A Unified Submodular Framework for Multimodal IC Trojan Detection

Farinaz Koushanfar, Azalia Mirhoseini, and Yousra Alkabani

Rice University, Electrical and Computer Engineering, Houston TX 77005, USA
{farinaz,azalia,yousra}@rice.edu

Abstract. This paper presents a unified formal framework for integrated circuits (IC) Trojan detection that can simultaneously employ multiple noninvasive measurement types. Hardware Trojans refer to modifications, alterations, or insertions to the original IC for adversarial purposes. The new framework formally defines the IC Trojan detection for each measurement type as an optimization problem and discusses the complexity. A formulation of the problem that is applicable to a large class of Trojan detection problems and is *submodular* is devised. Based on the objective function properties, an efficient Trojan detection method with strong approximation and optimality guarantees is introduced. Signal processing methods for calibrating the impact of inter-chip and intra-chip correlations are presented. We propose a number of methods for combining the detections of the different measurement types. Experimental evaluations on benchmark designs reveal the low-overhead and effectiveness of the new Trojan detection framework and provides a comparison of different detection combining methods.

1 Introduction

The prohibitive cost of manufacturing ICs in nano-meter scales has made the use of contract foundries the dominant semiconductor business practice. Unauthorized IP usage, IC overbuilding, and insertion of additional malware circuitry (*Trojans*) are a few of the major threats facing the horizontal IC industry where the IP providers, designers, and foundries are separate entities [8]. The Trojan attacker modifies the original design to enable an adversary to control, monitor, spy contents and communications, or to remotely activate/disable parts of the IC. Trojans are often hidden and are rarely triggered as needed.

A standing challenge for noninvasive IC testing and Trojan detection is dealing with the increasing complexity and scale of the state-of-the-art technology. It is hard to distinguish between the characteristic deviations because of the process variations and the alterations due to the Trojan insertion. What complicates the problem even more is that the space of possible changes by the adversary is large. Very little is known or documented about IC Trojan attacks. The possible adversaries are likely financially and technologically advanced and thus, intelligent attacks are possible. Because of the hidden functional triggering of Trojans, the logic-based testing methods are unlikely to trigger and distinguish

R. Böhme, P.W.L. Fong, and R. Safavi-Naini (Eds.): IH 2010, LNCS 6387, pp. 17–32, 2010.

the malicious alterations. The conventional parametric IC testing methods have a limited effectiveness for addressing Trojan related problems. Destructive tests and IC reverse-engineering are slow and expensive.

This paper formally devises a new unified framework that simultaneously integrates the results of several noninvasive measurement types. Each noninvasive measurement type is called a *modality*: *unimodal* detection employs a single measurement modality for finding the internal characteristics of the chip, while *multimodal* detection combines the measurements from several modalities to reveal the unwanted changes to the original design. We show that the detection objective for each modality is *submodular*. The submodularity property formalizes the intuition that inserting a Trojan would have a higher impact on a small circuit than inserting the same Trojan to a larger circuit that contains the smaller circuit as a subpart. The concept is demonstrated in Figure 1. The design

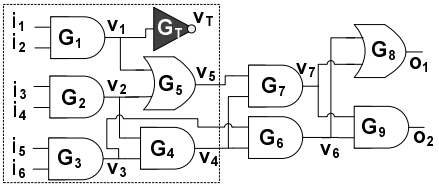

Fig. 1. The submodular property

consists of 9 gates $G_1, \ldots, G_9$, 6 inputs and 2 outputs. A Trojan gate G_T is added. Consider a subcircuit of this design composed of gates $G_1, \ldots, G_5$ in the dotted square that also includes G_T. Now, for one input vector applied to the circuit, the ratio of the current leaked by G_T to the rest of the circuit leakage would be higher for the subcircuit compared to the whole circuit. We exploit the theoretical results known for submodular functions to propose a near-optimal Trojan detection algorithm. Our contributions are as follows: (1) Proposing a unified noninvasive Trojan detection framework, (2) Formulating the optimization problem for simultaneous gate level profiles and Trojan detection for each modality, (3) Exploiting submodularity to achieve a near optimal solution for unimodular detection, and (4) Devising and comparing four methods for combining the results of multiple unimodal detections on benchmark designs.

2 Related Work

Hardware Trojan detection is a new and emerging research area. Agrawal et al. [1] use destructive tests to extract a fingerprint for a group of unaltered chips based on the global transient power signal characteristics. The other chips would be noninvasively tested against the extracted fingerprints by statistical Hypothesis testing. The overhead of destructive testing, sensitivity to noise and process variations, and lack of usage of the logical structure and constraints are the drawbacks of this method.

Banga et al. [4,3] propose a region-based testing that first identifies the problematic regions based on power signatures and then performs more tests on the region. The underlying mathematical and logical circuit structure or the process variations are not considered. Rad et al. [24,23] investigate power supply transient signal analysis methods for detecting Trojans. The focus is on test signatures and not on the lower-level components (e.g., the gate level characteristics). Rad et al. further improved the resolution of power analysis techniques to Trojans by carefully calibrating for process and test environment (PE) variations. The main focus of this research is on the evaluation of four experimental signal calibration techniques, each designed to reduce the adverse impact of PE variations on their detection method. They also investigated the sensitivity of their Trojan detection method in terms of determining the smallest detectable Trojan under conditions such as measurement noise.

Jin and Markis [11] extract the path delay fingerprints by using the well-known principal component analysis that is a statistical dimension reduction technique. They use Hypothesis testing against the delay fingerprints to detect the anomalies. This approach also does not consider the gate level components and would also require exponential path measurements in the worst case. Li and Lach propose adding on chip delay test structures for Trojan detection [16]. Gate level characterization for noninvasive post-silicon IC profiling [12] and for Trojan detection was used in [20,22,2,28]. However, the previous work did not provide a systematic algorithm with any kind of optimality, nor they addressed calibration, sensitivity, or multimodal combining. Our work provides the first rigorous treatment of the multimodal Trojan detection problem, near-optimal solutions, mathematical calibration. Even though a number of authors suggested the potential benefits of combining different measurement types, to the best of our knowledge no systematic approach with evaluation results on combining different test and measurement modalities was reported.

Our method exploits the concept and results of submodular function optimization [21]. The concept has been utilized earlier in a variety of contexts [13], including but not limited to: set cover [9], sensor networks [15], linear regression [7], graph problems [6], and social networks [18]. To the best of our knowledge, our work is the first to use submodularity for IC Trojan detection.

3 Preliminaries

In this section, we provide the necessary background and the measurement setup.

(i) Process variations. As the CMOS dimensions shrink, uncertainty in the device characteristics increases. The variations might be temporal or spatial. In controlled settings, the dominant source of difference between the chips is the spatial variation [27,17]. Spatial variation may be intra-die, or inter-die, and could be systematic or random. We use the widely adopted Gaussian variation models [17]. Timing and dynamic power variations are a linear function of the variations and follow the same Gaussian patterns, while the leakage has an

approximately lognormal distribution. Our approach works for the stationary process variation models.

(ii) Trojan threat model. From the conventional testing and inspections point of view, the Trojan IC has exactly the same set of I/O pins, has the same deterministic I/O response as the original plan, and has the same physical form factor. A Trojan causes a change in the statistical distribution of the estimated gate characteristics that otherwise follow the process variation distributions. Our method uses the likelihood of the post-silicon characteristics for detection. The intra-chip correlations are assumed to have a lower amplitude when compared to the impact of the Trojans on the estimated profiles. The nominal values for the gate characteristics are available via the factory provided simulation models needed to ensure design-time power control and timing closure.

(iii) Measurement setup. We use similar measurement set-up as the conventional testing. However, our assumption is that the chip has already passed the standard automatic test pattern generation (ATPG) tests, and does not include any of the standard faults. For timing measurements, we exploit the classic timing test and validation techniques: Given an input vector to a test chip, applying the test input vector with multiple clock frequencies can give us the pass/fail behavior of the chips. The path delay would be the shortest clock period for which the chip does not fail with the intended input vector. We use the testing pattern generation method described in [29].The leakage current can be measured via the commonly known IDDQ test methods, where IDDQ refers to the measurement of the quiescent power-supply current. The IDDQ tests are often done via the off-chip pins by the precision measurement unit (PMU) [26]. The dynamic current tests are referred to as IDDT tests. IDDT tests can be done by averaging methods that do not require high precision or high frequency measurement devices needed for capturing the transient signals [10].

4 Unimodal Trojan Detection

The basis of the unimodal Trojan approach is the gate profiling discussed in Section 4.1. In Section 4.2 we show the detection problem is submodular. We further discuss the complex structure of the general unimodal detection problem which cannot be optimally addressed. We opt to use our prior knowledge about the process variations and submodularity property to address the problem in a hierarchical way. The precursor for our hierarchical method is systematic calibration that is discussed in Section 4.3.

4.1 Gate Profiling

In this subsection, we show how the side-channel measurements can be decomposed to their gate level components post-silicon. One can exploit the linear relationship between the IC's gate level profile and the side-channel measurements (constrained by the logic relations) to estimate the gate level characteristics. We introduce a formal framework for this problem:

Problem. UNIMODAL GATE PROFILING (MODALITY M).

Given. A combinational circuit **C**, with N_I primary inputs $x_1, \ldots, x_{N_I}$, and N_O primary outputs $z_1, \ldots, z_{N_O}$, where the netlist and logic structure is fully available. The circuit consists of interconnections of single-output gates where each gate $G_k, k = 1, \ldots, N_g$ implements an arbitrary logic function. The nominal profile of G_k for the modality M for each possible combination of gate inputs is available from the technology libraries and simulations models.

Measurements. For the modality M, a set of input vectors (V's) that are each an N_I tuple $(v_1, v_2, \ldots, v_{N_I})$, where $v_j \in \{0, 1\}$ for $j = 1, \ldots, N_I$ are available. Component values of V are applied to the primary inputs $x_1, \ldots, x_{N_I}$ which changes the states of the internal gates. For one or more input vectors, the side channel measurement is recorded either from the output pins, from other external pins, or contactless. The side-channel measurement is a linear combination of the gate characteristics in measurement modality M and a measurement error.

Objective. Estimate the post-silicon profile of each gate for the modality M.

A key step in noninvasive profiling of the chips is generation of input vectors that can controllably change the states of the gates such that this modification is observable from the measurement medium. Note that generating the input patterns that can distinctively identify each gate's characteristics is known to be NP-complete and has been a subject of extensive research in circuit testing [10]. We use the best known methods in testing for generation of the input vector patterns that maximally cover all the gates. Although we are limited by the same constraints as testing in terms of gate coverage, the difference here is that we are not detecting a particular fault model or the worst-case behavior (e.g., critical paths or stuck at fault) but we are estimating the gate parameters that may incur a certain error.

In delay testing, the input vectors have to functionally *sensitize* the tested paths such that the output is observable at an output pin. Delay test generation methods for sensitizing paths that achieve a high coverage, i.e., exercising many paths are available. Since there is a linear relationship between the tested paths and the gate delays, the explored paths directly translate to high coverage of the gate delays. We use the path sensitizing method proposed by Murakami et al [19]. Similarly, we use the available high coverage test vector generation methods for IDDQ, and for IDDT testing [25]. The number of generated input vectors is linear with respect to the total number of gates.

(i) Timing modality. The noninvasive timing measurements are taken by changing the inputs and measuring the time propagation of input transition to the output nodes. In this paper we consider the gate delays and ignore the wires. However, we emphasize that since the wire timings are linearly added to the path delays (assuming that crosstalk is bounded by controlling the possible couplings), their inclusion in the linear formulations is straightforward.

One can test J different paths and write a linear system of delay equations. The gate delays ($T(G_{k_j})$) on each chip are the variables (because of the process variation and operation effect) and $T_{meas}(P)$'s are the path delays that are measured on each chip:

$$EQ_j : \Sigma_{k_j=1}^{K_j} T(G_{k_j}) = T_{meas}(P_j), P_j = \{G_{k_j}\}_1^{K_j}, \qquad 1 \leq j \leq J \quad .$$

Noninvasive gate profiling aims at solving the above system of equations in presence of measurement error. If measuring the path delay $T_{meas}(P_j)$ incurs the error $\epsilon_T(P_j)$, the optimization problem objective function (OF) and constraints (C's) can be written as follows:

$$OF : \min_{1 \leq j \leq J} \mathcal{F}(\epsilon_T(P_j)) \tag{1}$$

$$C's : \Sigma_{k_j=1}^{K_j} T(G_{k_j}) = T_{meas}(P_j) + \epsilon(P_j), P_j = \{G_{k_j}\}_1^{K_j}, \qquad 1 \leq j \leq J \quad .$$

where $\mathcal{F}$ is a metric for quantifying the measurement errors; commonly used forms of $\mathcal{F}$ are the maximum likelihood formulation, or the l_p norms of errors defined as: $l_p = (\sum_{j=1}^{J} |\epsilon_T(P_j)|^p)^{1/p}$, *if* $1 \leq p < \infty$, and $l_p = max_{j=1}^{J} |\epsilon_T(P_j)|$, *if* $p = \infty$.

The delay of one gate $T(G_k)$ can be further written in terms of the deviation from the nominal delay of this gate from the value specified in the technology files. If the nominal gate delay value for the gate type G_k is $T^{nom}(G_k)$ and the deviation from nominal for G_k for the chip under measurement is $\theta_T(G_k)$, then $T(G_k) = \theta_T(G_k)T^{nom}(G_k)$ and thus, the unknowns are $\theta_T(G_k)'s$ and $\epsilon(P_j)'s$. The variable $\theta_T(G_k)$ is called the *delay (timing) scaling factor* of G_k. If there were no path measurement errors, the number of equations (J) required to have a full-rank system would be the same as the number of variables (gate delays). In presence of errors, the number of required equations is slightly higher, but the order is still linear in terms of number of gates N_g.

(ii) Leakage power modality. The leakage measurements rely on the fact that leakage is a function of the gate input for each gate type. Since the supply voltage is fixed, the static power is only dependent on the leakage current. For each input vector in quiescent state, the external pin current can be measured and written in terms of the sum of the individual components. For example, for (v_1; v_2; v_3; v_4) =1111, the total measured leakage current can be written as: $\Phi_{meas}(1111) + \epsilon_\Phi(1111) = \Phi_{G_1}(11) + \Phi_{G_2}(11) + \Phi_{G_3}(00) + \Phi_{G_4}(00) + \Phi_{G_5}(11) + \Phi_{G_6}(11)$, where $\Phi_{G_k}(x_1x_2)$ is the leakage current for gate G_k for its incident input (x_1x_2), and $\epsilon_\Phi(.)$ denotes the measurement error for the incident input. Each gate's leakage can be further decomposed to the nominal leakage value for the gate type and a *leakage scaling factor* denoted by $\theta_\Phi(G_k)$, i.e., $\Phi_{G_k}(x_1x_2) = \theta_\Phi(G_k)\Phi_{G_k}^{nom}(x_1x_2)$. Therefore, the linear optimization can be written over the J leakage measurements:

$$OF : \min_{1 \leq j \leq J} \mathcal{F}(\epsilon_\Phi(X_j)) \tag{2}$$

$$C's : \Sigma_{k=1}^{N_g} \theta_\Phi(G_k)\Phi_{G_k}^{nom}(x_j) = \Phi_{meas}(X_j) + \epsilon_\Phi(X_j) \quad .$$

(iii) Dynamic power modality. The dynamic power is dependent on the input transition. The measured average dynamic current (Section 3(iii)) can be written as the sum of the gate dynamic currents. Thus, the linear optimization for J dynamic current measurements would be:

$$OF: \min_{1\leq j\leq J} \mathcal{F}(\epsilon_\Psi(X_{j\rightarrow j+1})) \tag{3}$$
$$C's: \Sigma_{k=1}^{N_g} \theta_\Psi(G_k)\Psi_{G_k}^{nom}(x_{j\rightarrow j+1}) = \Psi_{meas}(X_{j\rightarrow j+1}) + \epsilon_\Psi(X_{j\rightarrow j+1}) \quad .$$

where $\epsilon_\Psi(.)$ denotes the measurement error for a reading, $\theta_\Psi(G_k)$ is the gate *scaling factor for the dynamic current*, $\Psi_{G_k}^{nom}(.)$ is the nominal dynamic current value for gate G_k for the pertinent transition, $x_{j\rightarrow j+1}$ refers to input vector transition from the vector j to j_1, and $\Psi_{meas}(.)$ is the extracted dynamic current measured at the external pin.

We see that each of the modalities can be written in the unified format of a system of linear equations. In the remainder of the paper we use the following generic notations for the gate profiling over the different modalities.

(i) **OF:** $\min \mathcal{F}(\epsilon)$, **Constraints:** $A\theta = B+\epsilon$; where $A_{[J\times N_g]}$ and $B_{[J]}$ are given by the technology values and J measurements, $\theta_{[N_g]}$ is a vector of unknown scaling factors, and $\epsilon_{[N_g]}$ is a vector of measurement errors.

(ii) Alternatively, the optimization problem can be written as **OF:** $\max \mathcal{L}(\epsilon)$, **Constraints:** $A\theta = B+\epsilon$, where $\mathcal{L}$ is the likelihood that the variations are coming from a certain distribution, e.g., normal distribution. Under the assumption of normal error distribution, maximizing the likelihood corresponds to minimizing the $\mathcal{F} = l_2$ error norm.

4.2 Unimodal Detection

Let us assume that the gates are positioned at the locations $\mathcal{D}$ in the 2D layout space. For a single modality we can find an estimation of each gate's profile. As we described in Section 3(i), the profile of a benign gate can be modeled as the sum of its inter-chip and intra-chip systematic process variations, the random process variations, and measurement noise. The global objective of Trojan detection is to maximize the probability of Trojan detection (P_D) and to minimize the probability of false alarm (P_{FA}). However, explicit formulation of the two objectives is not plausible, since probability of detection/false alarm can only be determined for cases where we know the exact Trojan attack and the ground truth. Instead, our Trojan detection attempts at removing the impact of the anomalous gates by reweighing. The reweighing is done such that the likelihood of the remaining benign gate profiles being drawn from the process variation and noise distribution after mapping to the benign space is maximized. Based on this criteria, the objective here is to select a subset of gates $\Gamma \subseteq \mathcal{D}$ for the linear program and reweigh them such that the likelihood $\mathcal{L}(\mathcal{D}\backslash\Gamma, \epsilon)$ is maximized (i.e., the gate profiles fall into the benign space for maximizing the P_D), subject to

the cost constraint $\mathcal{Q}(\Gamma)$ for selecting the O gates as Trojan (for minimizing the P_{FA}). Reweighing is done by setting the gate scaling factor to its nominal value of unity, assuming the systematic variations are calibrated. Let q_u denote the budget for the cost $\mathcal{Q}(\Gamma)$. Thus, one can write the objective function and the constraints of the problem as follows:

$$OF: \max_{\Gamma \subseteq \mathcal{D}} \mathcal{L}(\mathcal{D} \backslash \Gamma, \epsilon) \tag{4}$$
$$C's: \mathcal{Q}(\Gamma) \leq q_u; A\theta_{\{\theta \in [(\Gamma \subseteq \mathcal{D}) \cup (\Gamma = 1)]\}} = B_{\{\theta \in [(\Gamma \subseteq \mathcal{D}) \cup (\Gamma = 1)]\}} + \epsilon \quad .$$

The first constraint corresponds to the cost budget for the number of Trojans. The second constraint set corresponds to the gate level profiling discussed in the previous section (after reweighing the anomalies). Assuming the distribution of random process variations is Gaussian and the systematic process variations follows a 2D Gaussian in the spatial domain, the above likelihood function will always be lowered if we select to reweigh the maximum number of anomalies q_u. This is because the reweighing can make the noisier observations more consistent and therefore improve the likelihood results. But this is not usually desirable since it would unnecessarily increase the P_{FA}. Notice that the OF in Equation 4 has two simultaneous goals, one is to find the location of the gates that maximize the likelihood, and the other is to maximize the likelihood of the estimation error ϵ. Generally speaking, detecting guaranteed anomalies in problems like ours where there is an uncertainty about the value and interval of the variables (dependent on the other variables values) was demonstrated to be NP-hard [14]. Thus, we can only hope for heuristics and approximations to address the problem.

Iterative hierarchical detection and profiling. To simultaneously address the two goals embedded in Equation 4, we take a hierarchical approach for solving the problem by separation of concerns paradigm and iterative evaluations. We first present the high level view of this algorithm and then propose a class of formulations for which we can derive tight bounds on the solutions obtained by our approach. Our method is presented in Algorithm 1.

The procedure iteratively increases the maximum allowed number of anomalies q_u, starting from zero (Step 1). The stopping criteria of the iterative algorithm is improvement above a certain threshold (Step 2). For each added value of q_u (Step 3), we follow a greedy selection and add the most discrepant gate o to the set Γ (Step 4)). Discrepancy is evaluated as the distance to the projection into the benign gate space (Steps 5-7). A new round of gate level profiling is done after adding the newly reweighed gate o (Step 5). Since the derived gate profiles contain both systematic and random variations, calibration is performed to adjust for the systematic variations (Step 6). Now, the benign gates would only have random variations. The anomaly detection criteria is evaluated for checking the stopping condition for the algorithm (Step 7).

Algorithm 1 - Unimodal anomaly detection

Input. Combinational circuit, noninvasive measurements for J inputs, nominal technology values;

Output. Scaling factors (θ) from gate level profiling (GLP); anomalous gate set (Γ);

```
1  Set Γ = ∅, q_u = 0; Perform an initial GLP;
2  While (improvement by anomaly reweighing) do
3      q_u++;
4      Select the gate o to reweigh (Γ = Γ ∪ {o});
5      Perform a GLP with the reweighted o;
6      Calibrate for systematic variations;
7      Evaluate the improvement criteria;
```

The complexity of Algorithm 1 can be computed as follows. Let N_g denote the total number of gates in the circuit. Assuming that solving the linear system and calibration are at most of polynomial complexity, the worst case complexity of the above algorithm would still be polynomial and is effectively dominated by the time the solver takes to find the gate level profiles. The exact form of the GLP objective function would determine the solver time. For example, maximizing the likelihood for Gaussian distribution corresponds to solving a quadratic optimization problem in each round. The number of iterations is much less than the number of gates N_g since not all gates will be reweighed and the improvement criteria has a diminishing return property that would decrease at each iteration. If the Trojan is so large that many gates need to be reweighed, then the problem becomes trivial: it is well known in statistics that the anomaly detection is only challenging when the outlier characteristics only slightly differ from noise [14].

Greedy anomaly detection. For evaluating the improvement criteria for reweighing the gates in Γ, we propose a formulation of the anomaly detection objective based on likelihood improvement. This objective aims at performing *penalty reduction*. The penalty reduction metric quantifies the expected reward obtained by reweighing a set of gates. The expected penalty reduction due to reweighing the gates in the set Γ is denoted by $\mathcal{R}(\Gamma)$ and is defined as:

$$\mathcal{R}(\Gamma) = \mathcal{L}(\mathcal{D}) - \mathcal{L}(\mathcal{D} \backslash \Gamma) \quad . \tag{5}$$

The above formulation has a number of important properties that we exploit in our framework. A set function $\mathcal{R}$ is called *submodular* if it satisfies the following properties: (i) the penalty will not be reduced if we do not reweigh a new anomalous gate, i.e., $\mathcal{R}(\emptyset) = 0$; (ii) $\mathcal{R}$ is a nondecreasing set function and thus, reweighing a new anomaly could just decrease the associated penalty, i.e., $\mathcal{R}(\Gamma_1) \leq \mathcal{R}(\Gamma_2)$, for $\Gamma_1 \subseteq \Gamma_2 \subseteq \mathcal{D}$; (iii) the set function satisfies the *diminishing return property*: if we reweigh a gate in a smaller set of gates with logic relations (denoted by $\mathcal{D}_s$), we improve the reward by at least as much, as if we reweigh in a larger set of gates (denoted by $\mathcal{D}_l$) with logic relations such that $\mathcal{D}_l \subseteq \mathcal{D}_s$.

Nemhauser et al. [21] have shown that a function $\mathcal{R}$ is submodular if and only if the following theorem holds:

THEOREM 1. For all detected and reweighed Trojans $\Gamma_1 \subseteq \Gamma_2 \subseteq \mathcal{D}$ and a new candidate point $o \in \mathcal{D} \backslash \Gamma_2$ the following holds:

$$\mathcal{R}(\Gamma_1 \cup \{o\}) - \mathcal{R}(\Gamma_1) \geq \mathcal{R}(\Gamma_2 \cup \{o\}) - \mathcal{R}(\Gamma_2) \quad . \tag{6}$$

It can be shown that the reward function $\mathcal{R}$ satisfies the above theorem [15]. Now our optimization problem over Γ can be expressed as:

$$OF: \max_{\Gamma \subseteq \mathcal{D}} \mathcal{R}(\Gamma) \qquad C's: \mathcal{Q}(\Gamma) \leq q_u \quad .$$

Since solving the above problem has been shown to be NP-hard for the most interesting instances [9], we address the above optimization problem by the greedy procedure described in Algorithm 1. This is because a key result states that for submodular functions, the greedy algorithm achieves a constant factor approximation:

THEOREM 2 [21]. For any submodular function $\mathcal{R}$ that satisfies the above three properties, the set Γ_G obtained by the greedy algorithm achieves at least a constant fraction $(1 - 1/e)$ of the objective value obtained by the optimal solution, or,$\mathcal{R}(\Gamma_G) \geq (1 - 1/e) \max_{|\Gamma \leq q_u|} \mathcal{R}(\Gamma)$. Perhaps more surprisingly, Feige has shown that no polynomial time algorithm can provide a better approximation guarantees unless P=NP [9]. Thus, for any class of submodular objective functions, the proposed greedy selection algorithm results in the best achievable solution.

4.3 Calibration

To perform the anomaly detection, it is required that we calibrate for the systematic variations after profiling the gates. As mentioned in Section 3(i), the systematic variations consist of inter-chip and intra-chip variations. The inter-chip variations are simply affecting the mean of the variations and can be adjusted for by shifting the mean extracted profile values to have a mean of unity. The intra-chip variations are in form of a spatial distribution, e.g., 2D Gaussian in our model. The key observation is that the spatial rate of change of the neighboring gate level profiles due to the systematic intra-chip variations (spatial correlations) is slower than the rate of change because of the Trojan insertion. The larger Trojans that would affect many gates in a larger area are trivial to detect and would not be a challenge to address. This suggests using a high-pass filter over the 2D discrete space of the gate layouts for the identification of the sharp edges that have high frequency components in their frequency transformation. In this paper, we use the 2D Discrete Cosine Transform (DCT).

5 Multimodal Trojan Detection

The next step of our approach is to combine the results for anomalous gate detection over the M modalities. While there are a number of possible methods

to accomplish this task, our goal is to combine the unimodal methods to optimize the P_D and P_{FA} results. Assume that $\mathcal{C}_m(G_k)$ is the anomaly vote for gate G_k in modality m:

$$\mathcal{C}_m(G_k) = \begin{cases} 1 \text{ for } G_k \text{ anomalous in modality } m; \\ 0 \text{ otherwise.} \end{cases}$$

We propose four methods for combining the results of different modalities.

(i) Unanimous voting: In this voting approach, the Trojan gates are those that have been marked anomalous by all the M modalities. For example, for the three modalities the following constraint should hold for marking a gate as Trojan: $\mathcal{C}_T(G_k)+\mathcal{C}_\Phi(G_k)+\mathcal{C}_\Psi(G_k) = 3$ where the subscripts T, Φ, and Ψ denote the timing, quiescent current, and dynamic current measurement modalities respectively. This voting method is likely to decrease P_D but improves P_{FA}. It would also give the minimum achievable P_{FA} (lower bound) by any linear combination of the unimodal detection methods.

(ii) Conservative voting: A gate that has been marked anomalous by any of the modalities is marked as a Trojan by the conservative voting method. In our case, the following constraint is necessary and sufficient for marking a gate G_k as Trojan by conservative voting: $\mathcal{C}_T(G_k) + \mathcal{C}_\Phi(G_k) + \mathcal{C}_\Psi(G_k) \geq 1$. This voting method is likely to increase P_{FA} but also increases P_D. It would also give the maximum achievable P_D (upper bound) by any linear combination of our anomaly detection methods.

(iii) Majority voting: Here, the Trojan gates are those that have been marked anomaly by at least $1 + \lfloor \frac{M}{2} \rfloor$ of the modalities. In our case, the majority voting translates to the following condition: $\mathcal{C}_T(G_k) + \mathcal{C}_\Phi(G_k) + \mathcal{C}_\Psi(G_k) \geq 2$. This method provides a useful trade-off between the P_D and P_{FA} values.

(iv) Weighed voting: The voting methods above assume that all the modalities have the same detection ability. However, this is not true. For example in our experiments we see that there is less controllability/observability for the timing modality. For example, assume that we give the weights S_k^T, S_k^Φ, and S_k^Ψ for gate G_k for timing, leakage, and dynamic current respectively. Now, the votes of the three unimodal detectors over an anomalous gates are combined as follows: $S_k^T\mathcal{C}_T(G_k) + S_k^\Phi\mathcal{C}_\Phi(G_k) + S_k^\Psi\mathcal{C}_\Psi(G_k) \geq$ threshold. If this expression is true, the gate G_k is marked as the Trojan. Changing the detection threshold introduces a tradeoff between P_D and P_{FA} values.

6 Experimental Evaluations

6.1 Evaluation Set-Up

We evaluate the performance of the unimodal detection and the unified multimodal framework on the widely used MCNC benchmark suite. The ABC synthesis tool from UCB was used for mapping the benchmark circuits to NAND2, NAND3, NAND4, NOR2, NOR3, NOR4, and inverter library gates. Placing

the gates on the layout placement is done by the Dragon placement tool from UCLA. The gates have different sizes and they are located on irregular grids. The process variation model is as described in Section 3(i). In a number of our experiments, the amount of variations are altered and the detection results are tested against the variation fluctuations. In cases where we do not change the variations, the random variation is 12%, intra-die variation correlation is 60% of the total variation [5]; 20% of the total variation is uncorrelated intra-die variation and the remaining variation is allotted to the inter-die variation. The noninvasive measurement setup was described in Section 3(iii).

To find the values for timing, leakage, and dynamic currents for each of the library gate files over various input states, we performed HSPICE simulations for the 65nm CMOS transistor technology. The linear optimization was performed using the MATLAB optimization toolbox. Several other internal MATLAB functions are used for computing the likelihood and for filtering to calibrate the systematic variations. Each of our reported numbers and statistics are averaged over 100 runs of the random circuit instances.

6.2 Unimodal Trojan Detection

Gate level profiling. Finding the gate level profiles is the essence of the proposed approach. Table 1 shows gate level dynamic power, static power, and timing profiling results on benchmark circuits. For these evaluations the number of measurements is the minimum of double the number of gates and the maximum number of tests that can be done on the circuit. The first column shows the name of the benchmark denoted by *ct*. The second column shows the number of gates in the benchmark denoted by N_g. The number of primary inputs and primary outputs denoted by $\#i$ and $\#o$ are shown in the third and fourth columns respectively. The next nine columns show the profile estimation l_2 errors in the presence of 3%, 5%, and 10% measurement error for the three modalities respectively. On the average, the gate level profile estimation error is 3.8%, 4.8%, and 9.6% for the different measurment errors in case of dynamic power. The error in static power profiling is 4%, 6%, and 10% for 3%, 5%, and 10% of the measurement error respectively. For the timing modality the profiling error is 4%, 6%, and 11%.

Table 1. Dynamic power, static power, and timing profile estimation error for MCNC benchmark circuits

				D. Power			**S. Power**			**Timing**		
ct	N_g	**#i**	**#o**	**3%**	**5%**	**10%**	**3%**	**5%**	**10%**	**3%**	**5%**	**10%**
C1355	512	41	32	7.8	9.1	11.5	8.5	10	12	4	8	12.3
c8	165	28	18	4.2	6.4	11.2	5.6	7	11.6	5.3	7	11.5
C3450	1131	50	22	3.5	6	9.5	4	5.9	9.8	2.9	4.1	9.2
C432	206	36	7	1.5	3.1	6.9	1.7	3.5	7.2	3.8	5.4	10.1
C499	532	41	32	2.2	4.2	8.8	2.9	4.5	9	5	6.5	12

Unimodal Anomaly detection. In this Section, we evaluate the performance of the unimodal anomaly detection over the three modalities under study. We study three scenarios: (i) a Trojan-free circuit, (ii) one extra NAND2 gate inserted as a Trojan, and (iii) a 3-gate comparator circuit is inserted. The Trojans are inserted in the empty spaces within the automatic layout generated by the Dragon tool. An important property for unimodal anomaly detection is how the diminishing return property changes. As we discussed earlier, this function should be monotonically decreasing assuming no random perturbations. We have tested the validity of our assumption on multiple benchmark circuits. An example result is shown in Figure 3 for the leakage modality for the C432 benchmark and 100 measurements. The results for the other two modalities are similar.

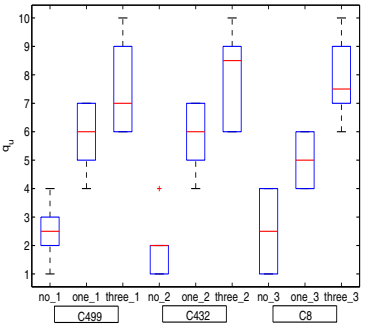

Fig. 2. Boxplots of q_u for Trojan free, 1 Trojan, and 3 Trojan gates

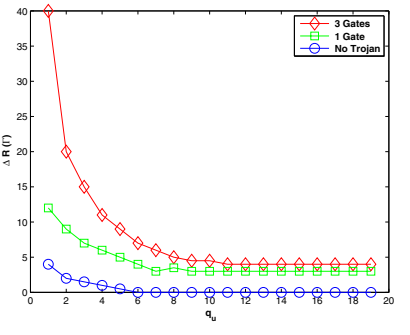

Fig. 3. The stepwise diminishing return improvement for leakage modality

As we can clearly see on the figure, the difference between diminishing returns of two consecutive steps of our algorithm is much higher for the larger Trojan, it becomes lower for the smaller Trojan and it is really low for the Trojan free case. The same phenomena is observed over all the modalities. This result is really important for adjusting the stopping criteria of our algorithm. Basically, when the Trojan circuit reaches the same diminishing return difference as the Trojan-free circuit, no further significant improvement is foreseen. The above results gave us an insight to set the stopping criteria for Algorithm 1. For example, we set it such that the diminishing return is decreased by more than 2 in each step. As can be seen on the Figure 3, this decision would result in an average 1 gate false alarm for this circuit with 206 gates, i.e., $P_{FA} = 1/206 = 0.5\%$ and about 2 gates are not detected in case of the smaller Trojan, about 1%. It should be noted that we detect more anomalous gates than what is inserted by the Trojan because the Trojan gates affect the side channel characteristics of the logically connected gates. This result can be used to help localize the Trojan, however, this is out of the scope of this work. Figure 2 shows the boxplot of the number of iterations (q_u) before our stopping criteria is reached for 100 runs over 3 benchmarks for leakage modality. The number of iterations corresponds

Table 2. The number of gates giving false alarm in a non Trojan circuit

ct	UNA	CON	MAJ	WD
C1355	0/512	4/512	2/512	2/512
c8	0/165	3/165	1/165	2/165
C3450	0/1131	3/1131	3/1131	3/1131
C432	1/206	2/206	1/206	1/206
C499	0/532	3/532	1/532	1/532

to the number of detected anomalous gates. We see that the number of detected anomalous gates increases as the Trojan size increases.

Table 2 shows the P_{FA} (in terms of number of gates giving false alarm in the circuit) for 100 chips with no Trojan. The false negatives ranged from 0 to 4 gates and are largest for the timing modality. The first column shows the name of the benchmark. The rest of the columns show the results for different multimodal methods: unanimous (UNA), conservative (CON), majority (MAJ), and weighed (WD). The unanimous voting performs best in terms of P_{FA}. This is because it can filter out the effect of the modalities that give more false positives.

We also studied the probability of detection P_D using the different voting methods. Our P_D results demonstrate an average of 86%, 99%, 98%, and 98% for unanimous, conservative, majority, and weighted voting respectively. The unanimous voting yields the worst P_D while as we described earlier, it resulted in the best P_{FA}. The conservative voting yields the best P_D at the expense of worsening the false alarm probability P_{FA}. On the other hand, the majority voting and weighed voting result in a good trade-off between the two probabilities. In addition, we observed that the weighted voting gives the best result when we assign the lowest weight to the timing modality. The inefficiency of timing modality is because the small Trojans would only affect a few of the tested timing paths, whereas many more sets of current tests would show the impact of the modified currents. The two power modalities are much more effective in detecting Trojans. Another interesting observation was that even though there is a good amount of independent information in the static and dynamic current tests, the outcomes of the two testing modalities demonstrate an average of 73% correlations on our benchmark circuits.

7 Conclusion

Our work presents a new unified formal framework for IC Trojan detection by noninvasive measurements from multiple test modalities. For each modality, a unimodal anomaly detection is built upon the gate level profiling. Since the problem is extremely complex, we devise an iterative detection and profiling method. Our objective function for detecting the abnormal gate level behavior is shown to be submodular. Because of the objective submodularity, our iterative greedy detection and profiling algorithm achieves a near optimal solution (within a constant fraction of the optimal) in polynomial time. We show a method to

calibrate the systematic variations. Our multimodal Trojan detection approach combines the unimodal detection results using a number of different techniques. Experimental evaluations on benchmark circuits using timing, leakage current, and transient currents show the effectiveness of our approach.

References

1. Agrawal, D., Baktir, S., Karakoyunlu, D., Rohatgi, P., Sunar, B.: Trojan detection using ic fingerprinting. In: S&P, pp. 296–310 (2007)
2. Alkabani, Y., Koushanfar, F.: Consistency-based characterization for ic trojan detection. In: ICCAD, pp. 123–127 (2009)
3. Banga, M., Chandrasekar, M., Fang, L., Hsiao, M.: Guided test generation for isolation and detection of embedded trojans in *ICs*. In: GLS-VLSI, pp. 363–366 (2008)
4. Banga, M., Hsiao, M.: A region based approach for the identification of hardware trojans. In: HOST, pp. 43–50 (2008)
5. Cao, Y., Clark, L.T.: Mapping statistical process variations toward circuit performance variability: an analytical modeling approach. In: DAC, pp. 658–663 (2005)
6. Chekuri, C., Pal, M.: A recursive greedy algorithm for walks in directed graphs. In: FOCS, pp. 245–253 (2005)
7. Das, A., Kempe, D.: Algorithms for subset selection in linear regression. In: STOC, pp. 45–54 (2008)
8. Defense Science Board (DSB) study on High Performance Microchip Supply (2005), `http://www.acq.osd.mil/dsb/reports/2005-02-HPMS_Report_Final.pdf`
9. Feige, U.: A threshold of ln n for approximating set cover. Journal of ACM 45(4), 634–652 (1998)
10. Jha, N., Gupta, S.: Testing of Digital Systems. Cambridge University Press, Cambridge (2003)
11. Jin, Y., Makris, Y.: Hardware trojan detection using path delay fingerprint. In: HOST, pp. 51–57 (2008)
12. Koushanfar, F., Boufounos, P., Shamsi, D.: Post-silicon timing characterization by compressed sensing. In: ICCAD, pp. 185–189 (2008)
13. Krause, A., Guestrin, C.: Near-optimal observation selection using submodular functions. In: AAAI, pp. 1650–1654 (2007)
14. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations. Kluwer Academic Publishers, Dordrecht (1997)
15. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: SIGKDD, pp. 420–429 (2007)
16. Li, J., Lach, J.: At-speed delay characterization for IC authentication and trojan horse detection. In: HOST, pp. 8–14 (2008)
17. Liu, F.: A general framework for spatial correlation modeling in VLSI design. In: DAC, pp. 817–822 (2007)
18. Mossel, E., Roch, S.: On the submodularity of influence in social networks. In: STOC, pp. 128–134 (2007)
19. Murakami, A., Kajihara, S., Sasao, T., Pomeranz, I., Reddy, S.M.: A test structure for characterizing local device mismatches. In: ITC, p. 376 (2000)
20. Nelson, M., Nahapetian, A., Koushanfar, F., Potkonjak, M.: Svd-based ghost circuitry detection. In: Katzenbeisser, S., Sadeghi, A.-R. (eds.) IH 2009. LNCS, vol. 5806, pp. 221–234. Springer, Heidelberg (2009)

21. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of the approximations for maximizing submodular set functions. Math. Programming 14, 265–294 (1978)
22. Potkonjak, M., Nahapetian, A., Nelson, M., Massey, T.: Hardware trojan horse detection using gate-level characterization. In: DAC, pp. 688–693 (2009)
23. Rad, R., Plusquellic, J., Tehranipoor, M.: A sensitivity analysis of power signal methods for detecting hardware trojans under real process and environmental conditions. IEEE Trans. on Very Large Scale Integration (VLSI) Systems 99 (2009)
24. Rad, R., Wang, X., Plusquellic, J., Tehranipoor, M.: Power supply signal calibration techniques for improving detection resolution to hardware trojans. In: ICCAD, pp. 632–639 (2008)
25. Chakravarty, S., Thadikaran, P.: Simulation and generation of iddq tests for bridging faults in combinational circuits. IEEE Trans. on Computers 45(10), 1131–1140 (1996)
26. Sabade, S., Walker, D.: IDDX-based test methods: A survey. ACM Trans. Design Automation of Electronic Systems 9(2), 159–198 (2004)
27. Srivastava, A., Sylvester, D., Blaauw, D.: Statistical Analysis and Optimization for VLSI: Timing and Power. Springer, Heidelberg (2005)
28. Wei, S., Meguerdichian, S., Potkonjak, M.: Gate-level characterization: Foundations and hardware security applications. In: DAC (2010)
29. Yang, K., Cheng, K.T., Wang, L.: TranGen: A SAT-based ATPG for path-oriented transition faults. In: ASPDAC, pp. 92–97 (2004)