Variance (NICV) [4]. If the possible values for X are $\{x_1, \ldots, x_m\}$ the NICV can be understood as the following hypothesis test:

$$H_0 : \mathbb{E}[Y|X = x_1] = \mathbb{E}[Y|X = x_2] = \ldots = \mathbb{E}[Y|X = x_m]$$
$$\text{vs. } H_{alt} : \mathbb{E}[Y|X = x_i] \neq \mathbb{E}[Y|X = x_j] \text{ for some } i, j \in \{1, \ldots, m\}, i \neq j.$$

This can be achieved by computing

$$F = \frac{\frac{1}{m-1} \sum_{i=1}^{m} n_i (\overline{y_{i\cdot}} - \overline{y_{\cdot\cdot}})^2}{\frac{1}{N-m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_{i\cdot}})^2}$$

where $\overline{y_{i\cdot}}$ is the mean of all observations such that $x = x_i$, y_{ij} is the j^{th} observation such that $x = x_i$ (there are n_i in total), and $y_{\cdot\cdot}$ is the mean of all observations for all values of x. Under the null hypothesis the statistic follows an F-distribution with degrees of freedom $(m - 1, N - m)$.

NICV is closely related to the correlation test as proposed by Durvaux *et al.* [15], which also looks at the explanatory power of group means at each point in the trace but does so via Fisher's z-transform of the correlation which has an approximate standard normal distribution under the null hypothesis of zero correlation. Suppose $M(\cdot)$ maps each plaintext/key X to its corresponding average leakage for the trace point to be tested[2], the method from [15] essentially amounts to an hypothesis test of the following form:

$$H_0 : Corr(M(X), Y) = 0 \text{ vs. } H_{alt} : Corr(M(X), Y) \neq 0. \tag{3}$$

Note that this implicitly assumes that $(M(X), Y)$ can be approximated by a bivariate Gaussian distribution, so that zero correlation implies (and is implied by) independence between them.

Since the F-statistic and the correlation need to be estimated across the entire input space, and since (at least, in software implementations) each clock cycle typically depends on only a portion of the state, the usual practice is to take X to be a byte of plaintext/key as opposed to the entire plaintext/key, and to perform the test in turn for each byte separately. Note that neither of these methods are suitable to detect dependencies that emerge after bytes have been effectively 'mixed' (at least, not without computing and testing for those later intermediates explicitly).

An alternative 'non-specific' method which partitions on the full input manages to do so by settling for a single fixed input to produce one of the partitions whilst letting the input vary at random to produce the other. This is known as the 'fixed-vs-random' test and is part of the popular TVLA framework of Goodwill *et al.* [22], employing the t-test for difference-of-means [23] to decide between the following hypotheses:

$$H_0 : \mathbb{E}[Y|X = c] = \mathbb{E}[Y|X \leftarrow R] \text{ vs. } H_{alt} : \mathbb{E}[Y|X = c] \neq \mathbb{E}[Y|X \leftarrow R]. \tag{4}$$

Despite the appeal and widespread adoption of this approach, a drawback is that it inevitably only covers a small part of a very large sample space, especially with respect to the 'fixed' input acquisition. The 'fixed-vs-fixed' variant of the test [15], designed with data efficiency in mind, does indeed have greater power but covers an even tinier fraction of the total sample space: failure to

[1] It can vary by time point as long as it doesn't vary *within* time point for different inputs.

[2] Note that this mapping implicitly incorporates the intermediate function of X on which the leakage depends.

find a leak merely gives an assurance about the indistinguishability of the two chosen fixed inputs (one pair of every possible $2^{2\times B}$, where B is the block size), and not of the data non-dependency of the implementation in general. We discuss this among a range of coverage challenges in Section 6.

Notice that, in stark contrast to the tests we looked at initially, all three of the NICV, the correlation test and the fixed-versus-random DoM can only detect leaks that are present in the first (univariate) moment. Existing publications overcome this deficiency by applying pre-processing of various forms (potentially combining multiple points) to the raw traces, 'forcing' distributional differences into the first moment. Whilst generally effective given 'enough' traces, this work-around undermines the statistical formalism of the tests which (as we shall explain in Section 4) has troubling, typically unacknowledged consequences for an evaluator's ability to meet the goals described in Section 2.

Detecting Leaks Related to Specific Intermediate Values The strategies above do not test intermediate values specifically, but seek to find arbitrary associations between the key or plaintext and the (univariate) distributions at each point in the trace. Assuming the chosen test attains this ideal functionality, such an approach minimises the number of different tests needing to be performed as well as the requirement for detailed knowledge of the implementation. An alternative strategy, which loses these advantages but avoids the computational complexity of the CMI test whilst gaining insight over the non-specific fixed-vs-random test, is to compute specific intermediate values (and relevant combinations thereof, in the case that the device is suspected to leak transitions between intermediate states) and look for associations between *these* and the measured leakages. The TVLA framework [22] proposes, for example, to perform t-tests on partitions constructed around known intermediate *bits*:

$$H_0 : \mathbb{E}[Y|X[b] = 0] = \mathbb{E}[Y|X[b] = 1] \text{ vs. } H_{alt} : \mathbb{E}[Y|X[b] = 0] \neq \mathbb{E}[Y|X[b] = 1], \tag{5}$$

or around known intermediate *bytes*, in a 'one value versus all other values' manner:

$$H_0 : \mathbb{E}[Y|X = c_1] = \mathbb{E}[Y|X = c_2] \text{ vs. } H_{alt} : \mathbb{E}[Y|X = c_1] \neq \mathbb{E}[Y|X = c_2]. \tag{6}$$

The NICV and correlation tests described above can be easily adapted to this 'specific test' setting, reducing the number of different tests needing to be performed in order to get reasonable coverage of the many potentially leaking states. (For example, in the latter case one F-test readily replaces 2^8 separate t-tests, and is more statistically rigorous – although it cannot distinguish between intermediates that are effectively permutations of each other, such as the input and the output of an S-box).

Note that all of these tests are once more only able to capture data-dependencies that exhibit in the first moment of the trace distribution, and that they all depend to a greater or lesser extent on the assumption of normality. The same pre-processing 'tricks' exist to capture higher-order (and multivariate) dependencies as are used for non-specific tests (in the case of correlation this entails applying a corresponding transformation to the power model $M(\cdot)$), with the same degradation of statistical rigour.

In the next three sections we identify a number of ways in which leakage detection, as commonly practised, falls short of the goals listed in Section 2. We discuss existing attempts to overcome these shortcomings, and propose some possible new options, but perhaps most importantly we seek to clarify the limitations of such efforts and the implications for the reliability of leakage detection as an evaluation tool.

4 Shortcoming 1: Difficulty Interpreting Negative Outcomes

Formally, a statistical hypothesis test either rejects the null hypothesis in favour of the alternative, or it 'fails to reject' the null hypothesis. It does not 'prove' nor even 'accept' the null hypothesis. Moreover, it does this with a certain probability of error.

Recall the significance level α (also known as Type I error rate). This is the probability of rejecting the null when it is true, i.e. concluding that there is a leak when there isn't. Because it is chosen as part of the test design it is fully transparent to the evaluator (as long as it is chosen correctly). However, the Type II error (usually denoted β) – the probability of not rejecting the null when the alternative is true, i.e. concluding that there is no leak when there is – is opaque to the evaluator without further effort. If this error is very high (equivalently, we say that the 'statistical power' $1 - \beta$ is low) then the failure of the test to detect an effect really doesn't mean very much at all.

If a test fails to reject the null of 'no leakage' in the context of an evaluation, at the very least a certifier (especially one with the second goal of 'certifying security' in mind) needs to be able to argue that the device was given a *fair trial*. That is, there needs to be some set of criteria to ensure that tests are comparable across targets and measurement set-ups. Otherwise, one device may pass as secure, and another fail to pass, simply because the latter was subjected to a more aggressive testing procedure.

Since all evaluations are inevitably subject to budget and time constraints, it may seem natural to fix the overall effort invested and to make comparisons on that basis. However, a number of factors – differing levels of expertise, quality of equipment, availability of *a priori* information from evaluations of similar targets – undermine the apparent 'fairness' of such an approach. In pursuit of a more objective framework we instead turn to the tools of statistical power analysis. We first explain the general methodology and then show how it can be applied in the case of leakage evaluation.

4.1 Statistical Power Analysis

Since the estimate of a test statistic is itself an observation of a random variable (with a sampling distribution of its own) conclusions drawn from statistical tests are subject to error. The decision to reject a null hypothesis when it is in fact true is called a Type I error (a 'false positive'). In the side-channel setting this corresponds to finding leakage when in fact there is none. An hypothesis test seeks to control this error rate at a **significance level** α. A Type II error is a failure to reject the null when it is in fact false (a 'false negative'), corresponding to failing to find leakage which is in reality present. The Type II error rate of an hypothesis test is denoted β and the **power** of the test is $1 - \beta$, that is, the probability of correctly rejecting a false null in favour of a true alternative. The two errors can be traded-off against one another, and mitigated (but not eliminated) by:

- Increasing the **sample size** N, intuitively resulting in more evidence from which to draw a conclusion.
- Increasing the minimum **effect size** of interest ζ, which in our case implies increasing the magnitude of leakage that one would be willing to dismiss as 'negligible'.
- Choosing a different statistical test that is more efficient with respect to the sample size.

For a given test (i.e. leaving aside the latter option) the techniques of **statistical power analysis** are concerned with the mutually determined relationship between α, $1 - \beta$, ζ and N. Appendix B gives a worked-through example for the simple case of a t-test with equal sample sizes

and population variances σ_1 and σ_2 and arrives at this expression for the minimum (total) sample sample required:

$$N = 2 \cdot \frac{(z_{\alpha/2} + z_{\beta})^2 \cdot ({\sigma_1}^2 + {\sigma_2}^2)}{\zeta^2} \tag{7}$$

where $\zeta = \mu_1 - \mu_2$ is the true difference in means between the two populations. Note that Equation (7) can be straightforwardly rearranged to alternatively compute any of the significance level, effect size or power in terms of the other three quantities. This becomes useful in the event that, for example, the sample size is constrained in practice and the analyst wishes to know the power of the test to detect the effect size of interest, or the minimum effect size that could be detected with a satisfactory power.

Ideally, such an analysis is performed even before data collection as an aid to experimental design; this is known as *a priori* power analysis and can help to ensure (e.g.) the collection of a large enough sample to detect data-dependencies of the expected magnitude with the desired probability of success [32]. If the desired probability of success is not achievable by a given test within a feasible number of traces there is simply no point in performing the test, as failures to find leakage will have no meaningful interpretation. Power analysis can be performed *after* data collection in order to make statements about the power to detect a particular effect size of interest, or the minimum effect size that the test would be able to detect with a certain power. This can be especially useful when it comes to responsibly interpreting the non-rejection of a null hypothesis. However, it is crucial that the effect sizes are chosen independently of the test, based on external criteria, as it has been shown that attempts to estimate 'true' effect sizes from the test data produce circular reasoning. In fact, there is a direct correspondence between the p-value and the power to detect the observed effect, so that 'post hoc power analysis' merely re-expresses the information contained already in the test outcome [24].

This requirement for information *about* the data sample which cannot be estimated *from* the data sample is the main obstacle to statistical power analysis. The choice of effect sizes for the computations can be guided by previous experiments (e.g., in our case, leakage evaluation on a similar device with a similar measurement set up) or (ideally) by some rationale about the practical implications of a given magnitude (e.g. in terms of loss of security). Note that we always *eventually* need some rationale of this latter type: what is ultimately of interest is not just whether we are able to detect effects but whether the effects that we detect are of practical concern. With a large enough sample we will always be able to find 'arbitrarily small' differences; the question then remains, at what threshold do they *become* 'arbitrary'?

Also needed in order to perform statistical power analysis are the population standard deviations of the partitioned samples, which may or may not be the same. These are usually assumed to have been obtained from previous experiments and/or already-published results, which can be especially tricky when approaching a new target for evaluation. Various rules-of-thumb exist, for example, dividing the expected range of values by 5 [10,46]. This reduces the reliance on *a priori* knowledge (at the cost of lower accuracy) but still requires expert input in the form of a meaningful approximation of the range. Whatever the information available, it is generally preferred to choose conservative estimates for the standard deviation (e.g. the largest among those previously observed) as underestimates will lead to overestimates of the power, risking a false sense of security.

It is convenient (and bypasses some of the reliance on prior information) to express effect sizes in standardised form. Cohen's d is defined as the mean difference divided by the pooled standard

deviation of two samples of (univariate) random variables A and B:

$$d = \frac{\overline{a} - \overline{b}}{\sqrt{\frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}}}$$

where $\overline{a}$, $\overline{b}$ are the sample means, s_A^2, s_B^2 are the sample variances and n_A, n_B are the sample sizes. Notice that this is essentially a measure of signal-to-noise ratio (SNR), closely related to (and therefore tracking) the various notions that already appear in the side-channel literature. The most common definition is the variance of the characterised data-dependent part of the leakage (for example, the Hamming weight, or a linear regression model) divided by the residual variance (i.e., the noise). This is not identical to d – which corresponds to a 'signal' that arises from the construction of the test, e.g. the difference between the partitioned means – but for a given leakage characterisation and partition the one can be determined from the other. The formula for the sample size required for the t-test can be expressed in terms of the standardised effect size as follows:

$$N = 4 \cdot \frac{(z_{\alpha/2} + z_\beta)^2}{d^2} \tag{8}$$

Cohen [9] proposed that effects of 0.2 or less should be considered 'small', effects around 0.5 are 'medium', and effects of 0.8 or more are 'large'. Sawilowsky [43] expanded the list to incorporate 'very small' effects of 0.01 or less, and 'very large' and 'huge' effects of over 1.2 or 2.0 respectively. The relative cheapness of sampling leakage traces (and subsequent large sample sizes) compared with studies in other fields (such as medicine, psychology and econometrics), as well as the high security stakes of side-channel analysis, make 'very small' effects of more interest than they typically are in other statistical applications.

This helps to put the analysis on a like-for-like footing for all implementations. But of course acquisitions vary greatly in the standardised effects they are prone to exhibit, so it doesn't remove the need for knowledge of the particulars of the device in order for meaningful interpretation.

4.2 Statistical Power Analysis for Leakage Detection

In order to apply the principles of statistical power analysis to evaluation design we need to decide what makes a test procedure 'fair'. Some options include:

Fixed sample size: The idea of a sample size threshold bears some similarity to that of a fixed overall 'effort', as the acquisition and processing of the trace measurements determines much of the time and computational complexity of the analysis. As a basis for fairness it has an intuitive appeal which bypasses the requirement to understand the technicalities of statistical power analysis. However, the power of the test will vary depending on the variance and the scale of the effects likely to be exhibited by a particular DUT. Applying tests of different power to different DUTs could be considered unfair; moreover, the power remains opaque to the analyst unless an explicit attempt is made to ascertain it, potentially covering up habitual shortfallings in chosen test procedures.

Fixed power: If an expected effect size for the target implementation can be stated, then the test can be designed to achieve a desired power $1 - \beta$ for the given probability of false detection α. However, this essentially amounts to working harder to attack a device that you suspect to be less vulnerable – in effect penalising security! Clearly, if one target requires fewer traces than another to detect with the same probability, this should be interpreted as an indicator of greater vulnerability.

Fixed effect size: Fixing the raw effect size is difficult/ill-advised as trace measurements can be differently scaled depending on the measurement apparatus and pre-processing. However, it is quite possible to fix the standardised effect size (e.g. Cohen's d, as in Section 4.1 above), and derive the required sample size to detect this effect with a power of $1 - \beta$. Of course, the *actual* effect sizes present in each case differ substantially, as does (consequently) the power to detect these. But it could be argued that this is precisely what makes one device more or less secure than another, and therefore that detection success *should* be allowed to depend on this difference.

We take this latter view – namely, that boosting tests against targets with smaller effects in order to achieve equal power is unfair and that tests should instead be designed to be capable of detecting the same (standardised) effect size. Then, if target A is found to be vulnerable while target B is not, at least there is a sound basis for reasoning that B is 'more secure' than A.

Of course, it is important to choose the standardised effect size carefully. It needs to be conservatively small in terms of what one expects to see and what would constitute a concerning risk. Expectations can be gauged by *observing the effect magnitudes* exhibited by a range of previous acquisitions; risk severity can be reasoned about on the basis of *worst-case adversarial resources*. We discuss each of these considerations in turn.

Observed effect sizes It is not straightforward to 'simply' observe magnitudes in existing acquisitions. This is because all differences will be non-zero, regardless of whether they represent actual leakage, and deciding which ones are 'meaningful' essentially corresponds to the task of detection itself. Choosing 'real' effects based on the outcomes of t-tests, and then using the magnitudes of those effects to make claims about 'detectable' effect sizes, amounts to circular reasoning, and depends on the choice of significance criteria. Fortunately the end goal of leakage detection provides us with a natural, slightly more objective, criterion for identifying 'real' effects, via the outcomes of key recovery attacks. That is, if leakage detection is geared towards identifying (without having to perform attacks) points in the trace which are vulnerable to attack, then an effect size which is 'large enough' to be of interest is one that can be successfully exploited.

We take this approach, and perform distance-of-means attacks on all 128 bits of the first round SubBytes output for three AES acquisitions, taken on an ARM board, an 8051 microcontroller and an RFID device. We also compute the sample effects for each of those bits, which enables us to report estimated effect sizes of interest. Two remaining draw-backs are 1) especially for small effects (i.e. those with a low SNR), the sample sizes may not be large enough for the estimates to be precise; and 2) more points may become vulnerable to attack given even larger samples (which is always the case). So smaller effect sizes may be important to evaluators concerned with more powerful adversaries.

Since there are only 256 subkey candidates, among thousands of distance-of-means attacks a proportion will inevitably rank the correct key first purely by chance (approximately $\frac{1}{256}$ of those on irrelevant points, i.e. where there is zero data-dependency). Adapting from [55], we take measures to confirm the stability of an attack outcome before classifying a point as 'interesting'. Our approach involves repeating the attack on 99% of the full sample and retaining only those points where the correct subkey is ranked first in both instances. (Even then the length of the traces makes it likely that some of the retained points will be 'false positives', a problem that affects multiple statistical tests in general, as we discuss in the next section).

Figure 1 shows the raw (top) and standardised (bottom) observed effect sizes (i.e. mean differences associated with an S-box bit) of first round AES traces measured from an ARM board, an 8051 microcontroller and an RFID device respectively. As expected, because of the different scales

of the measurements (arising from different pre-processing, etc), the raw effects are not necessarily useful to compare. The ARM effects range up to about 0.8, while effects on the 8051 and the RFID implementation range up to 3 and 2 respectively. The standardised effects are much more comparable (≈ 0.6 and ≈ 1 for ARM and 8051 respectively; ≈ 0.4 for the RFID, although this is for the second rather than the first S-box as the latter is less 'leaky' in this instance). [3]

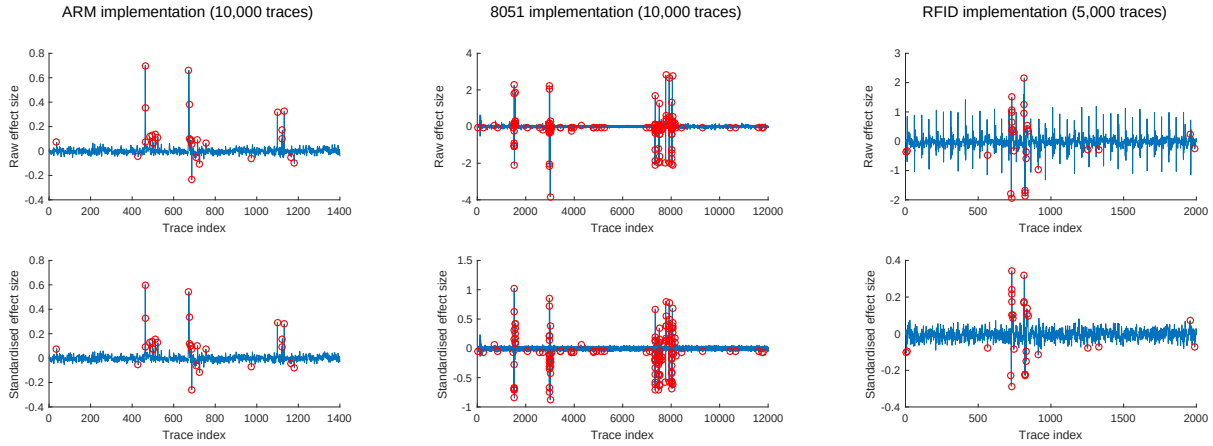

Fig. 1. Difference of means (top) and standardised equivalent (bottom) associated with the first bit of the first S-box of two software AES implementations and the first bit of the second S-box of one hardware implementation (chosen because the first S-box is not very 'leaky' in that particular acquisition). Red circles denote points where a distance-of-means attack achieves stable key recovery.

Table 1 summarises the standardised and raw effect sizes associated with distance-of-means key recoveries over *all* bits of all S-boxes. The smallest standardised effect detected is 0.0413 for the 8051 microcontroller; the ARM and RFID smallest effects are in a similar ballpark. One might therefore choose 0.04 as a minimum effect of interest – although, as we have warned, larger samples will reveal ever-smaller effects, if present. This motivates approaching the question with a 'worst case' adversary in mind, rather than relying only on the data we have access to ourselves.

Implementation	**Proportion**	**Standardised**			**Raw**		
	interesting	**Min**	**Max**	**Median**	**Min**	**Max**	**Median**
ARM	0.0226	0.0444	0.9087	0.1155	0.0388	1.0265	0.1073
8051	0.0150	0.0413	1.4265	0.1670	0.0254	5.3808	0.1469
RFID	0.0049	0.0624	0.3935	0.0933	0.2272	3.4075	0.3836

Table 1. Summary of effect magnitudes associated with stable distance-of-means key recovery attacks.

[3] In a non-specific fixed-versus-random experiment (and even more so in a fixed-versus-fixed setting) the differences depend on more than a single bit so, depending on the value of a given intermediate under the fixed input, can potentially be several times larger (see e.g. [45]) – or they can be smaller (e.g. if the leakage of the fixed intermediate coincides with the average case, such as the (decimal) value 15 in an approximately Hamming weight leakage scenario). It is typically assumed in the non-specific case that, as the input propagates through the algorithm, at least some of the intermediates will correspond to large (efficiently detected) class differences [15].

Adversary resources By considering a 'worst case' adversary with a certain number of traces one can determine the minimum standardised effect size that could be detected with the desired α and β, and use this to set the criteria. Of course, this becomes self-defeating if the envisaged adversary has more resources than a typical evaluator, as the effect will then be too small to detect reliably in an evaluation setting. So the minimum standardised effect size we can actually *test* for is necessarily additionally constrained by feasibility – that is, the number of traces available to an evaluator. If the smallest effect of interest is *not* testable for, at least an evaluator will be able to supply the appropriate caveats in their reported results.

Table 4.2 reports the standardised effect sizes detectable with balanced errors (i.e. power equal to $1-\alpha$) for a range of sample sizes and significance levels. For example, an adversary with 100,000 traces can, with power 0.99999, detect a standardised effect of 0.055 at a significance level of 0.00001 (the significance level implied by the threshold recommended for TVLA). We will revisit (and adjust) these computations in the next section, when we come to consider the implications of the fact that tests are not typically performed singly but multiple times against different points in the same trace. In any case, though, note that we have come around full circle to a solution very like fixing the sample size, albeit in order to derive a standardised effect size. In the absence of any objective criteria (e.g. a mechanism for converting quantifiable security losses into effect sizes), a degree of circularity is inevitable in the attempt to 'bootstrap' a useful solution. (Researchers in other disciplines have also acknowledged this limitation [29]).

α	**Number of traces**				
	10^2	10^3	10^4	10^5	10^6
0.05	0.721	0.228	0.072	0.023	0.007
0.01	0.980	0.310	0.098	0.031	0.010
0.00001	1.736	0.549	0.174	0.055	0.017

Table 2. Some example standardised effect sizes detectable with balanced error rates for different values of α and different sample sizes.

4.3 Discussion

- We have here proposed deriving an effect size of interest from an evaluator's best knowledge about the resources of a typical adversary. However, the sample size required for detection does not necessarily give any indication of the sample size required for an attack [51]. In particular, it is in the interests of coverage (see Section 6 below) to use test strategies that are as generic as possible; by contrast, an attacker is likely to have a specific target in mind, and may have some knowledge of the form of the leakage and/or the capability to choose inputs in a tailored way, all of which could considerably refine the data efficiency of an attack. Freed from the burden of comprehensive testing they might also be more willing to redirect resources towards more computationally costly methods such as mutual information (although this is not typically considered to be more data efficient, at least in most relevant scenarios [39]).
- It is relatively straightforward to derive power/sample size formulae for the t-test under the normality assumption (see Appendix B). However, since statistical power analysis requires characterising the test statistic distributions under both the null and a specific alternative hypothesis, it is non-straightforward in most other cases. Sample size and power derivations for quantities such as mutual information typically rely on methods such as randomised resampling [32], which

can provide useful (and hopefully transferable) insights, but do not permit the type of quick and cheap preliminary computations that incorporate neatly into automated procedures.

– As hinted in Section 3, the preprocessing steps by which higher order and/or joint data dependencies are 'shifted' into first order moments to be exploited by t-tests (and potentially correlation tests and F-tests) cause the assumptions underlying those tests to be considerably violated.(See Appendix A for some indicative analysis of this). This compromises the validity of any formal statistical inferences made, and therefore undermines the types of power analysis-based arguments we are relying on if we want to make a case for security out of the failure of a test to detect leakage. For example, t-tests are understood to be fairly *level*-robust to small deviations from assumptions (that is, Type I error rates are not hugely affected) and, even in the case of large deviations of the types produced by high order preprocessing, the *means* tend towards normality under the Central Limit Theorem (CLT) – a fact that has been used to justify the continued use of the TVLA t-test approach [16]. However, a) the rate of this convergence varies considerably depending on the underlying distributions (by the Berry–Esseen theorem for independent observations this is somewhat determined by the degree of skewness); and b) power and sample size computations become meaningless in both settings, even after convergence (as they still derive from the false assumption that the processed observations are themselves normal, not just the mean), so that nothing can be stated conclusively in the case that such a test *fails* to detect leakage.

5 Shortcoming 2: Unreliability of Positive Outcomes

Statistical hypothesis testing is generally introduced (as it is above) under the implicit assumption that a single null/alternative pair is up for consideration. Unfortunately, the conclusions of the test are no longer formally supported in the case that multiple tests are performed without modification as part of the same experiment. This is because each test has, by design, a probability α of falsely rejecting the null hypothesis when it is in fact true. Hence, the probability of rejecting *at least one* true null hypothesis across all N tests (that is, the *overall* false positive rate as opposed to the *per test* rate) might be as high as $\alpha_{overall} = 1 - (1 - \alpha_{per-test})^N$ if those tests are independent. (Otherwise, the rate will be lower but will depend on the form of the dependencies).

Most evaluation procedures operate on trace measurements acquired during the execution of a code sequence. These can be many thousands of data points in length with no (or only rough) *a priori* knowledge of where a particular leak is likely to manifest. Hence a test for a particular vulnerability is typically repeated multiple times. Ding et al. [13] point out that the detection threshold of 4.5 recommended by the TVLA framework [22], which implies a per-test false positive rate of ≈ 0.00001, corresponds to an overall rate of 0.0068 for 1,000 independent tests, 0.0661 for 10,000 tests, 0.4957 for 100,000 tests, and 0.9987 for 1,000,000. (Note, though, that the serial nature of trace acquisitions mean that the tests are unlikely to be independent in practice, so the true overall rates could be lower).

Thus, unless adjustments are made for trace length, a device producing long traces is more likely to be assessed as vulnerable than an equally secure device producing short traces, or where more *a priori* information is available to help select only the relevant part of the traces. Clearly, this is undesirable in a certification procedure. In Section 5.1 we explain some available options for conducting multiple tests and in Section 5.2 we explore the challenges of applying these in a leakage detection setting.

5.1 Multiplicity Corrections

There are two main approaches to correcting for multiple tests: controlling the *family-wise error rate* (FWER) and controlling the *false discovery rate* (FDR). Both of these were discussed and evaluated in the context of leakage detection by Mather *et al.* [32].

FWER-based methods work by adjusting the per-test significance criteria in such a way that the *overall* rate of Type I errors is no greater than the desired α level. For example:

- The Bonferroni correction [14] which simply derives a per-test significance level by dividing the desired overall significance level by the number of tests m, i.e. $\alpha_{per-test} = \frac{\alpha}{m}$. This controls the FWER for the 'worst case' scenario that the tests are independent, and is conservative otherwise.
- The Šidák correction [58] is able to be slightly less stringent by explicitly *assuming* independence, and that all null hypotheses are false, and setting $\alpha_{per-test} = 1 - (1 - \alpha)^{\frac{1}{m}}$. (These assumptions are not really appropriate in a leakage evaluation setting).
- The Holm procedure [25] gains power over the Bonferroni correction by adjusting the significance levels of each individual test in a 'step up' manner: having ordered the tests according to p-value (smallest to largest), it sets a criteria of $\alpha_i = \frac{\alpha}{m-i+1}$ for the i^{th} test.

It should be clear that any such downward adjustment to the per-test Type I error rates (in order to avoid concluding that there is a leak when there isn't) inevitably increases the rate of Type II errors (the probability of missing a leak which is present). Erring on the "safe side" with respect to the former criterion may not be at all "safe" in terms of the cost to the latter. The relative undesirability of the two error types depends on the goals of the evaluation and must be carefully considered. In particular, the Bonferroni correction may be simplest to implement and to analyse (by substituting $\alpha_{per-test}$ for α in the power analysis formulae) but as the most conservative option it is the most costly in terms of the amount of additional data required to retain adequate power.

FDR-based methods take a slightly different approach which is more relaxed with respect to Type I errors and subsequently less prone to Type II errors. Rather than minimise the probability of *any* false positives they instead seek to bound the proportion of total 'discoveries' (i.e. rejected nulls) which are false positives. Mather *et al.* [32] reasoned that this was well-suited to the task of leakage detection, where individual false positives can be tolerated as long as some of the discovered points are truly 'leaky'. Of course, this may be more or less the case depending on the particular evaluation goal, which should be borne in mind when reporting outcomes. The main FDR-controlling method, and the one that we will consider in the following, is the Benjamini–Hochberg procedure, which (like the Holm correction) operates in a 'step up' manner as follows:

1. For the ordered (small to large) p-values $p_{(1)}, \ldots, p_{(m)}$, find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject the null hypothesis for all tests $i = 1, \ldots, k$.

A recent proposal [13] takes an alternative third way, whereby the decision to collectively reject or not reject a set of null hypotheses is based on the *distribution* of the p-values output by the tests performed separately. Their method appears to be more powerful than tests using the Šidák correction, but the authors did not provide a comparison with tests controlling for the FDR. Moreover, the test is unable to conclude *which* of the null hypotheses are untrue, just that at least one of them is, so does not identify the location of the leakage. It also relies on the assumption that the tests are independent, which we would like to avoid. (As such, we omit this method from further analysis and focus only on the three FWER-based and one FDR-based methods described above).

In addition to the inevitable *loss* of power associated with all of the above adjustments, a substantial obstacle to their use in an evaluation setting is the difficulty of *analysing* (and controlling) the power which, as discussed in Section 4, is essential if we want to draw meaningful and comparable conclusions from test outcomes. In cases where a single per-test significance level $\alpha_{per-test}$ is derived (e.g. Bonferroni and Šidák), this can simply be substituted into the power analysis formulae to gain the per-test power. However, consensus is lacking when it comes to performing equivalent computations for the FDR-controlling procedures deemed more suitable for evaluation (compare, e.g., [5,17,30,38,56]). Moreover, multiple testing scenarios give rise to other notions than per-test power, the computation of which require *a priori* knowledge of the number of true effects and the correlation structure of the tests, which are typically unknown and can only be guessed at. (This is regardless of whether correlation is actively taken into account by the adjustment procedure, which isn't straightforward to achieve and is outside the scope of this current study). Porter [37] presents some of these different notions and describes a way of approximating them by simulating test statistics under certain assumptions about the underlying processes producing them. We discuss these in Section 5.4 after first looking at the impact of multiplicity adjustments in a leakage evaluation setting (Section 5.2) and comparing this with the TVLA's more ad-hoc solution for minimising false positives (Section 5.3).

5.2 Multiplicity Corrections for Leakage Detection

We consider the ARM board implementation of AES (partially) depicted in Figure 1. The top left corner of Figure 2 shows the mean of the power consumption (which has been mean-centred and reduced to one point per clock cycle). It is quite easy to visualise the rounds, so an analyst might choose to run a detection test targeting the output of the first S-box in the first round only (see bottom row of Figure 2). This implies much shorter traces and therefore fewer tests. We are interested in the consequences of this truncation for the error rates of the evaluation. The RHS of Figure 2 depicts the t-statistics when the data are partitioned by the least significant bit (LSB) of the first S-box output. Orange dots denote points that are significant at a 5% level without correction; yellow circles are those which remain significant after a simple Bonferroni correction; purple crosses are those significant under the Benjamini–Hochberg (BH) procedure for controlling the FDR. (The Šidák and Holm corrected tests closely align with the Bonferroni, and have been omitted to avoid cluttering the figure). It is clear that the FDR-controlling procedure retains more of the detected points than the FWER-controlling one. However, some of the points flagged as leakage look intuitively questionable given our *a priori* expectations (in particular, in the top figure the BH-adjusted test concludes that a trace point in a later round depends on the first round S-box).

Green dots (which cover over but always coincide with an orange dot) depict points that are significant at the TVLA-implied level of $\alpha = 0.00001$. In this instance, this set coincides exactly with the 20 points found significant in the truncated traces at an $\alpha_{overall} = 0.05$ level using the Bonferroni correction (one of which loses significance under Bonferroni when the full trace is taken into consideration). In fact, it is reasonable to suppose (though the original paper does not make this explicit) that the TVLA threshold criteria has been chosen with multiple comparisons in mind: for a trace length of 5,000, setting $\alpha_{per-test} = 0.00001$ is equivalent to setting $\alpha_{overall} = 0.05$ and making the Bonferroni correction. This suggests that the stringent TVLA criteria should perhaps be considered an *alternative* to multiplicity corrections, rather than applied in addition to them. Whether or not it is a good alternative is another question: recall that fairness requires being able to analyse and control the error rates from one test to another, which is better achieved by beginning with a more liberal significance level and adjusting according to the actual length of a given trace set.

Recall that the attack-based analysis in Section 4.2 found 30 'exploitable' S-box 1, bit 1 effects in the (truncated) dataset. By comparison, this hypothesis test-based analysis finds 113 when no correction is made, 20 when the FWER-controlling correction is made, and 34 when the FDR-controlling correction is made. There is no objective way of knowing which is closest to the 'true' set of leaky points as the attacks could also include false positives; however, it seems reasonable to suppose that there are at least 20 and possibly as many as 30 effects of size 0.04 or larger.

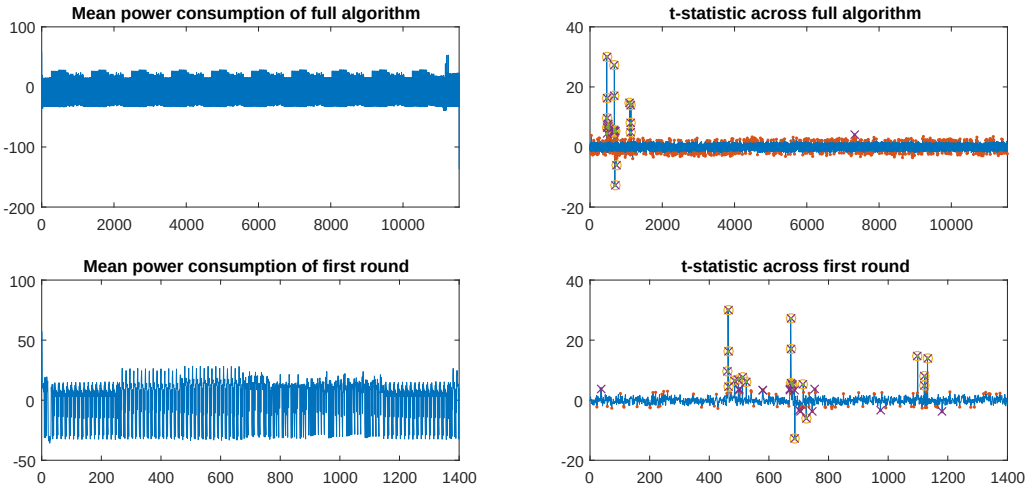

Fig. 2. Mean leakage and t-statistics for the full AES algorithm and the first round. Orange dots denote t-tests that are significant at the 0.05 level without adjustment; yellow circles denote t-tests that are significant at the 0.05 level with Bonferroni correction; purple crosses denote t-tests that are significant at the 0.05 level under the Benjamini–Hochberg procedure for controlling the FDR; green dots denote t-tests that are significant at the 0.00001 level without adjustment.

Figure 3 shows the number of detections as the sample size increases, for both the full and the truncated traces. As well as testing for the LSB of the S-box output, we also test for a random partition to show how the test behaves for a leak which isn't present. The top row relates to a significance level of 0.05, typical in the wider statistical literature. The left panel shows that the number of false detections of 'no leak' in the full traces is about level with the number of true plus false detections of the LSB leak when no adjustment is made for multiplicity. In the truncated traces there is a clearer margin between these two numbers – a difference of 45, which we might be tempted to read as a ball-park indicator of the number of true positives. Either way, the comparison suggests that the majority of the detected leaks are false.

Once adjustments are made (top middle and right panels) the detection rates go down considerably for both the full and the truncated traces and are still increasing as the sample size reaches the maximum available. Meanwhile, 'detections' in the no-leak scenarios are almost entirely eradicated. We can therefore be relatively confident that the discovered leaks are real, but less confident that we have found them all. Indeed, the fact that the FDR-controlling procedure finds considerably more vulnerabilities in the truncated dataset than in the full one, and more than the FWER-controlling procedure finds in either, demonstrates the superior power of that approach but also its sensitivity to the number of tests (for a fixed number of false nulls), which may be undesirable.

The bottom row relates to a significance level of 0.00001, chosen to correspond approximately to the recommended TVLA threshold. We can see from the left panel that this extremely small

choice effectively takes multiple comparisons into account without adjustment, and (in this case) without discernible penalty for the lack of flexibility with respect to trace length. The detection rates for the full and truncated traces now track each other perfectly, and attain the same maximum number of detections (20) observed for an adjusted overall significance level of 0.05. 'Detections' in the no-leak case are again nearly entirely eradicated. When corrections are applied to the already very small significance level (bottom middle and right) we lose around a quarter of the (believed true) detections. It is interesting to note that, under these stricter significance criteria, the FDR-controlling procedure performs comparably to the FWER-controlling one.

At least in our example scenario, the TVLA recommendation (without further adjustment) turns out to be a reasonable choice for controlling error rates, relative to a standard correction procedure. However, the original TVLA paper also proposes another measure for avoiding false positives, which we look at next.

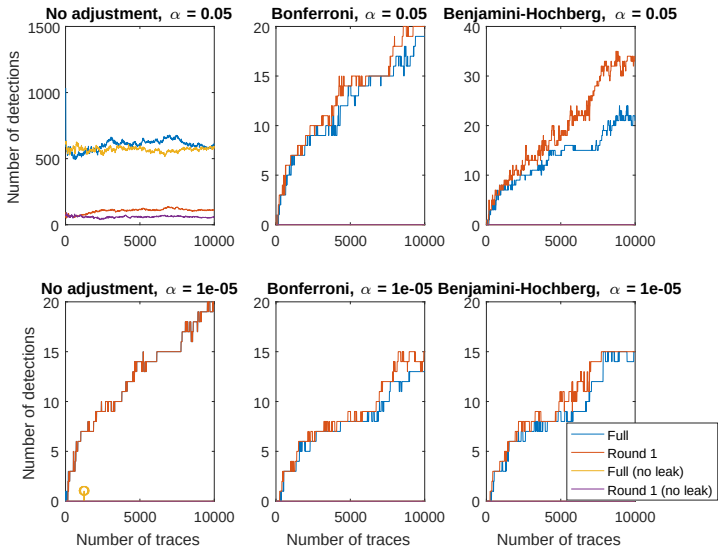

Fig. 3. Number of (false and true) detections made as the sample size increases.

5.3 Experiment Repetition

An often-overlooked recommendation of the original TVLA paper is to repeat the leakage detection on two independent samples (without adjusting the significance criteria), only retaining the points that are detected in both instances. Of course, this supposes perfect alignment between the two acquisitions. But practical challenges aside, this strategy of repetition can be very effective in reducing the risk of false positives.

If the per-test significance level is $\alpha_{per-test}$, then the probability of observing at least one false positive might be as large as $\alpha_{overall} = 1 - (1 - \alpha_{per-test})^N$ (where N is the number of points in a trace) if the tests are independent. The probability of observing at least one false positive in each of a pair of independent experiments is then $(1 - (1 - \alpha_{per-test})^N)^2$, or $(1 - (1 - \alpha_{per-test})^N) \times (1 - (1 - \alpha_{per-test}/2)^N)$ if the direction of the effect is required to match (since this is fixed by the first test and random differences in either direction are equally likely in the repeated step). However,

the probability of observing two false positives *in the same position* (and in the same direction) is $\alpha_{repeat} = 1 - (1 - \frac{\alpha_{per-test}^2}{2})^N$, which grows much slower as N increases.

Figure 4 shows the implications of this in practice. When the per-test significance level is controlled at 0.05, as per popular practice in the statistics literature, the probability of at least one false detection in each experiment is not much reduced relative to that of the single experiment (under a simplifying assumption of independence). However, it takes about 40 times as many traces to falsely observe a leakage at the *same index* with near certainty (although the probability of at least one coinciding detection is over a half once the length of the trace reaches 600). By contrast, under the original TVLA recommendations (which imply $\alpha_{per-test} \approx 0.00001$), the probability of a coinciding detection is close to zero even for traces that are millions of points long. (Only once the number of points is on the order of 10^{10} do coinciding false detections become non-negligibly probable).

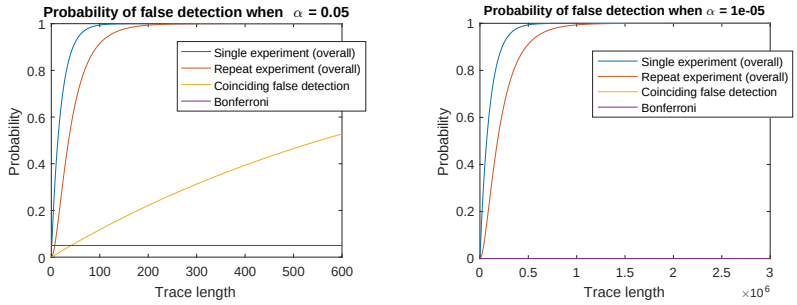

Fig. 4. Overall probability of a false positive as the length of the trace increases, for two different per-test significance levels.

At first glance, this may seem like 'problem solved' but actually there is a substantial trade-off with power that is not immediately obvious. The requirement to repeat the experiment effectively halves the total maximum resources (sample size, computational effort) obtainable to perform the evaluation (that is, because one needs to collect traces and perform the computations twice). On top of that, the stricter criteria imposed by the double-checking procedure of course impacts on true positives too: if the individual power to detect a true leak (of a particular magnitude) is $1 - \beta$ then the probability of detecting it twice is $(1 - \beta)^2 < 1 - \beta$.

The blue and red lines in Figure 5 compare the power for a single experiment with the power for a repeat experiment for two values of α and a fixed effect size as the *total* sample size increases. That is, in the case of the repeat experiment the two separate acquisitions add up to the sample size for a single experiment. The impact on the power for a given sample size is substantial. For a significance level of 0.05, a total sample size sufficient to achieve a power of 80% ($\approx 20,000$) if one test outcome suffices has only 26% probability of identifying a true positive in both of two tests when it is required to confirm the results with the same overall amount of resources. For a significance level of 0.00001 (requiring $\approx 70,000$ in the single experiment case), the probability drops even lower, to around 6%. Over twice as many traces total are needed to achieve the same power when the experiment is repeated. For comparison, we also show the power attained when the Bonferroni correction is applied to the traces (see yellow, green and purple lines).[4] When the overall significance level is large (0.05) the power of the Bonferroni strategy is close to that of the repeated experiment for a short (100 point) trace; for longer traces it drops below it. For the very

[4] The formulae do not readily extend to FDR controlling procedures, but in this context it is anyway more informative to compare with the most conservative of the tested options.

small significance level, the traces have to be much longer (on the order of 100,000,000) before the repeated experiment becomes the more powerful option.

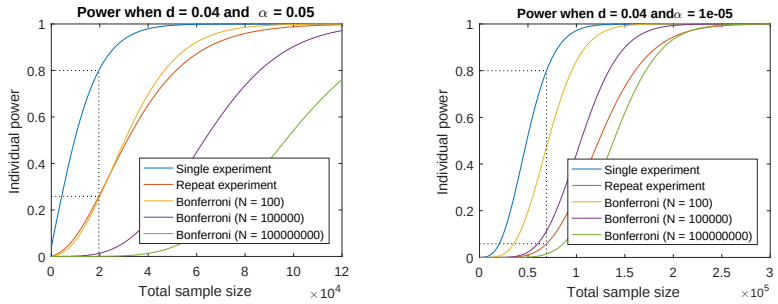

Fig. 5. Power to detect a standardised effect of size 0.04, for two different per-test significance levels, using a fixed total quantity of traces.

So far, we have only considered the ways that measures to avoid the inflation of false positive rates when multiple tests are performed impact on the individual power of each test to detect a real leak (of a certain minimum effect size) if present. In the next section, we explore the idea that multiple testing also gives rise to alternative notions of power that, depending on the goal of an evaluation, may in fact be more relevant.

5.4 Different Notions of Power

Just as multiple tests raise the notion of an 'overall' Type I error which is not equal to the per-test error, so we need to consider the 'overall' Type II error, and what precisely we mean by that. We have seen above that multiplicity corrections reduce the per-test power – the probability of detecting a true effect wherever one exists. Porter [37] describes this as 'individual' power, and contrasts it with the notion of 'r-minimal' power[5] – the probability of detecting at least r true effects. The relevant notion varies depending on the goal of the leakage evaluation: mapping the leakage or certifying the security (i.e. by finding no leaks having tested thoroughly) requires conserving the individual power of each test, while controlling the 1-minimal power may well be sufficient for certifying leakage or finding an attack, when what is important is that *some* leaks are found, not that *all* leaks are found.

In the case that the tests are independent, the probability of detecting *all* true effects (the 'complete power') is the product of the individual powers. (In a leakage scenario, we don't really expect independence so the product is likely to be conservatively low). The r-minimal power is naturally greater than or equal to this quantity. In particular, the 1-minimal power can actually be *higher* in a multiple testing scenario than in a single test – as long as the true number of false positives is greater than 1, each such test represents an additional opportunity to find an effect. So the situation for leakage detection, at least in the case that it is sufficient to simply show the existence of leakage, may not be as disheartening as the impact of multiplicity adjustments on individual power would imply.

We take the scenario observed in Sections 4.2 and 5.2 where there appeared to be around 30 true leakage points in an (AES software implementation) trace of length 12,000, and we assume that the TVLA repeat experiment method is used to guard against false positives. Figure 6 presents the individual power, the power to detect all 30 leaks (under a simplifying independence assumption),

[5] Porter uses the terminology d-minimal; we use r instead of d to avoid confusion with Cohen's d.

and the r-minimal power for $r = 1$ and $r = 10$. For $\alpha = 0.05$ the sample size required to map leakage (i.e. find all leakage points, with a probability of at least 95%) is around 12 times the sample size required to certify vulnerability (i.e. conclude that there is leakage with the same probability). For $\alpha = 0.00001$ the sample sizes are all considerably larger, but the number of traces to map leakage is only around 4 times the number needed to certify leakage.

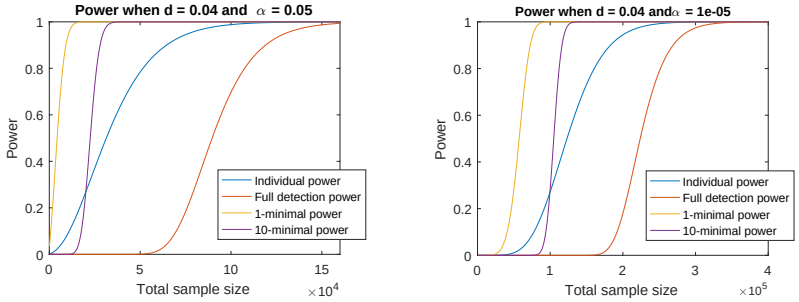

Fig. 6. Different notions of power to detect a standardised effect of size 0.04, for two different per-test significance levels, where the TVLA recommendations are followed. As per Sections 4.2 and 5.2 we suppose that there are 30 'true' effects in a trace set of length 12,000.

Porter suggests a way to approximate the different types of power by simulating large numbers of test statistics under a suitable alternative hypothesis, performing the multiplicity adjustments and simply counting the proportion of instances where 1, r, or all the false nulls are rejected (for the 1-, r-minimal and complete powers) as well as the total proportion of false nulls rejected (for the average individual power) [37]. The considerable limitation of this approach is that it requires a lot of information about the leakage scenario, which we do not typically have in a real evaluation. However, based on the dataset analysed in Sections 4.2 and 5.2 we construct a realistic set of null and alternative hypotheses and show how the different notions of power evolve as the sample size increases.

Suppose the t-statistics corresponding to a trace set of length 1,400 have the same correlation structure as the observed ARM traces, characterised by the covariance matrix Σ. The null hypothesis is that none of the points leak; the alternative is that there are 30 effects of standardised size 0.04, located as per the analysis presented in Figure 1, where $\mathcal{T}$ denotes the set of indices of successful attacks. Under the null hypothesis, for a large enough trace set (which we need anyway to detect such a small effect) the joint distribution of the t-statistics under the alternative hypothesis can be approximated by a multivariate normal with mean $\mu = [\mu_1, \ldots, \mu_{1400}]$ such that $\mu_t = 0.04$ for all $t \in \mathcal{T}$ and $\mu_t = 0$ for all $t \notin \mathcal{T}$, and covariance matrix Σ. By drawing repeatedly from this distribution and noting which of the (individual) tests, with and without correction, reject the null hypothesis and which do not, we can estimate the power and the error rates for tests in this particular scenario.

We performed the analysis for two different significance levels ($\alpha = 0.05$ and $\alpha_{TVLA} = 0.00001$) and six different methods: no correction, Bonferroni, Šidák and Holm corrections to control the FWER, the Benjamini–Hochberg procedure to control the FDR, and the experiment repetition (for a given overall sample size) as per TVLA recommendations. Figure 7 shows what we consider to be the most relevant results, based on 5,000 random draws from the distribution under the alternative hypothesis. (In particular, the three FWER-controlling corrections perform near-identically, and so we only present a single representative, whilst previous analysis has indicated that the TVLA recommendations to use a very small α_{TVLA} and repeat the experiment are best viewed as an

alternative to formal corrections rather than an additional measure). It is clear that the different approaches have substantially different characteristics in practice.

- For a significance level of 0.05, it takes 2.5 times as many traces to achieve 95% average power using the Bonferroni correction as it does using no correction at all. It takes less than twice as many using the BH procedure. The TVLA significance level of 0.00001 is slightly lower than the Bonferroni adjusted level when $\alpha_{overall} = 0.05$, and has fractionally lower power accordingly. Requiring the experiment to be repeated (at the TVLA-recommended significance level) more than doubles the total sample size of that needed to achieve 95% power in a single experiment.
- For a significance level of 0.05, it takes nearly twice as many traces to achieve 95% complete power as it does to achieve 95% average power using no correction. It takes twice as many again using Bonferroni, but only one and half times as many using the BH procedure. The TVLA significance criteria again has a complete power slightly below the Bonferroni correction; with the repetition step the complete power remains negligibly close to zero for traces of length below 180,000.
- For a significance level of 0.05 it takes just over 3,500 traces to achieve 95% 1-minimal power using no correction. (This is below the presented range and was computed in a separate experiment). Interestingly, the two correction procedures closely coincide for this type of power, each requiring about 9 times as many traces as the uncorrected tests to achieve 95% 1-minimal power. The repetition step requires over twice as many traces as the unadjusted test at the TVLA-recommended significance level.
- When no correction is used with a significance level of 0.05 (the blue line) there are false positives throughout the tested range, as we would expect. (We anticipate on average one false positive in every 20 tests). By contrast, with a TVLA-inspired significance level of 0.00001, the rate of false positives stays close to zero (and naturally stays as such when the experiment is required to be repeated). Bonferroni controls overall false positives at the α level, by design, but the BH procedure allows some. The rate increases as the sample size increases and seems to roughly stabilise at about 0.75 for 50,000 or more traces.
- The false discovery rate with no corrections is again close to zero under the TVLA criteria, but high for a significance level of 0.05, decreasing as the sample size increases and seeming to stabilise at around 0.75 for sample sizes of 25,000 or more. The BH procedure, as we would hope, successfully controls the false discovery rate at the α level. The Bonferroni correction, which is stricter about avoiding false positives altogether, has an even lower false discovery rate.

We repeated the experiment with a larger standardised effect (0.2) and observed very consistent outcomes, with the required sample sizes reduced to 25 ($= \left(\frac{0.2}{0.04}\right)^2$) times smaller across the board.

We also repeated the experiment assuming independence between the tests, and found that it made very little difference to either error rate. This is *not* to say that *taking the dependence structure into account in the tests themselves* would not improve the performance of the tests, but it does imply that (at least in this instance) a power analysis which assumes independence need not give a misleading account of the capabilities of the chosen tests.

5.5 Discussion

- We have seen that the strict TVLA threshold already corresponds loosely to a more liberal significance criterion coupled with a (conservative) multiplicity correction, so that it does not seem advisable to apply a correction *on top of* the TVLA recommendations. That said, for realistic acquisition lengths, even conservative multiplicity adjustments retain more power than the recommendation to repeat a test, so that actually, unless an evaluator is handling traces of

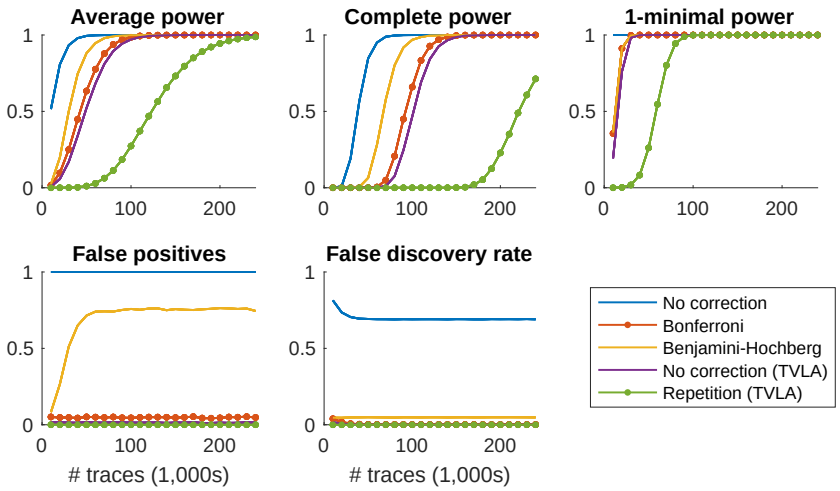

Fig. 7. Different types of power and error to detect 30 true effects of size 0.04 in a trace set of length 1,400, as sample size increases, for an overall significance level of $\alpha = 0.05$. (Based on 5,000 random draws from the multivariate test statistic distribution under the alternative hypothesis).

length on the order of 100,000,000, further corrections remain preferable to dividing resources between two independent experiments.

– More sophisticated methods for controlling overall errors are harder to analyse with respect to power and sample size, at least without considerable *a priori* knowledge of the joint distribution of the trace measurements. This prompts the suggestion that, for the same data complexity implied by test repetition, a two-stage exploratory/confirmatory analysis could instead be performed. The first acquisition could be used to learn about the covariance structure of the traces and the possible locations, sources and nature of leaks. This information could then be used to perform a pared-down confirmatory analysis, in which fewer and more carefully-formulated hypotheses are tested (so reducing the inflation of Type I errors) and insights about the data are used to make more tailored adjustments and to analyse them accordingly. We leave this as an avenue for further work.

6 Shortcoming 3: Impossibility of Achieving Exhaustive Coverage

As discussed previously, ideally an evaluator would like to rule out *any* possible sensitive dependency – of all distributional forms, for all points and tuples of points jointly, via all target functions and intermediate states – before judging a target device secure. In practice, no fully arbitrary single test has been proposed to this end. It is therefore important to understand the limits of each individual test, the best way of combining tests and organising experiments in order to cover as many eventualities as possible, and the limitations that remain even after such a best effort has been made. (It is trivial to note that only vulnerabilities that are tested for have a chance of being found).

6.1 Code Coverage

Coverage is a term from code testing referring (loosely speaking) to the extent to which everything that *could* be tested *has* been tested [33]. Typical metrics in this setting include code coverage (have all lines of code been touched by the test procedure?), function coverage (has each function been

reached?), and branch coverage (have all branches been executed?) [1]. In a hardware setting one might alternatively (or additionally) test for toggle coverage (have all binary nodes in the circuit been switched?) [54]. The appropriate choice of coverage metric can also depend on whether one assumes white- or black-box testing. The given examples all stem from white-box testing as they evidently assume access to the code. In black-box testing, lacking knowledge of and access to the source code, coverage tends to be defined in functional terms.

6.2 Side-Channel Coverage

In the context of side-channel evaluation there are different concepts of coverage that we might consider:

- Have all possible intermediates been tested? (Including via non-specific tests that aim to cover a class of intermediates all at once).
- Have all possible leakage forms been taken into account? For example, some circuits might leak in function of the intermediate values; some in function of the *transitions between* certain intermediate values (with the combinations not necessarily known *a priori*); some in combination of both. Some tests only capture differences in the means, while in reality the leakage might be present in higher order moments or best detected by comparing distributions.
- Have all possible locations in the trace been tested (with each intermediate and leakage form in mind)? This includes not just univariate points but tuples of trace points in the case where higher order leakage of protected intermediates is of concern. (In the case where a DUT has a claimed order of security, it may not be required to test for effects above that order).
- What proportion of the input space has been sampled? I.e. a single fixed-versus-random test might give very non-representative results. Or some keys might be more or less leaky than others, so the typical DPA-inspired approach which assumes a fixed key might be misleading. Moreover, depending on the attacker's capabilities, chosen combinations of inputs might produce more pronounced (and therefore more easily exploited) leakages for a given fixed sample size. With a total possible input space of, e.g. (in the case of AES-128) $2^{128} \times 2^{128} = 2^{256}$ (key, plaintext) pairs, and all the possible distributions on those pairs, it is unavoidable that one can only test a tiny fraction. It is important to be able to articulate the assumptions under which the test outcomes can be supposed to generalise to the wider population. (E.g., that each byte of the total state leaks similarly independently of its position, and/or that the leakage of pairs after key mixing has happened depends only on the output of the mixing function (the XOR between them in the case of AES; see the 'Equal Images under different Subkeys (EIS)' assumption [44])).
- Statistical power analysis as explained in Section 4.1 can itself be thought of as a matter of coverage: has the population been adequately *sampled* to detect the types of effects that are interesting if present?
- Have all possible side-channels been tested?! With most of the literature typically focused on power (and sometimes EM radiation [21,40]) it is easy to forget that other physically observable characteristics (timing [26], temperature [6], light [19,48] and sound [2,47] emissions) also exist and have been shown to be vulnerable. Narrow focus on particular channels risks not only overlooking other problems but maybe even creating them if the corrective measures taken lead to unintended side-effects in untested spaces (for example, asynchronous logic has been found to flatten out power leakage but at the same time to increase EM exploitability [20]).

The first three of these considerations imply huge numbers of different tests as part of the same evaluation. This further exacerbates the multiplicity problems addressed in Section 5.1. Where

corrections *have* been made they have typically related to a single test as performed on multiple trace points. However, it is also the case that performing lots of *different* tests (i.e. with different hypothesis pairs) on those same trace points also inflates the overall probability of false positives.

The priorities of coverage also incentivise the use of more 'comprehensive' methodologies such as CMI. An added advantage of limiting the number of tests needing to be performed is that it helps to mitigate for the multiplicity problem (i.e. it reduces the risk of false positives). But a downside is an increased difficulty of interpreting negative results, as CMI presents considerable challenges in terms of statistical formalism relative to the far simpler t-test (power analysis can only be achieved experimentally as far as we are aware [32]).

Another downside of CMI and other non-specific tests is that they do not provide any ultimate indication of exploitability. For example, Diehl et al. [12] observe an implementation which fails against a fixed-versus-random t-test in such a way that is demonstrably *not* revealing of any sensitive information.

Example of Inadequate Coverage Our example is based on a 'toggle count' power model, derived by counting the number of bit flips in a hardware implementation of the AES SubBytes operation. It has been used in the literature before [31], and is a good representative of potentially highly-nonlinear functions which nonetheless exhibit some amount of first order leakage [32].

We simulate a leakage scenario in which the twice-masked output of AES SubBytes leaks in parallel with the two masks, with all intermediates taking this functional form. We model the noise as Gaussian such that the signal-to-noise ratio relative to the total exploitable variance arising from all three intermediates is 10. (This is high in order to keep the experimental effort within reasonable bounds, but it is fundamental to statistical power analysis (see Section 4.1) that t-test outcomes scale in a well-understood manner as the standardised effect size changes). Figure 8 presents the detection rates of tests targeting the first and second bits of the S-box output as the order of the test and the number of traces increases.

The top left panel on the left hand side confirms our expectation that the first two moments do not leak in either case. The top right shows that bit 2 is detectable within a million traces, but that the most effective moment to target is the 6th one, rather than the 3rd one as we might expect. Bit 1 is not detectable within this range, either targeting the 3rd or any other moment (see also the bottom left panel). However, increasing the sample size ten-fold confirms that tests targeting the higher moments (especially the 8th one) are on an upward trajectory even if the detection rate remains low within the tested region.

This experiment illustrates the sensitivity of the testing procedure to the choice of target and the configuration of the test. It is plain that the S-box output under this leakage scenario *is* vulnerable, but the vulnerability could easily be missed if only the first bit were tested. Moreover, the 'most leaky' moment is not in this case the one we would expect given the known masking order, implying that leakage might be missed if an evaluator stopped after testing at the order of one greater than the number of masks, or at least that more data and work would be required than for a fortuitously chosen higher order test.

6.3 Discussion

The challenge of achieving coverage highlights the mutually detrimental impact of gains along one dimension of evaluation, due to the inflation of Type I errors when so many tests are performed, the degradation of power when multiplicity corrections are subsequently applied, and the increased difficulty of statistical formalism. Some types of leakage (e.g. higher order data dependencies) require larger datasets to detect than others, while jointly leaking tuples can only be searched with

substantial increase in computational effort (exponential, as the size of the tuple increases). This again implies the need for a revised approach.

7 Leakage Detection as a Multivariate Problem

Up until now we have been treating the tests against different points in a trace as separate and (for the most part) independent. However, methods exist to test a single multivariate null hypothesis via a single test statistic that takes into account the dependency structure of the sample data. Recent work by Bronchain et al. [34] proposes the use of Hotelling's T-squared test to decide whether or not to reject the hypothesis that none of the points in a trace depend on the sensitive data versus the hypothesis that at least one of them does.

The T^2-test is a multivariate extension of Student's t-test which compares *vectors* of means between samples of joint (i.e. vector) random variables $\mathbf{A}$ and $\mathbf{B}$.

$$T^2 = \frac{n_A n_B}{n_A + n_B}(\overline{\mathbf{a}} - \overline{\mathbf{b}})' \hat{\mathbf{\Sigma}}^{-1} (\overline{\mathbf{a}} - \overline{\mathbf{b}})$$
$$\frac{n_A + n_B - p - 1}{(n_A + n_B - 2)p} T^2 \sim F_{p, n_A + n_B - p - 1},$$

where p is the dimension of the cluster, n_A and n_B are the sizes of the samples from the two distributions, $\hat{\mathbf{\Sigma}}$ is the pooled covariance matrix estimate, and F_{p,n_A+n_B-p-1} is the CDF of the F-distribution with degrees of freedom $(p, n_A + n_B - p - 1)$. In the univariate case, T^2 is the square of the t statistic and the test is equivalent.

It is useful to introduce the Mahalanobis distance – the multivariate extension of Cohen's d, defined as $D = \sqrt{\mathbf{d}\hat{\mathbf{\Sigma}}^{-1}\mathbf{d}'}$, where $\mathbf{d} = \overline{\mathbf{A}} - \overline{\mathbf{B}}$ is the vector of mean differences.

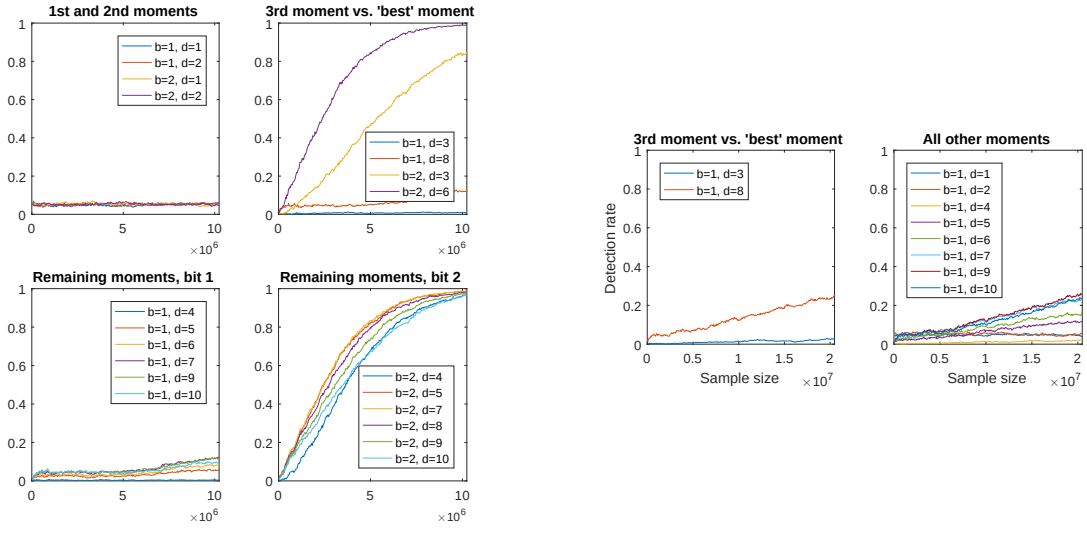

Fig. 8. Left: Comparison between t-tests against the first and second bits of the masked S-box output leaking in parallel with the two uniform random masks according to the 'toggle count' leakage model of Mangard *et al.* [31]. Right: Evolution of the bit 1 t-tests as the sample size increases further.

For the balanced case $n_A = n_B = \frac{n}{2}$ the power of the T^2 test to detect an effect of size $D = D_{alt} > 0$ (relative to a null distribution of $\mathbf{d} = \mathbf{0} \Rightarrow \mathbf{D} = \mathbf{0}$) can be computed as follows:

$$1 - \beta = 1 - F_{p,n-p-1;\lambda}\left(\frac{n-p-1}{(n-2)p} \times \frac{(n-2)p}{n-p-1} F^{-1}_{p,n-1-p}(1-\alpha)\right) \tag{9}$$

where $\lambda = \frac{n}{4} D^2_{alt}$ is the non-centrality parameter of the F distribution under the specific alternative hypothesis. (See Appendix C for details).

Note that the distributions involved in this expression depend on n, so that it is not possible to rearrange it into a neat analytical formula for the sample size in function of the power and effect size. However, sample sizes for scenarios of interest can be obtained numerically by computing the power for increasing values of n and identifying the threshold at which it attains the desired level. When $p = 1$ this equates to exact expression for the power of the t-test, although as we have seen above, in the univariate case it is convenient to take advantage of the normal approximation which enables computing the sample size directly.

Hotelling's T^2 as a tool for leakage detection can be seen to correspond far more directly than the t-test with goals 1 and 2, and with the prospect of bypassing all the problems presented by multiple comparisons. We note, though, that it is *not* a useful solution for either of goals 3 or 4, as it does not conclude on the location of the leakage without further analysis.

However, even for goals 1 and 2 the strategy is not without its limitations. Bronchain et al. [34] find that:

- When leaky points are *dense* in the traces, a multivariate approach detects with fewer traces than the t-test, and this advantage grows with the length of the traces. However, as leakage becomes sparser (for example, in protected implementations, or in the case where specific intermediate functions are targeted as opposed to fixed-versus-random or fixed-versus-fixed leakages) the multivariate approaches lose power until the t-test becomes the more efficient option.
- Statistical power analysis of the tests (in order to provide guarantees about the outcomes of an evaluation) requires a priori knowledge of the density, as well as the effect sizes and covariance structure.
- When leakage points are independent, the test statistic for the multivariate approach reduces to the sum of the squared individual t-statistics, simplifying the analysis. However, when they are not independent, the t-test is overly conservative w.r.t. false positives (thereby increasing the data complexity), the simplified multivariate approach leads to inflated false positives, and the proper Hotelling's test often cannot be implemented due to the non-invertibility of the high dimensional covariance matrices.

The authors suggest to reduce the sparsity as much as possible by dimensionality reduction such as peak extraction, and (if computationally feasible) to break up the traces into manageable chunks to be tested with a series of T^2 tests adjusting $\alpha_{per-test}$ accordingly (which is more conservative than a single T^2 test would be, but still less so than a higher-dimensional series of t-tests with a similar but more punishing adjustment).

We propose a methodology to build on this approach by attempting to *cluster* trace points into 'similarly leaking' groups (instead of equal sized blocks of adjacent points) before applying the T^2 tests. Our rationale is that implementations (particularly in software) comprise sequences of related operations, with values often recurring as the inputs or outputs to several instructions. It is therefore natural to suppose that power measurements at different points in a trace will have shared characteristics, such as proportional data-dependent leakage. Depending on the quality (i.e. the within group similarity and between group dissimilarity) of the cluster arrangement, it might be

expected that one can, in this way, (a) reduce the number of (T^2) tests to be performed relative to the t-test approach, thereby reducing the costs entailed by multiplicity corrections; and (b) minimise the dependency between the tests, so that simple corrections suffice and statistical power analysis can be performed.

In the remainder of this section, we introduce the methodology, present some experimental results, and reason about the power (and the challenges of assessing the power) of our approach relative to that of the standard univariate approach with multiplicity adjustments.

7.1 Detecting Leaky Clusters

Clustering is an unsupervised machine learning process that seeks to group related variables in multi-dimensional datasets. Imagine, for example, that we want to find points in a trace that are strongly cross-correlated with each other. We could set a correlation threshold above which points are considered 'related' – but this quickly gets more complicated than it sounds: e.g. what happens if two points that are not sufficiently correlated with each other are both correlated with a third point? This is essentially a hierarchical clustering problem; fortunately, algorithms exist to (heuristically) solve these. Agglomerative clustering incrementally links close singletons and clusters until all are grouped together; divisive clustering operates in the other direction, beginning with the whole group and incrementally dividing until all are separated. A decision is then made – based either on the desired number of clusters, or on some desired characteristic or quality metric – as to which level in the resulting tree (called a 'dendrogram') to treat the groups as distinct.

Unfortunately, clustering may be systematic, but it's far from objective. The resulting arrangement is highly dependent on user-specified parameters such as the linkage criteria and the nature and level of the threshold. The best parameters, and the outcomes they produce, differ from dataset to dataset, and there is no guarantee of arriving at 'neat' clusterings (rather than, e.g., a couple of really big groups and a large number of singletons). Thus, whilst we believe that a clustering approach is an informative avenue to explore, our theoretical analysis necessarily relies on simplifying assumptions about the form of the clusters, while our experimental results should be taken as indicative rather than conclusive.

Theoretical Performance We return to our running example scenario of the first round of an AES software implementation with 1,400 points-long traces, of which 30 points leak sensitive information. Note that this is much less dense than the fixed-versus-fixed leakage scenarios which were the primary focus of [34]. There is no a priori way of 'knowing' the features of the cluster arrangement necessary for performing statistical power analysis – the size and number of the found groups, the density of leakage within the groups, and the effect sizes and covariance structures – all of which could be envisaged in a number of ways.

For the purposes of this analysis we characterise the effects of interest via the Mahalanobis distance D, thus avoiding the need to separately specify the covariance structure and density of the discovered clusters and producing a standardised measure comparable to the d values used for the univariate analysis. A downside of this approach is that it defies direct comparison with the results of [34], which fixed the individual SNR of the points and varied the covariances and densities. However, the covariance scenarios were devised under the assumption that the groups for multivariate analysis were comprised of contiguous blocks of traces, which does not apply in the case of clustered data. Many more assumptions would be required in order to adapt the assumptions of [34] to our purposes; the use of a D value is easier to justify and more naturally generalises. We suppose that the number of leaks to be detected approximately scales relative to the number of

clusters, so that 30 of the 1,400 individual points leak, 3 of the 140 size 10 clusters and 2 of the 70 size 20 clusters.

Figure 9 shows the individual, complete and 1-minimal power of T^2 tests against trace point clusters with the *same multivariate effect size* $D = 0.04$, using the Bonferroni correction to control the overall false positive rate.[6] As the cluster size increases while D stays the same, the power to detect a leaky cluster goes down, in spite of the decreasing cost of multiplicity adjustments when fewer tests are performed. The concentration of false nulls in fewer tests means that the complete power can actually be higher for the cluster-based data than the raw traces, but for the same reason the 1-minimal power is considerably higher in the latter case. The individual, complete and 1-minimal power coincide in the case of a single Hotelling's T^2 test against the full trace set, but this power is very low due to the substantially increased dimension to effect size ratio.

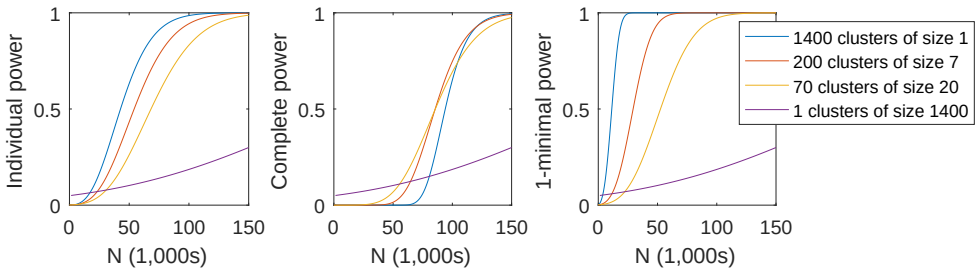

Fig. 9. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the clusters varies; effect size is fixed at $D = 0.04$.

Whether or not it makes sense to compare different sized clusters holding the effect size fixed is open to question. If the quality of the clustering is good, the addition of more leaky points potentially contributes to the multivariate effect size, but conversely the non-independence of the grouped points potentially detracts from it (as you are no longer adding 'new' information, per se). See Appendix D for further discussion, including an analysis of the covariance structures envisaged in [34] (which however, we note, do not naturally correspond to the types of relationships we would expect to see in clustered data).

Experimental Results We now test our idea against the real ARM traces, specifying that we want the 1,400 trace points relating to the first round of an AES implementation to be grouped into 200 clusters (Matlab's hierarchical clustering procedure either takes the desired number of clusters as input or a sensitivity threshold). We choose the correlation for our distance distance (as motivated by our original intuition) and average linkage (as seeming to give the best results). Table 7.1 summarises the cluster sizes produced by these parameters. As expected, the resulting clusters are of different sizes, meaning the true power of the Hotelling's tests will vary.

Figure 10 depicts the result of testing the individual points using Welch's t-test (LHS) and the clusters using Hotelling's T^2-test (RHS) for leakage of the first bit of the first S-box output. The dotted line on the LHS panel shows the threshold for the tests to be significant at a Bonferroni-corrected overall level of $\alpha_{overall} = 0.05$. It is the same for all trace points as the tests are all of the same dimensionality. The red circles depict the 11 trace points that are clustered together with

[6] Whilst it would be preferable to to extend the whole of Figure 7 to the T^2 test, this would require simulating draws from a multivariate F-distribution, which is much harder than the multivariate normal approximations made previously.

Mean	Min	25th percentile	Median	75th percentile	99th percentile	Max	# Singletons	# Large	% Large
7.0	1.0	4.0	6.0	10.0	19.0	21.0	10.0	2.0	2.9

Table 3. Summary of cluster structure (ARM data, 200 clusters).

the largest peak. Only five of these are above the threshold, implying that the cluster contains a mixture of leaky and non-leaky points in spite of the within-group similarity aimed for by the cluster procedure. Four of the points below the line are adjacent to the significant t-statistics, suggesting that the clustering has captured some serial relationship as we might expect. Two are completely separated, hinting towards the somewhat unreliable nature of clustering methods, which are highly sensitive to the choice of parameters and component processes and do not necessarily produce results that 'make sense'. (Our experiments were not able to substantially improve upon the arrangement we have chosen to report, while many attempts produced less coherent arrangements).

The threshold on the RHS panel decreases as the size of the cluster increases (cluster labels have been allocated in ascending size order). The cluster with the largest T^2 peak is the one that contains the individual point with the largest t value (though we do not advise attaching too much meaning to the magnitude of a test statistic). But 9 other clusters also contain evidence of leakage of the targeted bit, **implicating all the individual points in each cluster**, since the T^2 test does not produce separate conclusions for each component point. The 'leaky' clusters are of varying sizes.

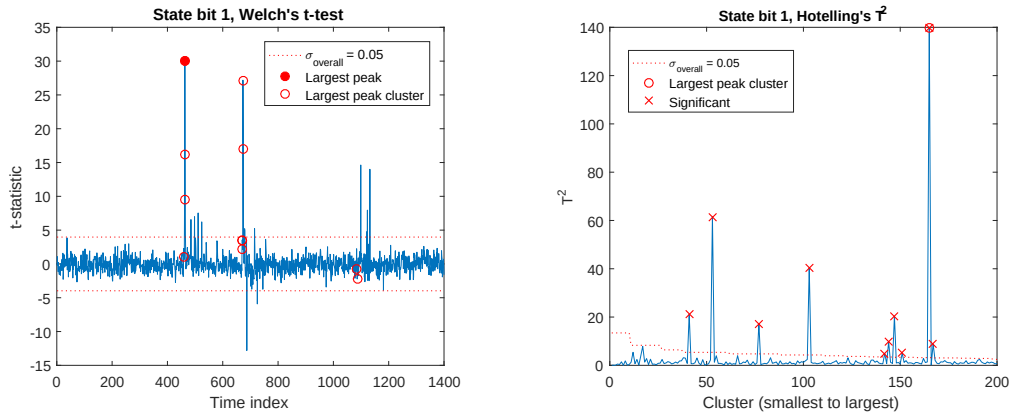

Fig. 10. Welch's t and Hotelling's T^2 tests for leakage of the first bit of the first S-box output, using the Bonferroni correction to control the overall significance level at $\alpha_{overall} = 0.05$.

We repeated the detection procedure for all 128 bits of the state after the first round SubBytes. Figure 11 summarises the distribution of the leaky intermediates across the 200 clusters. All 16 bytes are detected (via one or more bit) in at least 20 clusters, one in as many as 41. Over 120 clusters are associated with at least one of the 16 bytes; most of them with more than one (two clusters produce significant tests for as many as 11 of the 16 bytes). This suggests that our chosen clustering method is limited in its success at concentrating the leakage into a small number of clusters, and also at separating *different* intermediates into different clusters. It is quite possible that other clustering methods would produce more consistent and clear-cut results. However, the process of achieving an improved cluster arrangement becomes increasingly subjective, with the 'optimal' approach likely to vary substantially from one scenario to another.

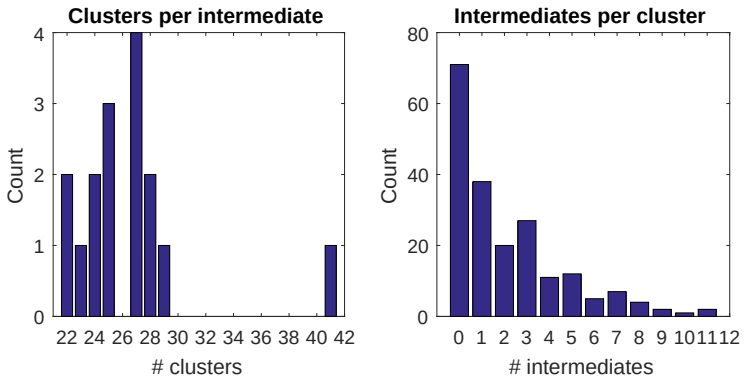

Fig. 11. Number of clusters associated with each intermediate (LHS) and number of intermediates associated with each cluster (RHS) over all first round S-box output bytes.

Another question we can ask is the extent to which the univariate and the multivariate approaches discover the same leaks. Table 4 cross-tabulates the points implicated by the t-test and those implicated by the T^2 test (over all 128 bits of the state after the first round SubBytes). The latter flags whole clusters as 'leaky' without any inbuilt facility to single out the individual points responsible for the leakage. Hence *all* the points in a leaky cluster are considered individually vulnerable, resulting in a larger number of individual detections (14,566) than those discovered by the t-test (3,894). If the clustering was more successful at separating leaky from non-leaky points then this excess of discoveries would be reduced. Meanwhile, 297 of the trace points implicated by the t-test are not detected via the T^2-test, suggesting perhaps that their influence is 'diluted' by their association with non-leaky points. (That is, the Mahalanobis distance for the cluster is smaller than the Cohen's d effect size of some of the points within it). This implies that the cluster-based approach cannot be relied upon to achieve the same coverage as the univariate approach, even while producing a number of *redundant* discoveries via the flagging of whole clusters.

		T^2-test		
		No detection	Detection	Total
t-test	No detection	164,337	10,969	175,306
	Detection	297	3,597	3,894
	Total	164,634	14,566	

Table 4. Cross-tabulation of the leaks identified via the T^2-test (taking all points within a leaky cluster to leak all detected intermediates) and those identified via the t-test.

We conclude that the appeal of the clustering approach – namely, the possibility to reduce the number of tests to perform whilst simplifying the assumptions required for multiplicity corrections – is, on the whole, outweighed by the difficulty of arriving at meaningful assumptions for the purpose of *a priori* power analysis and the extreme sensitivity to parameter and algorithm choice of the method in practice. That said, if more reliable clustering methods could be found, able to consistently separate leaky from non-leaky points (without supervision), then perhaps it would be an interesting avenue for further exploration.

8 Implications and Recommendations

We have shown that leakage detection tests, as typically applied, are limited in their capability to conclusively answer the questions posed by an evaluator. It has been a recurring theme that measures to help resolve one shortcoming typically serve to exacerbate another. Careful rigour is needed in order to reason convincingly that any of the identified goals have been met; some are more challenging to fulfil than others. We summarise the particular challenges and priorities for each below.

8.1 Implications and Recommendations for Certifying Vulnerability

This is the easiest goal to achieve (which is not to say that it is easy). Most important is that the test design provides assurances against *false positives* – the key challenge being that, for a given per-test rate of false positives, the *overall* rate can get very large as the length of the evaluated traces increases.

The TVLA recommendations (a very low implicit per-test false positive rate, plus the requirement to repeat the full test on a second independent sample) are very effective at minimising the Type I error rate, but very costly in terms of the Type II error rate. Methods to control the false discovery rate are unsuitable, as even just one false positive would compromise the goal of certifying vulnerability. Clustering the data prior to detection would help reduce the number of tests and increase the independence between them *in the ideal case that the resulting clusters are well-separated*, but this requirement is difficult to achieve. Our best recommendation is therefore to perform individual tests whilst controlling the family-wise error rate – not ideal, as the most popular (and easy to analyse) methods (e.g. the Bonferroni or Šidák corrections) are known to be over-conservative.

Fortunately, comprehensive coverage is not required for certifying vulnerability: judicious focus on likely targets is sufficient in the case that at least one of them is truly vulnerable. In the event that none of the likely targets evidence leakage, the task may shift towards certifying security, at which point more effort will be required, as we describe next.

8.2 Implications and Recommendations for Certifying Security

Certifying security is considerably more challenging as, in addition to protecting against false positives, the analyst must be able to reason convincingly that the non-detection of a vulnerability indicates that no vulnerability is present. From an error controlling perspective this means paying careful attention to the *1-minimal power* of a test, using the tools of statistical power analysis. It is necessary to quantify a minimum effect size of interest: since it is not known how to translate effect sizes into security losses, we recommend choosing a 'very small' (according to Sawilowsky [43]) standardised effect of 0.01. Even then, it is possible that an adversary who is better resourced, more strategic, and/or 'luckier' than the evaluator would be able to exploit a smaller effect.

For simple t-test based evaluations we have shown how to compute the 1-minimal power to detect a'very small' effect under particular assumptions about the number of leaks and the relationships between the tests. If the tests fail to find leakage then such an analysis can be used to argue that it was not simply down to the inadequacy of the method or sample size. *However* there are many caveats to this as a 'solution' to the goal of certifying security:

- Such *a priori* knowledge about the dependency structure and number of expected leaks (if leaky) is hard to obtain or test, and the analysis is sensitive to the correctness of these assumptions so that the certification is highly provisional at best.

– Pre-processing traces in order to perform higher-order detection via "t-tests" typically causes the distributional assumptions on which statistical power analysis depend to become invalid, and the conclusions untrue, making it impossible to 'certify' higher-order security (by known means).
– The simple formulae enabling easy analysis of t-tests do not exist for other evaluation methods such as those based on MI, or for carefully targeted attack strategies which (as hinted above) may prove more powerful than detection for a given amount of data.
– In addition to statistical rigour, coverage is crucial: arguing that everything possible has been done to find leakage requires testing for a comprehensive range of possible targets using all appropriate methodologies. *This is not feasible in practice*, therefore it is essential to report what has been covered and clarify what hasn't – as thoroughly as possible – as part of the conclusions of any evaluation.

In short, truly certifying security is not an attainable goal in practice. The best that can be achieved is to provide a sound theoretical basis for the design of the statistical tests (which should be made demonstrably consistent from one evaluation to another), and full transparency about the limitations of their scope.

8.3 Implications and Recommendations for Demonstrating an Attack

This shares many of the same constraints and requirements as the goal of certifying vulnerability, with two key differences:

1. On the one hand, it is 'easier' in the sense that the detected effect need not meet a formal statistical criteria for significance: what ultimately matters is the evaluator's success in exploiting it for key recovery or information extraction.
2. Meanwhile, it is 'harder' in the sense that the evaluator needs to be able to trace the point(s) selected in the detection step of the attack to a particular intermediate value and (at the very least) some *a priori* knowledge about the form of the data-dependency. Therefore, specific rather than non-specific leakage detection tests are preferable in this setting.

In terms of coverage, it is not so much the breadth of coverage that is interesting as the type of leaks which are under consideration, which ideally should be carefully chosen to produce the most effective attack strategies. Clustering trace points would not be desirable in this instance as multivariate tests implicate whole groups of indices rather than singling out those individually responsible for leakage and thus vulnerable to attack.

We recommend to use as much prior information about the target (or similar) device(s) as possible to search for likely (specific) candidates, and to attempt attacks against any that appear promising. Once a suitable target has been found, repeat experiments may be necessary in order to provide a suitable metric for attack success (e.g. guessing entropy, success rate, or global key rank after an attack). How to proceed in the event that *no* attackable point is found is trickier, as it is unclear then whether the fault is with the attack or with the detection. Verifying the presence of leakage in the absence of a demonstrable attack then becomes equivalent to the problem of certifying vulnerability, requiring increased statistical rigour. If no leakage at all is found we are back in the more challenging scenario of certifying security.

8.4 Implications and Recommendations for Highlighting Vulnerabilities

Highlighting (all) vulnerabilities, e.g. so that they can be addressed by designers in the development process, is by far the most difficult goal to achieve, as it requires high *individual* power. The

combined probability of missing one or more leakages is going to get huge even in very large trace acquisitions. We say outright that analysts should temper their expectations and settle to find and fix 'as many as possible', without the ambition to claim comprehensive coverage.

The key recommendation towards this end is to use specific rather than non-specific tests, as it is essential to be able to track leaky points back to specific causal features in the implementation if measures are to be taken to address the vulnerabilities. On a related note, the clustering approach is not suitable, as the inability of the multivariate tests to indicate individual points within a cluster would likewise leave the designers without the guidance necessary to take useful action.

Presumably, (a few) false positives are less of an issue in this setting, as long as false negatives are minimised; we therefore recommend using the higher-powered FDR-controlling procedures rather than FWER-controlling ones to deal with the problem of multiple testing. Of course, the downside of doing so is that it may lead to 'unfixable' leakages which are impossible to address as they are not really present, resulting in wasted effort on the part of the designers.

References

1. Paul Ammann and Jeff Offutt. *Introduction to Software Testing*. Cambridge University Press, New York, NY, USA, 1 edition, 2008.
2. Dmitri Asonov and Rakesh Agrawal. Keyboard acoustic emanations. In *IEEE Symposium on Security and Privacy*, pages 3–11. IEEE Computer Society, 2004.
3. Christian Berg. The cube of a normal distribution is indeterminate. *The Annals of Probability*, 16(2):910–913, 1988.
4. Shivam Bhasin, Jean-Luc Danger, Sylvain Guilley, and Zakaria Najm. Side-channel leakage and trace compression using normalized inter-class variance. In Ruby B. Lee and Weidong Shi, editors, *HASP 2014, Hardware and Architectural Support for Security and Privacy*, pages 7:1–7:9. ACM, 2014.
5. Ran Bi and Peng Liu. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics*, 17(1):146, Mar 2016.
6. Julien Brouchier, Tom Kean, Carol Marsh, and David Naccache. Temperature Attacks. *IEEE Security & Privacy*, 7(2):79–82, 2009.
7. Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical Measurement of Information Leakage. In *TACAS*, pages 390–404, 2010.
8. Tom Chothia and Apratim Guha. A Statistical Test for Information Leaks Using Continuous Mutual Information. In *CSF*, pages 177–190, 2011.
9. Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
10. MO Columb and MS Atkinson. Statistical analysis: Sample size and power estimations. *BJA Education*, 16(5):159–161, 2016.
11. Jean-Luc Danger, Guillaume Duc, Sylvain Guilley, and Laurent Sauvage. Education and open benchmarking on side-channel analysis with the DPA contests. In *NIST Non-invasive attack testing workshop*, 2011.
12. W. Diehl, A. Abdulgadir, F. Farahmand, J. P. Kaps, and K. Gaj. Comparison of cost of protection against differential power analysis of selected authenticated ciphers. In *2018 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, pages 147–152, 2018.
13. A. Adam Ding, Liwei Zhang, Francois Durvaux, Francois-Xavier Standaert, and Yunsi Fei. Towards sound and optimal leakage detection procedure. In Thomas Eisenbarth and Yannick Teglia, editors, *Smart Card Research and Advanced Applications*, pages 105–122. Springer International Publishing, 2018.
14. Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.
15. François Durvaux and François-Xavier Standaert. From Improved Leakage Detection to the Detection of Points of Interests in Leakage Traces. In Marc Fischlin and Jean-Sébastien Coron, editors, *Advances in Cryptology – EUROCRYPT 2016*, volume 9665 of *LNCS*, pages 240–262. Springer, 2016.
16. François Durvaux, François-Xavier Standaert, and Santos Merino Del Pozo. Towards Easy Leakage Certification. In Benedikt Gierlichs and Axel Y. Poschmann, editors, *Cryptographic Hardware and Embedded Systems – CHES 2016*, pages 40–60, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.
17. Bradley Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4):1351–1377, 08 2007.
18. Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007.

19. Julie Ferrigno and Martin Hlaváč. When AES blinks: Introducing optical side channel. *IET Information Security*, 2(3):94–98, 2008.
20. Jacques J. A. Fournier, Simon Moore, Huiyun Li, Robert Mullins, and George Taylor. Security Evaluation of Asynchronous Circuits. In Colin D. Walter, Çetin K. Koç, and Christof Paar, editors, *Cryptographic Hardware and Embedded Systems – CHES 2003*, pages 137–151, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
21. Karine Gandolfi, Christophe Mourtel, and Francis Olivier. Electromagnetic Analysis: Concrete Results. In Çetin Kaya Koç, David Naccache, and Christof Paar, editors, *Proceedings of CHES 2001*, volume 2162 of *LNCS*, pages 251–261. Springer, 2001.
22. Gilbert Goodwill, Benjamin Jun, Josh Jaffe, and Pankaj Rohatgi. A testing methodology for side-channel resistance validation. In *NIST Non-invasive attack testing workshop*, 2011.
23. William Sealy Gosset. The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908. Originally published under the pseudonym "Student".
24. John M Hoenig and Dennis M Heisey. The Abuse of Power. *The American Statistician*, 55(1):19–24, 2001.
25. Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6:65–70, 01 1979.
26. Paul C. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In Neal Koblitz, editor, *Advances in Cryptology – CRYPTO '96*, volume 1109 of *LNCS*, pages 104–113. Springer, 1996.
27. Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. Differential Power Analysis. In *CRYPTO*, pages 388–397, 1999.
28. A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell' Istituto Italiano degli Attuari*, 4:83–91, 1933.
29. Daniël Lakens. Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.
30. Peng Liu and J. T. Gene Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746, 2007.
31. Stefan Mangard, Norbert Pramstaller, and Elisabeth Oswald. Successfully Attacking Masked AES Hardware Implementations. In Josyula R. Rao and Berk Sunar, editors, *Cryptographic Hardware and Embedded Systems – CHES 2005*, pages 157–171, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.
32. Luke Mather, Elisabeth Oswald, Joe Bandenburg, and Marcin Wójcik. Does My Device Leak Information? An a priori Statistical Power Analysis of Leakage Detection Tests. In Kazue Sako and Palash Sarkar, editors, *Advances in Cryptology – ASIACRYPT 2013*, volume 8269 of *LNCS*, pages 486–505. Springer, 2013.
33. Joan C. Miller and Clifford J. Maloney. Systematic Mistake Analysis of Digital Computer Programs. *Commun. ACM*, 6(2):58–63, February 1963.
34. F.-X. Standaert O. Bronchain, T. Schneider. Multi-Tuple Leakage Detection and the Dependent Signal Issue. *IACR Transactions on Cryptographic Hardware and Embedded Systems (to appear)*, 2019.
35. Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.
36. Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
37. Kristin E. Porter. Statistical power in evaluations that investigate effects on multiple outcomes: A guide for researchers. *Journal of Research on Educational Effectiveness*, 0(0):1–29, 2017.
38. Stan Pounds and Cheng Cheng. Sample size determination for the false discovery rate. *Bioinformatics*, 21(23):4263–4271, 2005.
39. Emmanuel Prouff and Matthieu Rivain. Theoretical and practical aspects of mutual information-based side channel analysis. *IJACT*, 2(2):121–138, 2010.
40. Jean-Jacques Quisquater and David Samyde. ElectroMagnetic Analysis (EMA): Measures and Counter-measures for Smart Cards. In Isabelle Attali and Thomas Jensen, editors, *Smart Card Programming and Security*, volume 2140 of *LNCS*, pages 200–210. Springer Berlin / Heidelberg, 2001.
41. Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.
42. Oscar Reparaz, Benedikt Gierlichs, and Ingrid Verbauwhede. Selecting Time Samples for Multivariate DPA Attacks. In *CHES*, pages 155–174, 2012.
43. S. S. Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):597–599, 2009.
44. Werner Schindler, Kerstin Lemke, and Christof Paar. A Stochastic Model for Differential Side Channel Cryptanalysis. In Josyula R. Rao and Berk Sunar, editors, *Proceedings of CHES 2005*, volume 3659 of *LNCS*, pages 30–46. Springer, 2005.
45. Tobias Schneider and Amir Moradi. Leakage Assessment Methodology – A Clear Roadmap for Side-Channel Evaluations. In Tim Güneysu and Helena Handschuh, editors, *Proceedings of CHES 2015*, volume 9293 of *LNCS*, pages 495–513. Springer, 2015.

46. Walter R. Schumm, Michael Higgins, Lorenza Lockett, Shuyi Huang, Nadyah Abdullah, Abdullah Asiri, Kennedy Clark, and Keondria McClish. Does dividing the range by four provide an accurate estimate of a standard deviation in family science research? a teaching editorial. *Marriage & Family Review*, 53(1):1–23, 2017.
47. Adi Shamir and Eran Tromer. Acoustic cryptanalysis (website). `http://theory.csail.mit.edu/~tromer/acoustic/`. (Accessed 5th September 2012).
48. Sergei Skorobogatov. Using Optical Emission Analysis for Estimating Contribution to Power Analysis. In Luca Breveglieri, Israel Koren, David Naccache, Elisabeth Oswald, and Jean-Pierre Seifert, editors, *Fault Diagnosis and Tolerance in Cryptography – FDTC '09*, pages 111–119. IEEE Computer Society, 2009.
49. N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948.
50. François-Xavier Standaert, Benedikt Gierlichs, and Ingrid Verbauwhede. Partition vs. Comparison Side-Channel Distinguishers: An Empirical Evaluation of Statistical Tests for Univariate Side-Channel Attacks against Two Unprotected CMOS Devices. In *ICISC*, pages 253–267, 2008.
51. François-Xavier Standaert. How (not) to Use Welch's T-test in Side-Channel Security Evaluations. In *CARDIS 2018 (to appear)*, LNCS. Springer, 2018.
52. François-Xavier Standaert, Olivier Pereira, Yu Yu, Jean-Jacques Quisquater, Moti Yung, and Elisabeth Oswald. Leakage Resilient Cryptography in Practice. In Ahmad-Reza Sadeghi and David Naccache, editors, *Towards Hardware-Intrinsic Security: Foundations and Practice*, pages 99–134. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
53. Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 12 2007.
54. Serdar Tasiran and Kurt Keutzer. Coverage metrics for functional validation of hardware designs. *IEEE Des. Test*, 18(4):36–45, July 2001.
55. Adrian Thillard, Emmanuel Prouff, and Thomas Roche. Success through Confidence: Evaluating the Effectiveness of a Side-Channel Attack. In Guido Bertoni and Jean-Sébastien Coron, editors, *Cryptographic Hardware and Embedded Systems – CHES 2013*, pages 21–36, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
56. Tiejun Tong and Hongyu Zhao. Practical guidelines for assessing power and false discovery rate for fixed sample size in microarray experiments. *Statistics in medicine*, 27:1960–72, 05 2008.
57. Nicolas Veyrat-Charvillon and François-Xavier Standaert. Mutual Information Analysis: How, When and Why? In *CHES*, pages 429–443, 2009.
58. Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.
59. Carolyn Whitnall, Elisabeth Oswald, and Luke Mather. An Exploration of the Kolmogorov-Smirnov Test as a Competitor to Mutual Information Analysis. In *CARDIS*, pages 234–251, 2011.

A Degradation of Distributional Assumptions after Pre-processing

The t-test makes the assumption that the underlying data are normally distributed. In the case of higher-order detection, where the data points are powers or products of (zero-mean) data points, we know that this is almost certainly not the case. The questions are then: to what extent do the resulting distributions diverge from normal? and, how much does it matter for the purposes of fair and conclusive evaluation?

We take for illustration the simplest case of a univariate higher order test in which the (mean centred) measurements at a single trace point are raised to a power m in order to detect data-dependent differences in the m^{th} moment. (A more complex scenario would be a test reducing separate points to a univariate quantity via a combining function, typically multiplication). Suppose further that the measurements are standardised to have a variance of 1. In practice, the decision to test for differences in moments of 3 or above implies an expectation that, thanks to the joint distribution of the shares of the intermediate, the partitioned distributions are *not* normal, as otherwise there would be nothing to detect (normal distributions are fully determined by their mean and variance). However, for the purposes of being able to say something indicative about the impact of pre-processing we assume a normal initial distribution.

The density of $Y = Z^m$ where $Z \sim \mathcal{N}(0, 1)$ is a standard normal random variable is as follows (see, e.g., [3]):

$$f_Y(y) = \begin{cases} \frac{1}{m\sqrt{2\pi}} \cdot |y|^{(1/m)-1} \exp\left\{-\frac{1}{2}|y|^{2/m}\right\} & \text{if } m \text{ is odd} \\ \\ \frac{2}{m\sqrt{2\pi}} \cdot y^{(1/m)-1} \exp\left\{-\frac{1}{2}y^{2/m}\right\} \mathrm{I}_{\{y>0\}} & \text{if } m \text{ is even} \end{cases}$$

where $\mathrm{I}_{\{A\}}$ is the indicator function, i.e.

$$\mathrm{I}_{\{A\}} = \begin{cases} 0 & \text{if } A \text{ is false} \\ 1 & \text{if } A \text{ is true.} \end{cases}$$

Figure 12 plots these densities for increasing values of m. Table 5 presents the moments. Some things to note:

- For all powers $m \geq 2$ the density has a singularity at zero.
- For even powers the density is zero for all negative values (among other things this means they have non-zero means and are positive skewed; the skewness increases with the order). For odd powers the density remains symmetric.
- The kurtosis increases with the power.
- All odd order standardised moments are zero for odd powers and increasing for even powers.
- All even order standardised moments are positive and increasing for all powers.

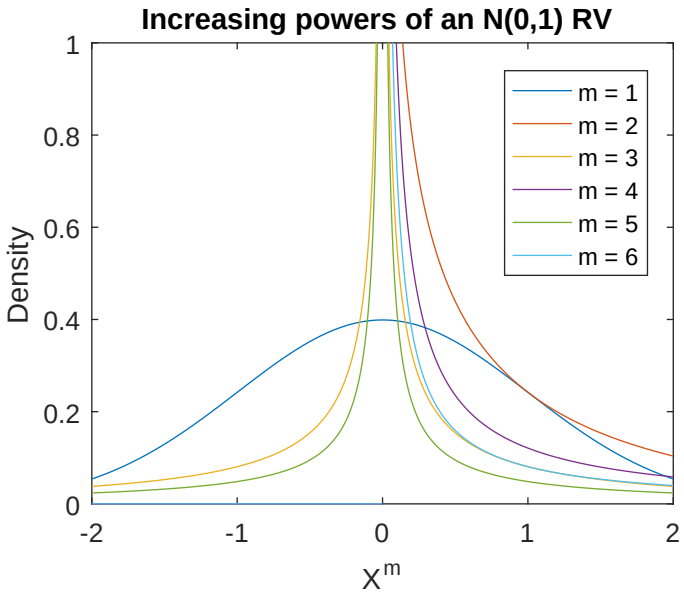

Fig. 12. Density of a standard normal random variable raised to increasing powers.

It is clear by visual inspection and by computation of the moments that raising a (mean-centred) normal to positive integer powers produces distinctly non-normal distributions. However, the usual argument made in support of the t-test to detect differences in thus pre-processed distributions is that the sampling distributions of their means tend to normal under the central limit theorem (CLT; see, e.g. [16]) – implying that the Type I error is correctly controlled once the sample size is 'large enough'. One problem with this reasoning is that the rate of convergence is known to be very variable depending on the underlying distribution.

Distribution	Mean	Var	Skew	Kurt	SM5	SM6	SM7	SM8
m = 1 (Normal)	0	1	0	3	0	15	0	105
m = 2 (χ^2)	1	2	3	15	96	755	6.98e+03	7.44e+04
m = 3	0	15	0	46	0	1.02e+04	0	6.25e+06
m = 4	3	96	10	207	6.92e+03	3.44e+05	2.39e+07	2.2e+09
m = 5	0	945	0	733	0	7.34e+06	0	4.01e+11
m = 6	15	1.02e+04	33	3.04e+03	5.91e+05	2.1e+08	1.23e+11	1.11e+14

Table 5. Moments of distribution as power increases.

To get a sense of the factors affecting the convergence of sample means we use the Shapiro–Wilk test [7] (the most powerful of those tested in a 2011 study by Razali and Wah [41]) to test a null hypothesis of normality of the sample mean as the sample size increases. As above, the initial distributions we consider are all zero-mean normal, now allowing for the standard deviation to vary.

The Shapiro–Wilk test decides whether or not to reject a null hypothesis that the distribution is in fact normal. Figure 13 shows the rejection rates as the sample size increases, based on 1,000 experiments each with 1,000 draws of the sample to estimate the mean. With a sample size of about 10,000 the sample means of the squared and cubed distributions, have more or less converged to the significance level (i.e. false rejection rate) of the test which we set at 0.05. The distributions raised to the powers of 4 and 5 take a bit longer. The skewness produced by raising the distributions to even powers results in a slower convergence to normality than that associated with the distributions obtained by adding 1 to the power to make it even (recall that, as per the Berry–Esseen theorem, the constant in the convergence rate depends on the third normalised moment). Within odd and even powers, though, the convergence is slower as the power increases. It is striking that the convergence seems unaffected by the standard deviation, implying that in noisy scenarios where large samples will anyway be needed, the asymptotic distributional assumptions may be sufficient for the purposes of controlling the Type I errors for leakage detection. However, this gives no assurances with respect to the Type II errors, as statistical power analysis (e.g. to determine the required sample size for a given power) derives from the distributional assumptions of the raw measurements, not from the sampling distribution of the mean. Unless and until a tailored solution can be found to derive and control the statistical power of t-tests performed on pre-processed univariate traces, such methods remain incapable of drawing fair and conclusive conclusions for the purposes of evaluation.

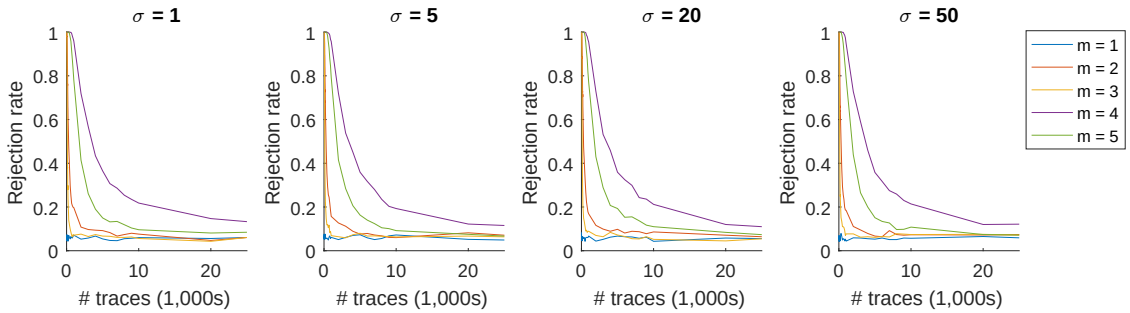

Fig. 13. Rejection rate of the null hypothesis that the distribution of the sample mean is normal, as the number of traces increases.

[7] Implemented in Matlab as `swtest(·)`, by Ahmed BenSaïda.

The distributions of the products of multiple normal random variables are even more complex in form, depending also on the dependencies between the initial distributions. Further work would be needed to ascertain the convergence of such distributions under realistic assumptions.

B Sample Size for the *t*-Test

We begin with a simple visual example that illustrates the concepts of α and β values and their relationship to the sample size.

Consider the following two-sided hypothesis test for the mean of a Gaussian-distributed variable $A \sim \mathcal{N}(\mu, \sigma)$, where μ and σ are the (unknown) parameters:

$$H_0 : \mu = \mu_0 \text{ vs. } H_{alt} : \mu \neq \mu_0. \tag{10}$$

Note that, in the leakage detection setting, where one typically wishes to test for a non-zero difference in means between *two* Gaussian distributions Y_1 and Y_2, this can be achieved by defining $A = Y_1 - Y_2$ and (via the properties of the Gaussian distribution) performing the above test with $\mu_0 = 0$.

Suppose the alternative hypothesis is true and that $\mu = \mu_{alt}$. This is called a 'specific alternative'[8], in recognition of the fact that it is not usually possible to compute power for *all* the alternatives when H_{alt} defines a set or range. In the leakage detection setting one typically chooses $\mu_{alt} > 0$ to be the smallest difference $|\mu_1 - \mu_2|$ that is considered of practical relevance; this is called the effect size. Without loss of generality, we suppose that $\mu_{alt} > \mu_0$.

Figure 14 illustrates the test procedure when the risk of a Type I error is set to α and the sample size is presumed large enough (typically $n > 30$) that the distributions of the test statistic under the null and alternative hypotheses can be approximated by Gaussian distributions. The red areas together sum to α; the blue area indicates the overlap of H_0 and H_{alt} and corresponds to β (the risk of a Type II error). The power of the test – that is, the probability of correctly rejecting the null hypothesis when the alternative in true – is then $1 - \beta$, as depicted by the shaded area.

There are essentially three ways to raise the power of the test. One is to increase the effect size of interest which, as should be clear from Figure 14, serves to push the distributions apart, thereby diminishing the overlap between them. Another is to increase α – that is, to make a trade-off between Type II and Type I errors – or (if appropriate) to perform a one-sided test, either of which has the effect (in this case) of shifting the critical value to the left so that the shaded region becomes larger. (In the leakage detection case the one-sided test is unlikely to be suitable as differences in either direction are equally important and neither can be ruled out *a priori*). The third way to increase the power is to increase the sample size for the experiment. This reduces the standard error on the sample means, which again pushes the alternative distribution of the test statistic further away from null (note from Figure 14 that it features in the denominator of the distance).

Suppose you have an effect size in mind – based either on observations made during similar previous experiments, or on a subjective value judgement about how large an effect needs to be before it is practically relevant (e.g. the level of leakage which is deemed intolerable) – and you want your test to have a given confidence level α and power $1 - \beta$. The relationship between confidence, power, effect size and sample size can then be used to derive the minimum sample size necessary to achieve this.

[8] The overloading of terminology between 'specific alternatives' and 'specific' TVLA tests is unfortunate but unavoidable.

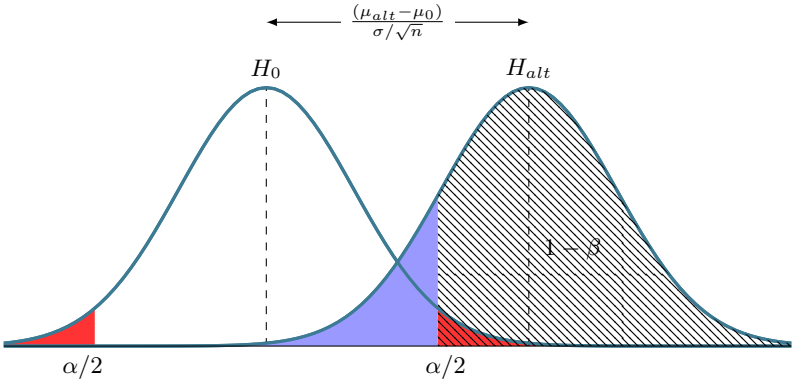

Fig. 14. Figure showing the Type I and II error probabilities, α and β as well as the effect size $\mu_{alt} - \mu_0$ for a specific alternative such that $\mu_{alt} > \mu_0$.

The details of the argumentation that now follows are specific to a two-tailed t-test, but the general procedure can be adapted to any test for which the distribution of the test statistic is known under the null and alternative hypotheses.

For the sake of simplicity (i.e. to avoid calculating effectively irrelevant degrees of freedom) we will assume that our test will in any case require the acquisition of more than 30 observations, so that the Gaussian approximations for the test statistics hold as in Figure 14. Without loss of generality we also assume that the difference of means is positive (otherwise the sets can be easily swapped). Finally, we assume that we seek to populate both sets with equal numbers $n = |Y|/2$ of observed traces.

Theorem 1. *Let Y_1 be a set of traces of size $N/2$ drawn via repeat sampling from a normal distribution $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y_2 be a set of traces of size $N/2$ drawn via repeat sampling from a normal distribution $\mathcal{N}(\mu_2, \sigma_2^2)$. Then, in a two-tailed test for a difference between the sample means:*

$$H_0\text{: } \mu_1 = \mu_2 \text{ vs. } H_{alt}\text{: } \mu_1 \neq \mu_2, \tag{11}$$

in order to achieve significance level α and power $1-\beta$, the overall number of traces N needs to be chosen such that:

$$N \geq 2 \cdot \frac{(z_{\alpha/2} + z_\beta)^2 \cdot ({\sigma_1}^2 + {\sigma_2}^2)}{(\mu_1 - \mu_2)^2}. \tag{12}$$

Note that Equation 12 can be straightforwardly rearranged to alternatively compute any of the significance level, effect size or power in terms of the other three quantities.

C Sample Size for Hotelling's T^2-Test

As a first step towards deriving an expression for the power (in the balanced case $n_A = n_B = \frac{n}{2}$) we first derive the threshold at which the null hypothesis is rejected:

$$
\begin{aligned}
&\mathbb{P}(T^2 > \text{cv}|H_0) = \alpha \\
\Rightarrow &\mathbb{P}\left(\frac{(n-2)p}{n-p-1}F_{p,n-p-1} > \text{cv}\right) = \alpha \\
\Rightarrow &\mathbb{P}\left(F_{p,n-p-1} > \frac{n-p-1}{(n-2)p}\text{cv}\right) = \alpha \\
\Rightarrow &\mathbb{P}\left(F_{p,n-p-1} \leq \frac{n-p-1}{(n-2)p}\text{cv}\right) = 1-\alpha \\
&\Rightarrow \text{cv} = \frac{(n-2)p}{n-p-1}F^{-1}_{p,n-1-p}(1-\alpha)
\end{aligned}
$$

This value of cv can then be plugged into the power computation. Under a specific alternative hypothesis H_{alt} such that $D = D_{alt} > 0$ (by comparison with the null H_0 of $\mathbf{d} = \mathbf{0} \Rightarrow \mathbf{D} = \mathbf{0}$) the test statistic has non-central F distribution with non-centrality parameter $\lambda = \frac{n}{4}D^2_{alt}$ (note that this is particular to the $n_A = n_B = \frac{n}{2}$ case) [18]. Then the power can be computed as follows:

$$
\begin{aligned}
\text{power} = 1-\beta &= \mathbb{P}(T^2 > \text{cv}|H_{alt}) \\
&= \mathbb{P}\left(T^2 > \text{cv}\middle|T^2 \sim \frac{(n-2)p}{n-p-1}F_{p,n-p-1;\lambda}\right) \\
&= \mathbb{P}\left(\frac{n-p-1}{(n-2)p}T^2 > \frac{n-p-1}{(n-2)p}\text{cv}\middle|\frac{n-p-1}{(n-2)p}T^2 \sim F_{p,n-p-1;\lambda}\right) \\
&= 1-\mathbb{P}\left(\frac{n-p-1}{(n-2)p}T^2 \leq \frac{n-p-1}{(n-2)p}\text{cv}\middle|\frac{n-p-1}{(n-2)p}T^2 \sim F_{p,n-p-1;\lambda}\right) \\
&= 1-F_{p,n-p-1;\lambda}\left(\frac{n-p-1}{(n-2)p}\text{cv}\right) \\
&= 1-F_{p,n-p-1;\lambda}\left(\frac{n-p-1}{(n-2)p} \times \frac{(n-2)p}{n-p-1}F^{-1}_{p,n-1-p}(1-\alpha)\right)
\end{aligned}
$$

D Evolution of D as Cluster Covariance and Density Varies

Bronchain et al. [34] consider three covariance structures for their theoretical analysis, designed to capture increasing dependency between the trace points. These take the form $\Sigma = (\sigma^2_{i,j}) \in \mathbb{R}^{n_c \times n_c}$ where:

1. $\sigma^2_{i,j} = 0, \quad i \neq j$ (i.e., the points are independent).
2. $\sigma^2_{i,j} = \max(1 - 0.1|i-j|, 0)$.
3. $\sigma^2_{i,j} = \max(1 - 0.02|i-j|, 0)$.

Because we are interested in *standardised* (univariate) effects, d (the original authors worked in terms of the SNR), we have that $\sigma^2_{i,i} = 1 \quad \forall i = 1, \ldots, n_c$ in all three cases.

The rationale for the above choices is that points closest to each other will have greater similarity than points further away, which of course doesn't directly translate to our setting, where the groups

are supposed to have been formed according to some similarity metric rather than according to order in the trace. Still, we consider these scenarios for the sake of comparison with previous work.

The authors also allow the *density* of the effects – that is, the proportion of trace points with non-zero leakage – to vary. However, they do this under the assumption of independence only (that is, they do not allow the density and covariance to vary together). We consider two values for the density of the clusters: 1 and 0.5. (Although, in the case where all trace points are jointly tested, we cap the total number of leaky points to 30, as per our hypothetical scenario). In covarying clusters of density < 1 we suppose that the trace points are organised such that the non-zero leaks appear at the start, followed by the zero leaks.

Table 6 shows the D values associated with each of the scenarios. In all cases, as the number of points increases, the effect size increases. However, for the non-independent clusters the increases are considerably smaller. The mixed clusters (density < 1) are surprising in that, despite having fewer leaky points, they exhibit larger D values than the fully leaky clusters when dependencies come into play. This phenomenon is associated with the dependencies between leaky and non-leaky points, and is reversed if the two 'types' of points are supposed to be independent (i.e. if the corresponding entries in the covariance matrices are set to zero), as we show in the final three columns of the table.

Cluster	**Density=1**			**Density=0.5**					
size	Σ_1	Σ_2	Σ_3	Σ_1	Σ_2	Σ_3	Σ_1'	Σ_2'	Σ_3'
1	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
7	0.106	0.048	0.041	0.080	0.093	0.201	0.080	0.043	0.041
20	0.179	0.068	0.044	0.126	0.096	0.201	0.126	0.054	0.042
1,400 (capped at 30 leaks)	0.219	0.217	0.278	0.219	0.217	0.278	0.219	0.079	0.047

Table 6. Multivariate effect sizes under different assumptions about the density and covariance structure of the clusters.

Figures 15 to 20 show, by way of illustrative example, the statistical power of cluster-based tests with the Bonferroni correction in the scenarios represented in the first three and final three columns of Table 6.

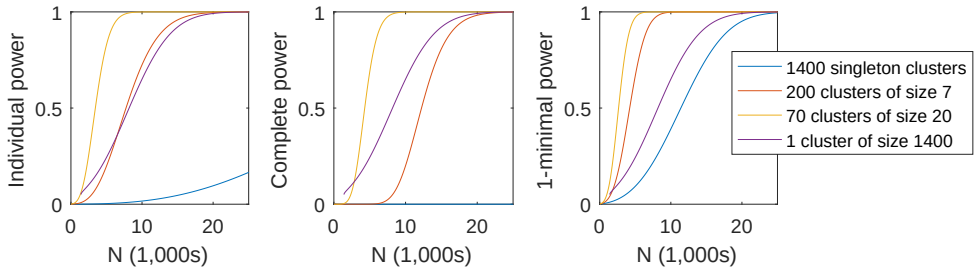

Fig. 15. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (perfect) clusters varies; covariance scenario 1 (independent trace points).

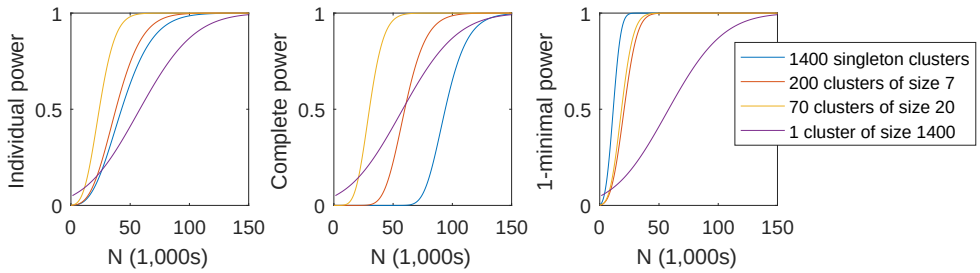

Fig. 16. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (perfect) clusters varies; covariance scenario 2 (medium dependency).

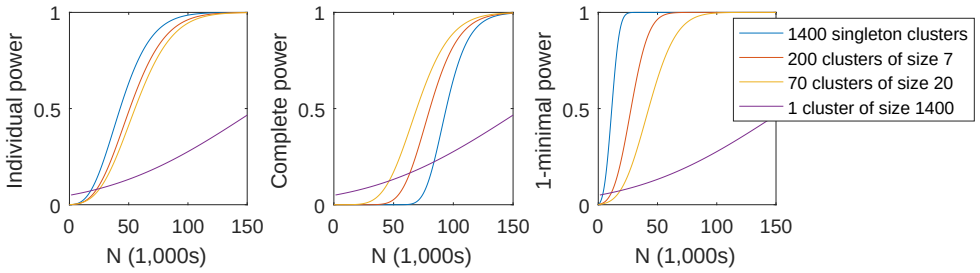

Fig. 17. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (perfect) clusters varies; covariance scenario 3 (high dependency).

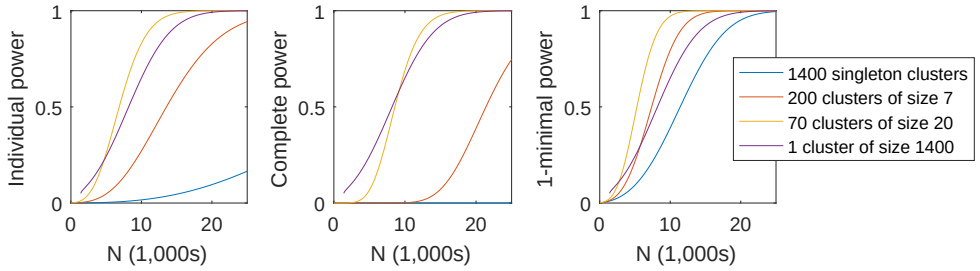

Fig. 18. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (mixed) clusters varies; covariance scenario 1' (independent trace points).

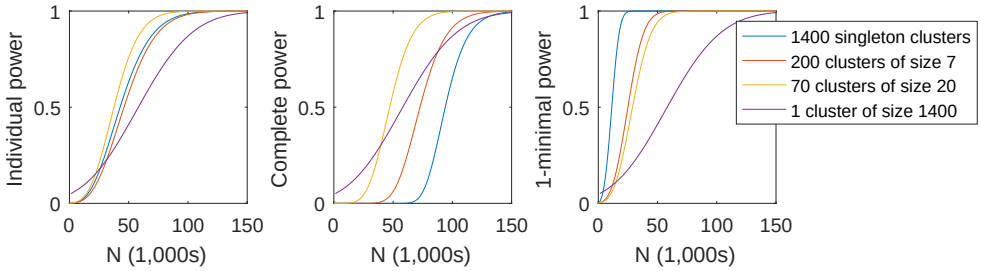

Fig. 19. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (mixed) clusters varies; covariance scenario 2' (medium dependency).

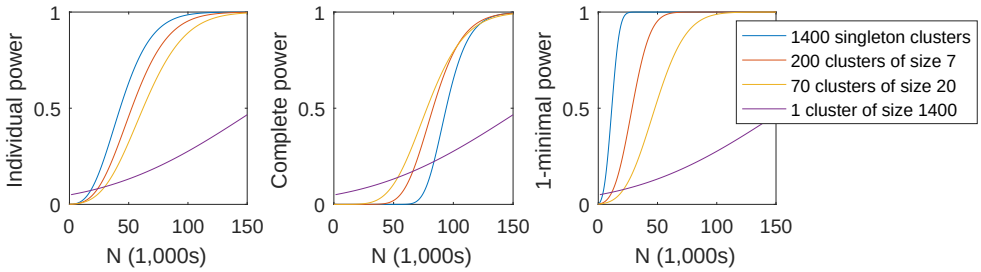

Fig. 20. Different types of power for Bonferroni-adjusted Hotelling's tests as the size of the (mixed) clusters varies; covariance scenario 3' (high dependency).