A Unified Framework for Multimodal Submodular Integrated Circuits Trojan Detection

Farinaz Koushanfar, *Member, IEEE*, and Azalia Mirhoseini, *Student Member, IEEE*

***Abstract*—This paper presents a unified formal framework for integrated circuits (ICs) Trojan detection that can simultaneously employ multiple noninvasive side-channel measurement types (modalities). After formally defining the IC Trojan detection for each side-channel measurement and analyzing the complexity, we devise a new *submodular* formulation of the problem objective function. 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 interchip and intrachip correlations are presented. We define a new *sensitivity metric* that formally quantifies the impact of modifications to each existing gate that is affected by Trojan. Using the new metric, we compare the Trojan detection capability of different measurement types for static (quiescent) current, dynamic (transient) current, and timing (delay) side-channel measurements. We propose four methods for combining the detection results that are gained from different measurement modalities and show how the sensitivity results can be used for a systematic combining of the detection results. 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.**

***Index Terms*—Change detection algorithms, circuit Trojan detection, gate-level characterization, hardware malware detection, hardware security and trust, submodular functions, timing/power tests.**

I. Introduction

THE prohibitive cost of manufacturing integrated circuits (ICs) in nano-meter scales has made the use of contract foundries the dominant semiconductor business practice. Unauthorized intellectual property (IP) usage, IC overbuilding, and insertion of additional malware circuitry (a.k.a., *Trojans*) are a few of the major threats facing the horizontal IC industry where the IP providers, designers, and foundries are separate entities [1]–[3]. Since the ICs form the core computing and communication kernels for the governments, defense, and industries today, ensuring IC trust in the presence of an untrusted foundry is of paramount importance. The Trojan embedder 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 (functionally activated) as needed.

A standing challenge for noninvasive side-channel IC testing and Trojan detection is dealing with the increasing complexity and scale of the state-of-the-art technology. Internals of the complex chips are inherently opaque. Scaling to the physical device limitations and mask imprecisions cause nondeterminism in chip characteristics. It is hard to distinguish between the characteristic deviations because of process variations and alterations due to Trojan insertion. What complicates the problem even more is the large space of possible Trojan attacks by potentially advanced adversaries. Very little is known or documented about IC Trojan attacks. Because the functional triggering of the Trojans may be hidden [2], the logic-based testing methods are unlikely to trigger and distinguish 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.

Recently a number of effective methods for IC Trojan detection were proposed. Comprehensive reviews can be found in [2] and [3]. Several authors used the power supply transient current signal for Trojan detection [4]–[7]. Using the timing signal signature by testing multiple path signatures was suggested [8]. Gate level characterization for Trojan detection was proposed [9]–[13] but no systematic formal analysis or optimal algorithms were discussed. The available methods are either based on ad-hoc measurements, heuristics for detection, calibration and test, or they use costly and slow destructive tests. No systematic method for IC malware detection, combining multiple side-channel modalities with optimality guarantees, or a mathematical formulation of calibration is presently available.

Our results are complementary to the existing literature in Trojan detection using side-channels. Essentially, using submodularity, for any given set of test vectors we formally demonstrate: 1) the best polynomial-time detection algorithm with constant factor approximation guarantee compared to the theoretically optimal achievable solution; and 2) an upper bound for the optimal detection metric, which could be used for bounding the quality of detection metric by any other (polynomial or nonpolynomial) heuristic detection methods, which may perform better than our constant factor polynomial-time approximation. Note that our bounds can be converted to any combinations of circuit components that are evaluated by other (i.e., not gate-level) detection methods.

Manuscript received July 22, 2010; revised September 24, 2010; accepted November 09, 2010. Date of publication December 03, 2010; date of current version February 16, 2011. This work was supported in part by the Office of Naval Research YIP (Grant R16480) and in part by the National Science Foundation CAREER award (Grant R3A530). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ramesh Karri.

The authors are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: farinaz@rice.edu; azalia@rice.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2010.2096811

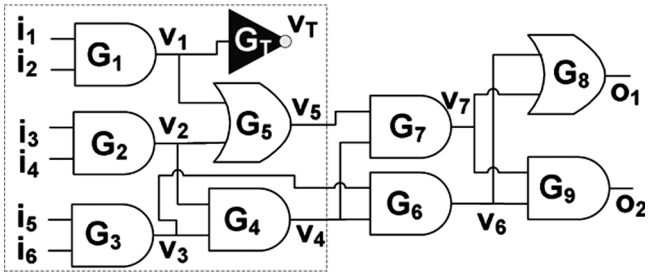

Fig. 1. Submodular property.

A. Motivational Example

Let us demonstrate the methods using a small example shown in Fig. 1. The design consists of nine gates $G_1, \ldots, G_9$, six inputs, and two outputs. A Trojan gate G_T that has an impact on the side-channel is added to the circuit. Consider a subcircuit of this design composed of gates $G_1, \ldots, G_5$ in the dotted square that also includes G_T.

We consider the cases where we are given a set of input vectors to test a side channel in the circuit. In this example, the side-channel is the circuit's static current (IDDQ testing). For each input vector, the total current drawn from the circuit is the sum of the individual gate leakages (assuming that we have lumped the wire and parasitic leakages into a gate leakage model). Now, for one input vector applied to the circuit, the ratio of current leaked by G_T to the rest of the circuit leakage would be higher for the subcircuit compared to the whole circuit. This behavior of the Trojan side-channel (i.e., a more significance on smaller subcircuits) can be abstracted by a submodular objective function which aims at minimizing the detection error. By exploiting theoretical results for submodular function detection, we propose a near-optimal polynomial Trojan detection algorithm for the given set of test vectors.

B. Contributions

Our contributions are as follows:

1) A new unified noninvasive multimodal Trojan detection framework is proposed. The framework translates the abnormal behavior disclosed by different measurement modalities to the gate level profiles and analyzes the results. For each modality, we formulate the optimization problem for simultaneous gate level profiling and Trojan detection and show the problem is NP-hard.
2) We demonstrate a formulation of the unimodal Trojan detection objective as a *submodular* function. The objective function submodularity is exploited for devising an iterative polynomial-time optimization algorithm that achieves a near-optimal unimodal detection (within a constant fraction of the optimal solution) for NP-hard detection. Our solution also defines an upper bound for quality of detection (i.e., lower bound for detection error) by any other non-polynomial detection methods.
3) A new calibration method for mitigating the impact of interchip and intrachip process variations is introduced.
4) We discuss how the cumulative statistics and detection results can be used for classifying ICs based on the Trojan symptoms and for speeding-up detection.
5) We introduce a new *sensitivity metric* formally quantifying the change in system response for each gate's change.
6) We devise and compare four methods for combining the results of multiple unimodal detections based on their probability of detection (P_D) and the probability of false alarm (P_{FA}). The effectiveness of the new methods are confirmed by extensive evaluations in presence of process variations and measurement noise on benchmark circuits.

II. Preliminaries

In this section, we provide the necessary background to make the paper self-contained.

A. Process Variations

As CMOS dimensions shrink, uncertainty and variation in device characteristics compared to their nominal values increases. CMOS circuits exhibit a high variability in both delay and power consumption that monotonically increases with scaling. In controlled settings, i.e., the same voltage, temperature and light levels as the simulation models, the dominant source of difference between chips is spatial variation [14]. Spatial variation may be intradie, or interdie, and could be systematic or random. In this paper, we use the Gaussian variation models [14]. Our approach works for the stationary process variation models.

B. 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. The measurements are considered to be stationary with an i.i.d. Gaussian noise. A Trojan causes a change in the statistical distribution of gate characteristics. Intrachip variations are assumed to have a lower amplitude when compared to Trojan impact. We call the gates with modified characteristics anomalous gates. In our case, the nominal values for gate characteristics are extracted from technology simulation files needed for design-time power estimation and timing closure. This is a standard method for finding the nominal values and was used in a number of earlier works [6]. Our nominal gate-level values could contain the side-channel value for the gate (e.g., delay) lumped together with the other side-channel parasitics and wire delays that are between the two circuit nodes [15].

Our detection framework can detect a Trojan as long as it affects the side-channel value beyond the noise level. For example, if a Trojan is power-gated, none of the power-based side-channel Trojan detection methods would be able to find it because its current cannot be measured. Since we assume the test vectors are given, a Trojan that is never sensitized by the available test vector set cannot be detected. We emphasize that this paper's contribution is not in providing a new set of test vectors that can sensitize rare Trojans, but in finding the best achievable detection performance by a given set of test vectors.

C. Sequential versus Combinational Circuits

This paper discusses detection of the Trojans for combinational circuits but we emphasize that sequential circuit Trojan detection can also be addressed by our approach. It is well-

known that high coverage tests of the sequential circuits are not feasible in practice, unless scan chains are used. Since one cannot usually control the present state lines of a sequential circuit nor directly observe the next state lines, a sequence for setting the circuit to the desired state is required [16]. In the test scan-chain mode, all flip flops (FFs) form a shift register so that the test input for the combinational logic part between two FFs can be scanned in (shifted in) and the output is scanned out (shifted out). Scan-chains are an essential part of conventional sequential designs today. From the point view of testing by supplying the input vectors and measuring the outputs, the scan-in and scan-out FFs can be considered the input and output pins. We can apply our test vectors by scanning them in to the target FFs and scanning out from the proper FFs. The combinational circuitry in between the scan-in and scan-out FFs can be tested by our method for detecting the Trojans. Therefore, our method can be scaled to larger sequential designs that are normally equipped with scan chains for conventional testing purposes.

D. Measurement Test Vectors

We use available test vector set generation tools for Trojan detection purposes. For timing measurements, we exploit a known delay-fault test and validation technique [17]. Extensive previous research in delay test vector generation has demonstrated that it is possible to generate a set of path test vectors that can sensitize all the gates (as much as the controllability and observability allows) [16]. It has been shown that the number of sensitized paths (testable path set) is the same order as the number of gates [15].

The leakage current can be measured via the commonly known IDDQ test methods often done via the off-chip pins by the precision measurement unit (PMU) [18]. The dynamic current tests are referred to as IDDT tests that can be done by averaging methods that do not require high precision or high frequency measurement devices [16]. Our Trojan detection method attempts at finding the current deviation at the gate-level where the Trojan impact modifies a set of gate currents. The mechanisms for testing all the gates' states are similar to the bridging/leakage fault testing [18].

A comprehensive survey of IDDT and IDDQ testing techniques can be found in [18]. In this work, we use available methods for test vector generation for current-based tests [19]. Test application time and power for the set of given test vectors could be reduced by conventional methods including test vector ordering and continuous scanning. We emphasize that the novelty of this work is not in introducing a high coverage test method, but to find the best achievable solution given a set of test input vectors.

III. Unified Framework

Given an IC, the original layout, and GDS-II (the design file submitted to foundry), a set of postsilicon, noninvasive, and nondestructive measurements for each modality m, $m = 1, \ldots, M$, where each measurement is taken over an input vector or for a transition between two input vectors,

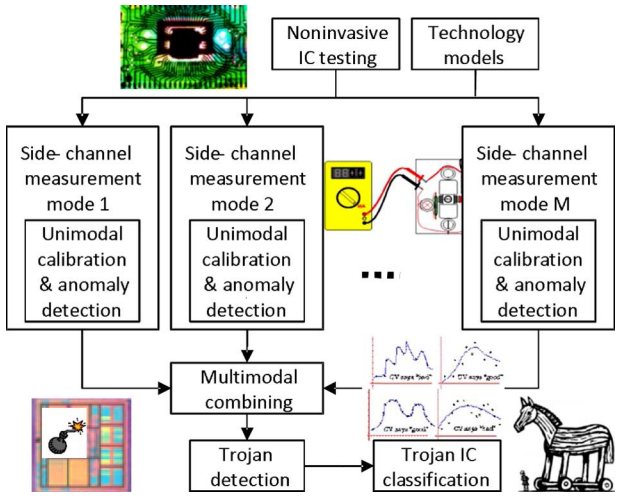

Fig. 2. Unified multimodal Trojan detection framework.

the goal is to identify the abnormal components on the chips postmanufacturing and packaging.

We introduce a unified format of Trojan detection problem that can be applied to any modality that measures a parametric function of the gates or other component characteristics on a chip. Fig. 2 presents the overall IC Trojan detection framework introduced in this paper. The gate level anomalies are detected for each measurement modality. After that, the decisions by the different modalities on each gate are fused together by using various combining methods. A formal sensitivity metric is introduced and utilized to quantify the impact of potential changes for each measurement modality. The multiple ICs are also classified for finding the IC groups that are modified in a similar way. The remainder of the paper discusses the details of the unimodal and multimodal malware (Trojan) detection.

The focus of this paper is on evaluation of the gate profiles by using three important externally observed measurement modalities of delay (T), quiescent current (IDDQ), and dynamic current (IDDT), but we emphasize that the proposed framework is generic and can be used for other features such as electromagnetic emanation measurements. In this paper, we do not consider the interconnect delay and power consumption. Since the wire impact is linearly added to the path delay and power consumption, it can be integrated in our framework in a straightforward way.

IV. Unimodal Trojan Detection

The basis of the unimodal Trojan approach is the gate profiling discussed in Section IV-A. In Section IV-B, we formulate the detection problem in a unified format regardless of the modality and further discuss the complex structure of the general NP-complete unimodal detection problem. We opt to use our prior knowledge about the process variations and submodularity to address the problem and to gain a near optimal solution.

The precursor for our hierarchical method is systematic calibration that is discussed in Section IV-C. We define a sensitivity metric for Trojan detection in Section IV-D. We also perform cumulative unimodal profiling for Trojan detection in Section IV-E.

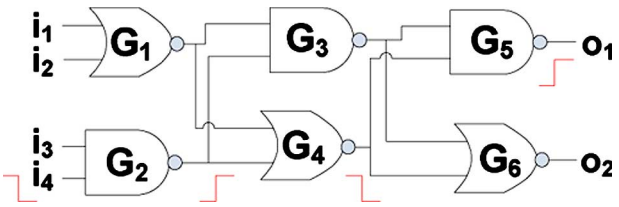

Fig. 3. Example circuit. Sensitizing a path by changing the input from 1111 to 1110.

A. Gate Profiling

In this subsection, we show how side-channel measurements can be decomposed to their gate level components postsilicon. One can exploit the linear relationship between the IC's gate level profile and the side-channel measurements (constrained by logic relations) to estimate the gate level characteristics. Note that even though we do not consider the wire effect, the impact can be linearly added to our framework. We introduce a formal framework for this problem.

Problem: Unimodal Gate Profiling

Given: A combinational circuit $\mathbf{C}$, 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 a 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 primary inputs $x_1,\ldots,x_{N_I}$ which changes the states of 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 gate characteristics for the modality M and a measurement error.

Objective: Estimate the postsilicon profile of each individual gate for the modality M.

Generating the input patterns that can distinctively identify each gate's characteristics is known to be NP-complete [16]. Although we are limited by the same constraints as testing in terms of gate coverage, the difference 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.

The details of the generic problem above are slightly different for each measurement modality. Perhaps the best way to demonstrate the details of postsilicon characterization for one of the measurement modalities employed in this paper (i.e., delay) is by an example shown in Fig. 3. The design consists of six gates $(G_1,\ldots,G_6)$, four inputs (x_1,x_2,x_3,x_4), and two outputs (z_1,z_2).

1) 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. Assume that the input vector transitions from $(v_1;v_2;v_3;v_4)=1111$ to 1110. This input vector sensitizes the I/O path from v_4 to z_1, denoted by P_1. $D(P_1)$ shows the delay of the sensitized path P_1 and can be written as the sum of the low to high delay transition at G_2, high to low delay at G_4, and low to high delay at G_5 (denoted by $T_{\mathrm{LH}}(G_2)$, $T_{\mathrm{HL}}(G_4)$, and $T_{\mathrm{LH}}(G_5)$ respectively). We assume that the delay of a gate $G(k_j)$ is the same for both low to high and high to low transitions and is denoted by $T(G_{k_j})$, but it is worth noting here that adding both transition sides is just a matter of introducing two variables for each gate and following the same steps.

Similarly, one can test J different paths and write a linear system of delay equations. Noninvasive gate profiling aims at finding the gate delay values in presence of measurement error. If measuring the path delay $T_{\mathrm{meas}}(P_j)$ incurs the error $\epsilon_T(P_j)$, the optimization problem objective function (OF) and constraints (C's) can be written as follows:

$$\begin{aligned}
&\mathrm{OF}: \quad \min_{1\le j\le J} \mathcal{F}(\epsilon_T(P_j)) \\
&\mathrm{C's}: \quad \Sigma_{k_j=1}^{K_j} T(G_{k_j}) = T_{\mathrm{meas}}(P_j) + \epsilon(P_j) \\
&\qquad\quad P_j = \{G_{k_j}\}_1^{K_j}, \quad 1\le j\le J
\end{aligned} \tag{1}$$

where $\mathcal{F}$ is a metric for quantifying the measurement errors; the commonly used form of $\mathcal{F}$ is the l_2 norm of errors.

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 $T^{\mathrm{nom}}(G_k)$ and the deviation from nominal for G_k for the chip under measurement is $\theta_T(G_k)$, then $T(G_k)=\theta_T(G_k)T^{\mathrm{nom}}(G_k)$ 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.

Similar methods can be applied to linearly relate the overall measured quiescent and transient current modalities to gate-level values. For the sake of brevity, we refer the readers to earlier literature [10], [20].

B. Unified Detection Formulation

In Section IV-A, we mentioned that each of the modalities can be written in a unified format of a system of linear equations. We showed the detailed linear equations for timing modality. In the remainder of the paper, we use the following generic notations for gate profiling over different modalities:

$$\mathrm{OF}: \quad \min \mathcal{F}(\epsilon) \mathrm{C's}: \quad A\theta = B + \epsilon \tag{2}$$

where in matrix $A_{[J\times N_G]}$, elements of each row are the gates' nominal profile values of the input corresponding to that row. $\theta_{[N_G]}$ is the vector of unknown scaling factors, $B_{[J]}$ is the vector

of measurement values, and $\epsilon_{[J]}$ is the corresponding vector of measurement errors.

Solving the above linear system, for $|\epsilon| < 0.5\theta_n$ ($\theta_n \in \theta$ for $1 \leq n \leq N_G$) would require less than $2N_G$ constraints [21]. Since in circuit testing scenarios the cost is often dominated by the number of measurements taken from the device, the overall cost of performing our Trojan detection method is $O(N_G)$ measurements. Note that for a given set of test vectors, preprocessing could be used for selecting a limited number of linearly independent test inputs that can be applied to the chips to minimize the test time and cost.

Since the measured gate profile (delay, dynamic, or static power) is an additive function of the circuit components, and the Trojan is not known to the simulation models, the impact of an inserted Trojan would change the estimates of the nearby gate profiles.

Problem: Trojan Detection by Gate Profile Estimation

Given: The same inputs as Section IV-A.

Objective: Estimate the profile of each gate on the IC and identify the anomalous gates based on abnormal values.

An abnormal gate is the one which has large deviations in its measured characteristics compared with its nominal value. Thus, an abnormal gate would have a scaling factor largely deviating from its expected value 1. To address the above problem, we form an optimization problem that attempts at finding the set of anomalous gates and remove their negative impact to achieve the maximum likelihood for the estimation error. Removing the negative impact of an abnormal gate G_k (with scaling factor largely deviating from 1) means reweighing the gate's scaling factor (θ_k) by setting it to 1.

The nature of the estimation problems in the presence of errors is such that removing the impact of anomalous gates by reweighing improves the overall objective function since it improves the estimation error. We use the notation $\mathcal{G}$ to refer to the set of all N_G gates in the original design and we also define a set Λ which contains the anomalous gates.

Assume that there is a penalty associated with selecting the gates in Λ as the Trojan set denoted by $\mathcal{T}(\Lambda)$, where the maximum allowable penalty is $\mathcal{T}_{\max}$. The penalty is defined for keeping the probability of false alarms (P_{FA}) low, because, the global objective of Trojan detection is both to maximize the probability of Trojan detection (P_D) and to minimize the probability of false alarm (P_{FA}). Using the unified format, the objective function and the constraints of the problem can be defined as follows:

$$\begin{aligned} \mathrm{OF}: &\quad \min \mathcal{F}(\epsilon) \quad \text{for} \quad \mathcal{G}\backslash\Lambda \\ \mathrm{C's}: &\quad A\theta = B + \epsilon \quad \text{for} \quad \mathcal{G}\backslash\Lambda, \\ &\quad \mathcal{T}(\Lambda) \leq \mathcal{T}_{\max}. \end{aligned} \tag{3}$$

The OF and the first constraint set are the same as before, but this time only defined over the set of nonanomalous (benign) gates in $\mathcal{G}\backslash\Lambda$ after removal of the anomalous gates. The last constraint is the cost for selecting the set Λ. Notice that the OF in (3) has two simultaneous goals; one is to find the location of the gates in Λ, and the other is to minimize 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 [22]. Thus, we can only hope for heuristics and approximations to address the problem.

1) Objective Submodularity: Modifications to existing gate to address the optimization in 3, we propose a new form for the detection objective function. Our new OF is denoted by $\mathcal{R}$ and is called the *reward function*. $\mathcal{R}$ quantifies the expected benefit from reweighing the set of anomalous gates in Λ: $\mathcal{R}(\Lambda) = l_2(\mathcal{G}) - l_2(\mathcal{G}\backslash\Lambda)$. Thus, instead of optimizing the detection error, we study how much the detection error can be improved by removing the anomalous components.

The key property of the function $\mathcal{R}$ is its *submodularity*. A function $\mathcal{R}$ defined over a set is submodular if it has the following three properties:

1) $\mathcal{R}(\emptyset) = 0$, meaning that there is no improvement in reward, if we do not reweigh any anomalies.
2) $\mathcal{R}$ is nondecreasing, for $\Lambda_1 \subseteq \Lambda_2 \subseteq \mathcal{G}$, $\mathcal{R}(\Lambda_1) \leq \mathcal{R}(\Lambda_2)$. Thus, reweighing a new anomaly always improves the associated reward.
3) $\mathcal{R}$ satisfies the *diminishing return* property that considers the gate estimation problem over two sets of gates $\mathcal{G}_1$ and $\mathcal{G}_2$ where $\mathcal{G}_1 \subseteq \mathcal{G}_2$. Assuming that the linear constraints in $\mathcal{G}_1$ are a subset of the linear constraints in $\mathcal{G}_2$ (i.e., $\mathcal{G}_1$ is a subcircuit of $\mathcal{G}_2$), reweighing the subset of gates $\Lambda \subseteq \mathcal{G}_1$ would improve the reward function over the subcircuit $\mathcal{G}_1$ by at least as much as reweighing the gates in Λ for the larger set $\mathcal{G}_2$.

The first and second property of the function $\mathcal{R}$ can be concluded from the fact that reweighing a new anomaly would decrease the l_2 estimation error since reweighing in essence changes the l_2 distance of the outlier measurement to the estimated values. The third property is satisfied because the absolute (therefore positive) error values of the gates are added to compute the reward function. Thus, after reweighing a same subset of gates, the reward function improvement for $\mathcal{G}_2$ which has a larger number of gates cannot be more than that of $\mathcal{G}_1$. Note that it has been proven that a reward set function $\mathcal{R}$ is submodular if and only if it satisfies the Theorem below [23].

Theorem 1: For all reweighed Trojans $\Lambda_1 \subseteq \Lambda_2 \subseteq \mathcal{G}$, for a candidate anomalous gate $G_k \in \mathcal{G}\backslash\Lambda_2$, the following holds: $\mathcal{R}(\Lambda_1 \cup \{G_k\}) - \mathcal{R}(\Lambda_1) \geq \mathcal{R}(\Lambda_2 \cup \{G_k\}) - \mathcal{R}(\Lambda_2)$.

Using the above theorem (submodular property of the transformed objective $\mathcal{R}$), the OF in (Problem 3) can be reformulated as: $\max_{\Lambda \subseteq \mathcal{G}} \mathcal{R}(\Lambda)$.

Perhaps not surprisingly, addressing the above optimization problem was shown to be NP-complete as well [24]. However, we address the above optimization problem by the greedy procedure that will be 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 [23]: For any submodular function $\mathcal{R}$ that satisfies the above three properties, the set Λ_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,

$$\mathcal{R}(\Lambda_G) \geq \left(1 - \frac{1}{e}\right) \max_{|\Lambda| \leq |\mathcal{T}_{\max}|} \mathcal{R}(\Lambda). \tag{4}$$

Moreover, Feige has proven that under the assumptions above, no algorithm of polynomial time complexity could provide a better approximation guarantee than the greedy algorithm, unless the NP-complete problem could be solved in polynomial time by an algorithm [24]. Note that the given bound above has two applications. One is to guarantee the minimum performance of $(1-1/e)$ compared to the optimal solution. The other one is to give a bound for the best achievable solution for other heuristic approaches that might perform better than the greedy algorithm.

2) Detection Algorithm: Using the submodular property of the reward function $\mathcal{R}$, we propose a greedy algorithm for addressing the Trojan detection problem formulated in (3). The linear penalty function for selecting the gates in Λ as anomalies is set to be Λ's cardinality, i.e., $\mathcal{T}(\Lambda) = |\Lambda|$. The details of the greedy algorithm are shown in Algorithm 1. Recall that the inputs to the problem are the combinational circuit, noninvasive leakage measurements for J input vectors, the nominal gate leakage values, the minimum required improvement in reward $\Delta_{\min}$, and the maximum number of allowed anomalous gates $\mathcal{T}_{\max}$. The outputs of the algorithm are the gate leakages in form of scaling factors (θ) and the set of anomalous gates (Λ).

Algorithm 1 : Trojan Detection

1 Set $\Lambda = \emptyset$, $|\Lambda| = 0$;

2 Use the inputs to estimate the gate leakages;

3 Calibrate the scaling factors for interchip and intrachip leakage correlations;

4 While ($\Delta\mathcal{R} > \Delta_{\min}$ or $|\Lambda| \leq \mathcal{T}_{\max}$)

5 Select the gate G_k with the highest l_2 error;

6 Reweigh G_k and add G_k to Λ;

7 Re-estimate the gate leakages;

8 increase $|\Lambda|$ by 1;

9 Find $\Delta\mathcal{R}$;

In Step 1, the set of anomalous gates is initiated to be an empty set and the number of detected Trojans is initially set to 0. The inputs are used for finding an initial estimate of the gate leakages as discussed in Section IV-A in Step 2. Step 3 calibrates the gate leakage scaling factors for interchip and intrachip correlations. The details of our calibration method is presented in Section IV-C. The stopping criteria for the Trojan detection algorithm is evaluated in Step 4: the algorithm stops when the difference in reward function compared to the previous step $\Delta\mathcal{R}$ is less than a minimum predefined value $\Delta_{\min}$, or the number of detected anomalous gates $(|\Lambda|)$ is more than the maximum set by the penalty criterion $\mathcal{T}_{\max}$. Steps 5–9 contain one iteration of the algorithm, where at each iteration the gate G_k with the highest l_2 distance to the estimation is selected as anomaly, reweighed, and added to Λ. After G_k is reweighed, the gate leakages will be re-estimated, and the number of anomalies T is increased by 1, and the function $\Delta\mathcal{R}$ is calculated for checking the stopping criteria.

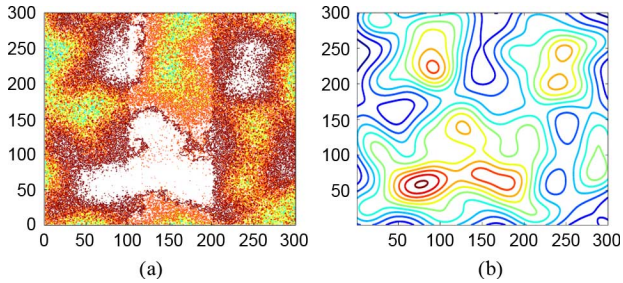

Fig. 4. (a) A 2-D circuit profile containing both random and systematic variations. (b) Filtered out 2-D systematic variations.

The submodularity of detecting an added side-channel current was discussed in Fig. 1. Because of the submodular property, Theorem 2 suggested that a greedy detection algorithm finds the best achievable polynomial time detection. Therefore, the method in Algorithm 1 iteratively selects the gate with the highest deviation from its nominal value in each iteration and reweight its impact until the deviation in nominal value is below the measurement noise threshold. The number of iterations of Algorithm 1 is in the order of the number of inserted Trojans $(|\Lambda|)$. In each iteration, a linear optimization is solved. As mentioned earlier, the dominant cost of our detection is in the measurement phase since testing actual devices is much more expensive than postprocessing computations in Algorithm 1. In Section VI, we experimentally show the convergence rate of $\mathcal{R}$ on benchmarks.

C. Calibration

To perform the anomaly detection, it is required that we calibrate for the systematic variations after profiling the gates. As mentioned in Section II-A, the systematic variations consist of interchip and intrachip variations. The interchip 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 intrachip variations are in the form of a spatial distribution, e.g., 2-D Gaussian in our model. The key observation is that the spatial rate of change of the neighboring gate level profiles due to the systematic intrachip variations (spatial correlations) is slower than the rate of change because of the Trojan insertion or random variations. 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 fact can also be observed by the submodularity property and the rate of diminishing returns. The impact of added/deleted components is very sharp on the nearby gates as we have experimentally observed. Fig. 4(a) demonstrates the profile of a circuit with 9K gates containing both random and systematic variations in the leakage power of the gates, but the systematic variations form a pattern with high correlations among the neighboring gates (slow frequency components) as shown in the filtered out systematic variations by the DCT method in Fig. 4(b).

The above observation suggests using a high-pass filter over the 2-D discrete space of the gate layouts for the identification of the sharp edges that have high frequency components in their

frequency transformation. Years of research in signal processing has introduced a wide-range of such filters that can be possibly used for calibration purposes. In this paper, we use the 2-D discrete cosine transform (DCT). The DCT translates a 2-D signal from a spatial representation into a frequency representation. The DCT has been shown to be very effective and computationally efficient in identifying the low frequency basis functions when compared to several other available transforms.

Since the DCT assumes uniform grid points and the gate layouts are not always uniform in the space, to enable the transform we impose a finer regular grid on the layout and assign to the grid points the average value of their nearest neighbors. DCT, filtering, and inverse discrete cosine transform (IDCT) are performed on the regular grid and are then interpreted on the actual layout.

D. Sensitivity Analysis

In this section, we discuss a method for performing sensitivity analysis. Our proposed method examines how sensitive a measurement is to a change in a gate's nominal value. A gate is considered to be more sensitive if the variation of its nominal value has a larger impact on the l_2 norm of the error. We define the following metric:

$$S_k = \frac{\Sigma_{j=1}^{J} S_{jk}}{J} \qquad \text{where} \qquad S_{jk} = \left(\frac{A_{jk}}{\Sigma_{k=1}^{N_G} A_{jk}} \right). \quad (5)$$

We call S_k the sensitivity factor of the gate G_k. A_{jk} is the gate G_k's nominal value for the jth input. Below we explain the relation between the defined metric and the l_2 norm of error. In our detection method, a gate is reported as anomalous if removing it minimizes the objective function which is the l_2 norm error

$$l_2(\epsilon) = l_2(|A\theta - B|) = \sqrt{\Sigma_{j=1}^{J} |\Sigma_{k=1}^{N_G} A_{jk}\theta_k - b_j|^2}. \quad (6)$$

The following inequality which comes from the Schwartz criteria gives a lower bound for the error:

$$\begin{aligned} &\sqrt{\left(\Sigma_{j=1}^{J} \frac{1}{(\Sigma_{k=1}^{N_G} A_{jk})^2} \right) \left(\Sigma_{j=1}^{J} |\Sigma_{k=1}^{N_G} A_{jk}\theta_k - b_j|^2 \right)} \\ &\quad \geq \left| \Sigma_{j=1}^{J} \left(\Sigma_{k=1}^{N_G} S_{jk}\theta_k - \frac{b_j}{\left(\Sigma_{k=1}^{N_G} A_{jk} \right)} \right) \right| \\ &\quad = \left| J\Sigma_{k=1}^{N_G} S_k\theta_k - \Sigma_{j=1}^{J} \frac{b_j}{\left(\Sigma_{k=1}^{N_G} A_{jk} \right)} \right|. \end{aligned} \quad (7)$$

A variation in gate G_k's nominal value changes its scaling factor θ_k which multiplied by the sensitivity factor S_k, changes the lower bound for the error. Thus for the same $\Delta\theta$, a higher sensitivity factor yields a larger lower bound for the error. Hence the gate with a high sensitivity is more sensitive in our detection method.

E. Cumulative Unimodal Profiling

Once the gate level profile is extracted for a number of ICs over multiple modalities, it is possible to study and analyze the cumulative statistics of the profiled chips. Such cumulative studies have a number of advantages, including: 1) The identified characteristics can aid classification of similarly tainted chips: it is unlikely that one type of exploit is only exercised on a single chip or a small number of chips. 2) The possibility of employing the statistics of the studied groups' characteristics to find signatures for testing the remaining chips, and thereby, speeding up the detection process. 3) The extracted statistics of gate level profiles in each modality advance our understanding of postsilicon characteristics including interchip, intrachip, and random variations. 4) The diminishing property of the reward function $\mathcal{R}$ enables identification of the magnitude of the Trojan modification imposed on the circuit. 5) The extracted characteristics can determine where further focused (and perhaps exhaustive) tests should be made in the marked regions for a better Trojan characterization.

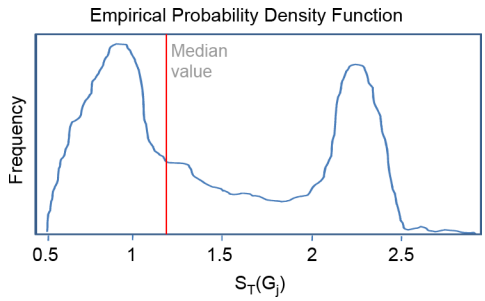

Fig. 5. Empirical distribution of a gate's scaling factors extracted from multiple ICs.

Let us describe more details of the above cumulative studies over multiple ICs. Because of space constraints, a full analysis and evaluation of the cumulative statistics is not included in the paper beyond this section.

1) Assume that there is a certain Trojan contamination that affects a group of chips. In this case, the characteristics of the tainted gates would have a specific density of scaling factors—for a single modality—over different ICs. This distribution would have more than one peak. Fig. 5 shows an example. As can be seen in the figure, the estimated scaling factors for one gate over a group of chips (composed of Trojan and Trojan free instances) has two peaks. It is likely that the same type of Trojan has been applied to the group of ICs whose pertinent gate falls within the peak that is further from the nominal scaling factor. The distance and the valley between the two peaks indicates that the second peak is not due to the manufacturing variability.
2) One can systematically use the above observation to form a compact signature for IC families with similar profiles that are clustered around the same value. Assuming that the signature of all the IC groups (i.e., ICs with similar anomalous exploits) are already extracted, all that is needed for classifying a new IC is a comparison of its signature to the known ones. For example, for the ICs whose gates are placed around the right peak in Fig. 5, we can find the timing paths that pass those tainted gates. For each modality, the same test vectors

that have the best coverage on the chips would be applied to all the chips from one design. Now, one can directly use the statistics of the measured values for identifying the gate level Trojan without even going through the linear optimization for each new test. For example, in timing tests, after the cumulative studies are done on a group of chips, one would only test the paths which include the problematic gates with two peaks. The distance of the chips can be rapidly found in terms of the distance of their problematic path signatures. Note that using the cumulative statistics of multiple path delays and dimension reduction of this statistics by PCA was used earlier for Trojan detection [8]. However, the prior work did not provide a systematic way of classifying the chips over multiple modalities, or introduce methods for tracing back such anomalies to the underlying abnormal gate characteristics.

3) In Section IV-C, we discussed how the high-pass DCT filter can be used for filtering the impact of intrachip correlations. The removed low frequency components, after all the contaminants are identified, can be used for finding an empirical model for the 2-D spatial variations on the chip.
4) Since the affected area's boundary can be determined, more focused tests on suspicious gates and possibly other side channel tests (e.g., EM measurements) can be also used for identifying more characteristics of the anomalies.

V. Multimodal Trojan Detection

A. Multimodal Trojan Identification

The next step of our approach is to combine the results for anomalous gate detection over 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 $P_{\rm FA}$ 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.

1) **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 improve $P_{\rm FA}$. It would also give the minimum achievable $P_{\rm FA}$ (lower bound) by any linear combination of the unimodal detection methods.
2) **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 $P_{\rm FA}$ 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.
3) **Majority voting**: Trojan gates are those that have been marked anomaly by at least $1+\lfloor M/2 \rfloor$ modalities. In our case, 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 tradeoff between P_D and $P_{\rm FA}$.
4) **Weighed voting**: The voting methods above assume that all modalities' votes are combined with equal weights. As we will show in our experimental results, this is not exactly the case. For example, timing tests inherently have less controllability and observability than current-based tests, since the path delay results are typically harder to trace at the output pins. In case of power, the changes in the internal alter the overall drawn current from the source and the changes are observed at the current supply pin. For example, assume that after anomaly detection phase where a number of tainted gates are removed from the equations, the sensitivity value for gate G_k to be S_k^T, S_k^Φ, and S_k^Ψ for timing, leakage, and dynamic current, respectively. Now, the votes of the three unimodal detectors over an anomalous gates are combined as follows: $S_k^T\mathcal{C}_T(G_k) + S_k^\Phi\mathcal{C}_\Phi(G_k) + S_k^\Psi\mathcal{C}_\Psi(G_k) \geq \text{threshold}$.

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 $P_{\rm FA}$ values.

VI. Experimental Evaluations

A. Evaluation Setup

The MCNC benchmark suit was used for evaluating the performance of unimodal detection and the unified multimodal framework. We used the ABC synthesis tool from Berkeley to map the benchmark to a library consisting of inverter, NAND2, NOR2, NAND3, NOR3, NAND4, and NOR4 gates. We used the UCLA Dragon placement tool. We described the process variation model in Section II-A. In a few of our experiments, we study the impact of fluctuating variations. In experiments where the variations are fixed, random variation is 12%, and intradie variation correlation is 60% of the total variation [25]. About 20% of the total variation is uncorrelated intradie variation and the remaining 80% is allotted to the interdie variation. The noninvasive measurement setup was described in Section II-C. HSPICE simulations for 65-nm technology was used for extracting the timing, static, and dynamic currents of each gate in the library for the possible input states. We used the MATLAB optimization toolbox for linear equation solving, and other MATLAB functions for calibration, and likelihood estimation. We report the average over 100 runs of random circuit instances for each gate-level profiling.

TABLE I
DYNAMIC POWER PROFILE ESTIMATION ERROR

ct	size	#inputs	#outputs	3%	5%	10%
C1355	512	41	32	7.8	9.1	11.5
c8	165	28	18	4.2	6.4	11.2
C3450	1131	50	22	3.5	6	9.5
C432	206	36	7	1.5	3.1	6.9
C499	532	41	32	2.2	4.2	8.8

TABLE II
STATIC POWER AND TIMING PROFILE ESTIMATION ERROR

	Static Power			Timing		
ct	**3%**	**5%**	**10%**	**3%**	**5%**	**10%**
C1355	8.5	10	12	4	8	12.3
c8	5.6	7	11.6	5.3	7	11.5
C3450	4	5.9	9.8	2.9	4.1	9.2
C432	1.7	3.5	7.2	3.8	5.4	10.1
C499	2.9	4.5	9	5	6.5	12

TABLE III
UPPER BOUND (UB) AND LOWER BOUND (LB) OF AVERAGE DETECTION ERROR (%) FOR RANDOM PROCESS VARIATIONS WITH $\sigma_{\text{rand}} = 5, 10, 20(\%)$ OVER THE MODALITIES (AVERAGED OVER C1355, C8, C3450, C432, C499 BENCHMARKS)

	σ_{rand}=5%		σ_{rand}=10%		σ_{rand}=20%	
Error Bound	**LB**	**UB**	**LB**	**UB**	**LB**	**UB**
Static	3.9	6.2	6.3	10.0	13.9	22.3
Dynamic	3.6	5.7	6.0	9.5	13.3	21.1
Timing	3.9	6.2	7.0	11.0	17.1	27.6

B. Unimodal Trojan Detection

1) Gate Level Profiling: We first report the evaluation results for gate level profiling. The dynamic current profiling for benchmark circuits are shown in Table I. On each benchmark, we applied as many test vectors at least as twice the number of gates. We used the available input vector generations and the number of new test vectors is determined by the available generator. The first column demonstrates the benchmark name (*ct*). The second column is the number of gates in the benchmark (*size*). The third and fourth columns show the number of primary inputs (*#inputs*) and outputs (*#outputs*). The l_2 norm for characterization estimation error for 3%, 5%, and 10% measurement error are shown in the last three columns. The static current and timing characterization results on the benchmark circuits are demonstrated in Table II. The error ranges are similar to the dynamic power results shown in Table I.

2) Calibration and Detection Bounds: Calibration smooths the intrachip correlations with a slower rate of spatial fluctuations than the modifications by the Trojans. We observed that in case of interchip variations, there is a wide-enough disparity between the available simulation models and the correlated scaling factors of the Trojan's neighboring gates. Thus, calibration is always more than 98% accurate for static, and more than 99% accurate for timing and dynamic current. Newer technology nodes may affect the accuracy of our calibration. Table III shows the average detection error for different random process variations (after calibration of the systematic variations) averaged over a set of benchmark circuits. The lower bound is given for any polynomial algorithms and the upper bound is the best a heuristic polynomial can achieve (4).

3) Sensitivity Analysis: We studied the sensitivity of gates for different modalities and their correlations. We computed the gates' sensitivity factors for several benchmarks and over the three modalities. For each benchmark, we formed a histogram of its gates' sensitivities. On this histogram, we find the 50th (median), 75th, and 95th percentiles for the sensitivity factors. The values are shown in Table IV. For example, a 95 percentile equal to 0.8 means that 95% of the gates have sensitivity factors less than 0.8 of the maximum sensitivity factor in that modality. It is seen that the corresponding percentile values for the leakage and dynamic power modalities are higher than the timing modality. We observe that in the first two modalities, a high percentage of the gates have sensitivity factors comparable to the maximum sensitivity factor in that modality. It shows that in these two cases, most of the gates are sensitive. Meaning that a change in these gates' values effectively alters the error as it was explained in Section IV-D, hence anomalies in them can be detected. But in the timing modality, the sensitivity factors of a high percentage of gates are considerably small. The low sensitivity of timing to gate delay changes shows that the anomalies cannot be detected well in this modality.

TABLE IV
PERCENTILE VALUES OF THE GATE SENSITIVITIES

Benchmark	Percentile	Leakage	Dynamic Power	Timing
C8	50	0.5415	0.5123	0.0656
	75	0.6772	0.5952	0.1605
	95	0.8002	0.7044	0.4709
C432	50	0.4314	0.5093	0.0277
	75	0.4986	0.6089	0.0616
	95	0.6181	0.7623	0.2581
C499	50	0.2843	0.5375	0.0247
	75	0.3506	0.6342	0.0564
	95	0.4305	0.7400	0.1614
C1355	50	0.2974	0.4539	0.0564
	75	0.3465	0.5819	0.1269
	95	0.3853	0.6250	0.3497
C3540	50	0.2371	0.3397	0.0070
	75	0.2842	0.4199	0.0331
	95	0.4667	0.7026	0.1464

TABLE V
PEARSON CORRELATION COEFFICIENTS. Φ, Ψ, AND T SHOW LEAKAGE, DYNAMIC, AND TIMING MODALITIES, RESPECTIVELY

Pearson Coeff.	C8	C432	C499	C1355	C3540
$\rho(\mathbf{\Phi}, \mathbf{\Psi})$	0.7556	0.5727	0.8270	0.8695	0.6376
$\rho(\mathbf{\Phi}, \mathbf{T})$	0.1113	-0.0158	-0.0138	-0.0802	0.0886
$\rho(\mathbf{\Psi}, \mathbf{T})$	-0.0932	0.0391	-0.0024	-0.0906	-0.0522

Table V shows the Pearson's linear correlation coefficient for each pair of the modality sensitivities. The interpretation of each coefficient is such that the closer it is to one, the more linearly correlated the two modalities are. It follows from Table V that the gate sensitivity factors of the leakage and dynamic power modalities are highly correlated for all the benchmarks, but the sensitivity factors of the timing modality are rarely correlated with those of the other two modalities.

The relationship between the sensitivity factors of different modalities can also be observed in Fig. 6(a) and (b). Sensitivity factors of the leakage modality are plotted in an ascending order x-axis and sensitivity factors of the corresponding gates of the dynamic power and timing modalities are plotted respectively y-axis. A gate which is less sensitive in one modality can be more sensitive in the other one. The fact that there is a little correlation between the sensitivity factors of the timing

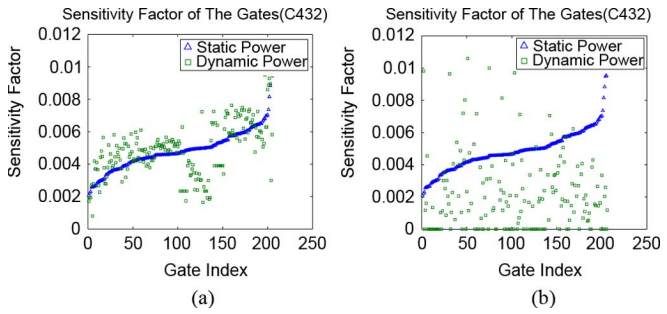

Fig. 6. Comparison of sensitivity factors for modalities. (a) Static and dynamic power. (b) Static power and timing.

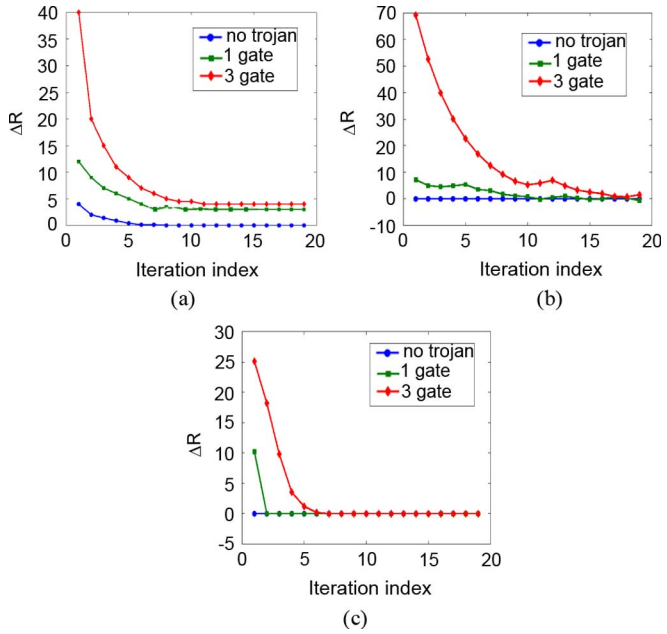

Fig. 7. Stepwise diminishing return improvement for a Trojan free, a 1-gate, and a 3-gate Trojan. (a) Leakage modality. (b) Dynamic power. (c) Timing.

modality and the two other modalities suggests that using the timing modality along with one of the other modes can improve the detection results.

4) Unimodal Anomaly Detection: We now evaluate the effectiveness of unimodal anomaly detection for the three modalities. Three scenarios are considered: i) a Trojan-free circuit, ii) one extra NAND2 gate inserted as a Trojan, and iii) a three-gate comparator circuit is added. We inserted the Trojans within the empty spaces of the automatic layout generated by the Dragon tool. We first study the key property of the changes in diminishing return at each iteration. As discussed, the diminishing return would be monotonically decreasing assuming no random perturbations. Fig. 7(a)–(c) demonstrate the diminishing return versus iteration number for the static, dynamic, and timing modalities for the C432 benchmark. The figures show that as we go through iterations, the stepwise change in diminishing return becomes smaller. Also, the step-wise change is higher for larger Trojans and is very low for no Trojans. After the Trojan circuit reaches a similar diminishing return difference as the Trojan-free case, no further significant change is observed. Similar trends are observed in all modalities.

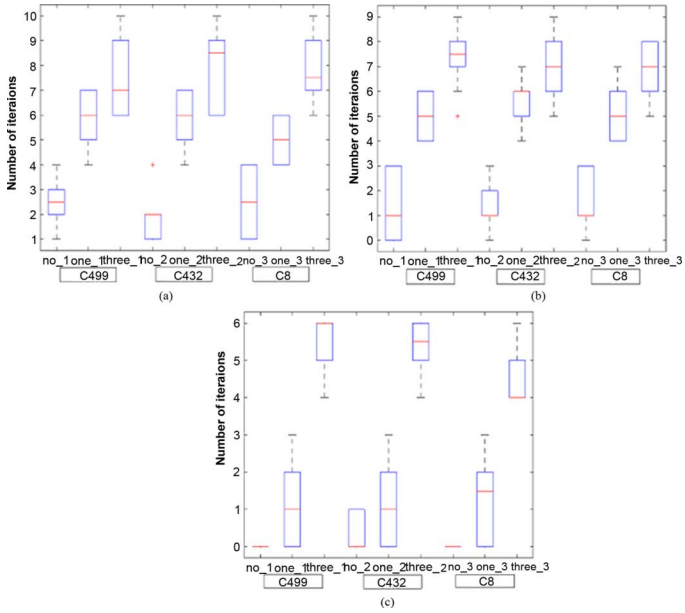

Fig. 8. Boxplots showing the final number of iterations or equivalently $|\Lambda|$ for the Trojan free, 1- and 3-gate Trojan. (a) Leakage modality. (b) Dynamic power. (c) Timing.

We exploit the diminishing return results to define stopping criteria for Algorithm 1. For instance, a criteria that would halt the algorithm once the step-wise decreasing improvement reaches 2. We show in Fig. 7 that this stopping criteria results in an average false alarm for one gate (the benchmark has 206 gates) meaning that $P_{\text{FA}} = 1/206 = 0.5\%$. In case of the smaller Trojan, about two gates are not detected, meaning that $P_D = 1\%$. We see that additional gates are reported because the Trojan gates impact the side channel measurements of the logically connected gates. One could exploit this observation to help localize the Trojan, but further localization is outside the scope of this paper. The stopping criteria is faster reached by the timing modality compared to the other modalities. This is because even if multiple gates are impacted by Trojan, only the ones with a high sensitivity will affect the reward function. In Section IV-D we observed that the timing modality would result in less gate sensitivities compared to other modalities and, therefore, a smaller number of gates would be detected. Selecting the stopping criteria results in a tradeoff between P_{FA} and P_D. On larger benchmarks, we observed that the P_{FA} is higher than the small circuits for no Trojan case. Note that the P_D and P_{FA} results for the multimodal case inherently include the unimodal case, so we decided to only report it in the multimodal subsection.

The boxplots for the number of iterations (before reaching a stopping criteria) on three benchmark circuits and for the three modalities are shown in Figs. 8(a) and (b), respectively. The number of iterations is related to the number of detected anomalous gates. We observe that by increasing the Trojan size, more anomalous gates would be detected. Note that the number of detected gates does not correspond to the number of gates in the Trojan circuitry. The anomalous gates are the ones that are impacted by a nearby Trojan. It can also be seen that for the delay modality, the number of the detected anomalous gates is less

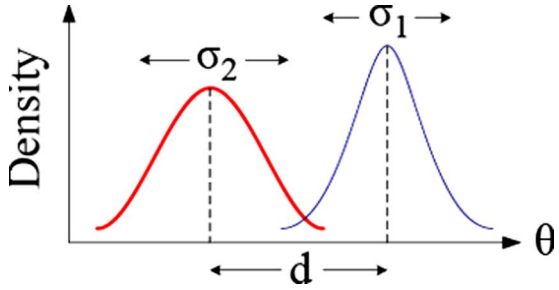

Fig. 9. Scaling factor distribution (2 peaks).

TABLE VI
DISTANCE BETWEEN MEANS AND VARIANCES FOR THE TWO-PEAK DISTRIBUTION

ct	d	σ_2	σ_1
C1355	0.49	0.2	0.22
c8	0.5	0.3	0.21
C3450	0.55	0.25	0.26
C432	0.48	0.18	0.22
C499	0.46	0.2	0.18

TABLE VII
P_{FA} IN TERMS OF THE DETECTED ANOMALOUS GATES IN A TROJAN-FREE CIRCUIT

ct	i	ii	iii	iv
C1355	0	4	2	2
c8	0	3	1	2
C3450	0	3	3	3
C432	1	2	1	1
C499	0	3	1	1

than the other two modalities. Our tests illustrate that the timing modality is more sensitive to the Trojan's location in the circuit. A Trojan that cannot impact multiple paths would result in a low number of detected anomalous gates.

5) Cumulative Unimodal Profiling: A cumulative profile of a chip is built for a 1000 randomly generated instances for each benchmark. For one benchmark, two cases could be correctly classified: the one-gate Trojan versus the Trojan-free. The shape of the derived distribution with two peaks is shown in Fig. 9. We approximate each distribution by a Gaussian where the two distributions are centered about a different peak. The peaks are separated by a distance d and the variances of the two distributions is denoted by σ_1 and σ_2. These values are reported in Table VI for different benchmarks (leakage modality). The distances are large enough to distinguish between the two Gaussians.

C. Multimodal Trojan Detection

In Table VII, we represent the false alarm (in terms of the number of gates) over 100 Trojan-free chips. The false alarm is the highest for the timing modality. After showing the benchmark names in the first column, the remainder of the columns represent the results for: i) unanimous, ii) conservative, iii) majority, and iv) weighed. The lowest P_{FA} was achieved by the unanimous voting that can moderate the impact of the modalities with a high false alarm. The probability of detection (percentage) for a one-gate Trojan over 100 chips is demonstrated in Tables VIII and IX. The second to fifth columns represent the cases where the Trojan is placed at a location of high sensitivity (for all modalities). The last four columns show P_D for a Trojan

TABLE VIII
P_D FOR A 1-GATE TROJAN INSERTED AT HIGH AND LOW SENSITIVITY LOCATIONS FOR THE THREE MODALITIES

	High sensitivity				Low sensitivity			
ct	**i**	**ii**	**iii**	**iv**	**i**	**ii**	**iii**	**iv**
C1355	98	100	100	100	25	48	48	32
c8	97	100	100	100	30	46	46	34
C3450	95	100	100	100	27	50	50	40
C432	97	100	100	100	24	49	49	34
C499	98	100	100	100	32	49	49	38

TABLE IX
P_D FOR A 1-GATE TROJAN INSERTED AT LOCATIONS WITH OPPOSITE LEVELS OF SENSITIVITIES FOR TIMING MODALITY AND THE OTHER TWO MODALITIES

	Low timing sensitivity				High timing sensitivity			
ct	**i**	**ii**	**iii**	**iv**	**i**	**ii**	**iii**	**iv**
C1355	25	100	100	100	47	96	47	77
c8	30	100	100	100	49	98	49	74
C3450	27	100	100	100	52	95	52	76
C432	24	100	100	100	44	95	44	69
C499	32	100	100	100	49	99	49	74

gate inserted at low sensitivity positions (for all modalities). Columns two to five of Table IX represent P_D where the Trojan is placed at a location with low timing sensitivity and a high power sensitivity. The last four columns show P_D where the Trojan is placed at a location with a high timing sensitivity and low power sensitivity. Note that the static and dynamic modalities are highly correlated. For highly sensitive gates, P_D is close to 100% except for the unanimous voting that minimizes P_{FA} as opposed to P_D.

The sensitivity results can also be used for finding the locations where the Trojans can be best hidden. Our formal sensitivity metric directly corresponds to P_D. Since our methods reach a detection bound for the available test vectors and for each modality, for improving the detection results methods that increase controllability and observability or newer test modalities should be adopted.

VII. RELATED WORK

Hardware Trojan detection is a new and emerging research area. Agrawal *et al.* [4] 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.* [7], [26] 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.* [5], [6] investigate power supply transient signal analysis methods for detecting Trojans. An IC's supply current is measured from multiple supply ports to deal with the small Trojan-signal-to-background-current ratios. The calibration technique transforms the measured currents for each IC to match those produced from a golden, Trojan free simulation model. 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.

Jin and Markis [8] 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 [27]. Gate-level characterization was used for postsilicon profiling [28]–[30] and its use for IC Trojan detection was first proposed in [9], [31] and also used in [10]–[13], [32], [33]. However, optimality guarantees (bounds), calibration, sensitivity, and multimodal combining were not discussed in the literature. Our work provides the first rigorous treatment of the multimodal Trojan detection problem, near-optimal solutions, mathematical calibration and introduction of a sensitivity metric. 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.

Note that the linear dependence of the path delays on gate characteristics is a well-known fact in traditional testing. But since in testing the path delays are important and the fault models are radically different from our Trojan models, this linear relationship is only exploited for finding the basis path sets [15] that is the smallest path set such that every other path in the circuit graph is linearly dependent on it. Our noninvasive profiling decomposes the measurements to gate level components as opposed to only generating a statistical signature. Thus, not only does it have a linear test time with respect to the number of gates (as opposed to exponential number of timing paths or power test vectors) but also it provides a better insight and Trojan detection capabilities.

A multimodal Trojan detection approach (concurrent to our work) was proposed in [34]. This work combines the IDDT measurements with $F_{\max}$ measurements as a multimode test. Aside from the multimode testing, our work is drastically different since we exploit other test modalities, pursue different objectives, and have a distinct approach.

Our method exploits the concept and results of submodular function optimization [23]. The concept has been utilized earlier in a variety of contexts [35], including but not limited to: set cover [24], sensor networks [36], and graph problems [37]. Our work is the first to formulate and use the submodularity concept for IC Trojan detection.

An earlier version of this work appeared in [20]. The new aspects of this paper include sensitivity analysis and cumulative unimodal profiling. Our new sensitivity metric evaluates and compares the efficiency of our detection method over the modalities. We also show how the cumulative detection data can be used for categorizing ICs based on the Trojan symptoms and for accelerating detection. calibration is presented with much more details. Extensive results are provided for the newly added sections and experimental analysis are expanded.

VIII. 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. To address the complex problem, we devise an iterative detection and profiling method. Our detection objective function is shown to be submodular. Because of 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. We introduce a new sensitivity metric which quantifies the impact of altering each gate on the overall detection result. Experimental evaluations on benchmarks for timing, leakage current, and transient currents show the effectiveness of the proposed approach. To the best of our knowledge, this is the first systematic unified IC Trojan detection framework. The emergence of testing techniques and more exact measuring equipment could improve the detection capabilities of the unified multimodal framework.

Acknowledgment

The authors acknowledge M. Majzoobi for help in proof reading, and D. Shamsi and Y. Alkabani for help in programming and experiments.

References

[1] M. Potkonjak, "Synthesis of trustable ICs using untrusted CAD tools," in *Proc. Design Automation Conference (DAC)*, 2010, pp. 633–634.

[2] M. Tehranipoor and F. Koushanfar, "A survey of hardware Trojan taxonomy and detection," *IEEE Des. Test Comput.*, 2010, DOI: 10.1109/MDT.2009.159 .

[3] R. Karri, J. Rajendran, K. Rosenfeld, and M. Tehranipoor, "Trustworthy hardware: Identifying and classifying hardware Trojans," *IEEE Computer*, vol. 43, no. 10, pp. 39–46, Oct. 2010.

[4] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, and B. Sunar, "Trojan detection using IC fingerprinting," in *IEEE Symp. Security and Privacy (S&P)*, 2007, pp. 296–310.

[5] R. Rad, J. Plusquellic, and M. Tehranipoor, "Sensitivity analysis to hardware Trojans using power supply transient signals," in *Proc. Int. Symp. Hardware Oriented Security and Trust (HOST)*, 2008, pp. 3–7.

[6] R. Rad, X. Wang, J. Plusquellic, and M. Tehranipoor, "Power supply signal calibration techniques for improving detection resolution to hardware Trojans," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, 2008, pp. 632–639.

[7] M. Banga, M. Chandrasekar, L. Fang, and M. Hsiao, "Guided test generation for isolation and detection of embedded Trojans in ICs," in *Great Lakes Symp. VLSI (GLS-VLSI)*, 2008, pp. 363–366.

[8] Y. Jin and Y. Makris, "Hardware Trojan detection using path delay fingerprint," in *Proc. Int. Symp. Hardware Oriented Security and Trust (HOST)*, 2008, pp. 51–57.

[9] M. Potkonjak, A. Nahapetian, M. Nelson, and T. Massey, "Hardware Trojan horse detection using gate-level characterization," in *Proc. Design Automation Conf. (DAC)*, 2009, pp. 688–693.

[10] Y. Alkabani and F. Koushanfar, "Consistency-based characterization for IC Trojan detection," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, 2009, pp. 123–127.

[11] S. Wei, S. Meguerdichian, and M. Potkonjak, "Gate-level characterization: Foundations and hardware security applications," in *Proc. Design Automation Conf. (DAC)*, 2010, pp. 222–227.

[12] M. Nelson, A. Nahapetian, F. Koushanfar, and M. Potkonjak, "SVD-based ghost circuitry detection," in *Proc. Information Hiding (IH) Conf.*, 2009, pp. 221–234.

[13] S. Wei and M. Potkonjak, "Scalable segmentation-based malicious circuitry detection and diagnosis," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, 2010, pp. 483–486.

[14] F. Liu, "A general framework for spatial correlation modeling in VLSI design," in *Proc. Design Automation Conf. (DAC)*, 2007, pp. 817–822.
[15] M. Sharma and J. Patel, "Bounding circuit delay by testing a very small subset of paths," in *VLSI Test Symp. (VTS)*, 2000, pp. 333–341.
[16] N. Jha and S. Gupta, *Testing of Digital Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[17] A. Murakami, S. Kajihara, T. Sasao, I. Pomeranz, and S. M. Reddy, "Selection of potentially testable path delay faults for test generation," in *Proc. Int. Test Conf. (ITC)*, 2000, pp. 376–384.
[18] S. Sabade and D. Walker, "IDDX-based test methods: A survey," *ACM Trans. Design Automation Electron. Syst.*, vol. 9, no. 2, pp. 159–198, 2004.
[19] S. Chakravarty and P. Thadikaran, "Simulation and generation of IDDQ tests for bridging faults in combinational circuits," *IEEE Trans. Comput.*, vol. 45, no. 10, pp. 1131–1140, Oct. 1996.
[20] F. Koushanfar, A. Mirhoseini, and Y. Alkabani, "A unified submodular framework for multimodal IC Trojan detection," in *Proc. Information Hiding (IH) Conf.*, 2010.
[21] E. Bai, H. Cho, R. Tempo, and Y. Ye, "Optimization with few violated constraints for linear bounded error parameter estimation," *IEEE Trans. Autom. Control*, vol. 47, no. 7, pp. 1067–1077, Jul. 2002.
[22] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*. Dordrecht: Kluwer, 1997.
[23] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of the approximations for maximizing submodular set functions," *Math. Programming*, vol. 14, pp. 265–294, 1978.
[24] U. Feige, "A threshold of ln n for approximating set cover," *J. ACM*, vol. 45, no. 4, pp. 634–652, 1998.
[25] Y. Cao and L. T. Clark, "Mapping statistical process variations toward circuit performance variability: An analytical modeling approach," in *Proc. Design Automation Conf. (DAC)*, 2005, pp. 658–663.
[26] M. Banga and M. Hsiao, "A region based approach for the identification of hardware Trojans," in *Proc. Int. Symp. Hardware Oriented Security and Trust (HOST)*, 2008, pp. 43–50.
[27] J. Li and J. Lach, "At-speed delay characterization for IC authentication and Trojan horse detection," in *Proc. Int. Symp. Hardware Oriented Security and Trust (HOST)*, 2008, pp. 8–14.
[28] D. Shamsi, P. Boufounos, and F. Koushanfar, "Noninvasive leakage power tomography of integrated circuits by compressive sensing," in *Proc. Int. Symp. Low Power Electronics and Designs (ISLPED)*, 2008, pp. 341–346.
[29] F. Koushanfar, P. Boufounos, and D. Shamsi, "Post-silicon timing characterization by compressed sensing," in *Proc. Int. Conf. Computer-Aided Design (ICCAD)*, 2008, pp. 185–189.
[30] Y. Alkabani, F. Koushanfar, N. Kiyavash, and M. Potkonjak, "Trusted integrated circuits: A nondestructive hidden characteristics extraction approach," in *Proc. Information Hiding (IH) Conf.*, 2008, pp. 102–117.
[31] F. Koushanfar and M. Potkonjak, "CAD-based security, cryptography, and digital rights management," in *Proc. Design Automation Conf. (DAC)*, 2007, pp. 268–269.
[32] A. Vahdatpour, M. Potkonjak, and S. Meguerdichian, "A gate level sensor network for integrated circuits temperature monitoring," *IEEE Sensors*, to be published.
[33] A. Vahdatpour and M. Potkonjak, "Leakage minimization using self sensing and thermal management," in *Proc. Int. Symp. Low Power Electronics and Designs (ISLPED)*, pp. 1–6.
[34] S. Narasimhan, D. Dongdong, R. Chakraborty, S. S. Paul, F. Wolff, C. Papachristou, K. Roy, and S. Bhunia, "Multiple-parameter side-channel analysis: A non-invasive hardware Trojan detection approach," in *Proc. Int. Symp. Hardware Oriented Security and Trust (HOST)*, 2010, pp. 13–18.
[35] A. Krause and C. Guestrin, "Near-optimal observation selection using submodular functions," in *Proc. Nat. Conf. Artificial Intelligence (AAAI)*, 2007, vol. 2, pp. 1650–1654.
[36] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. Conf. Knowledge Discovery and Data Mining (KDD)*, 2007, pp. 420–429.
[37] C. Chekuri and M. Pal, "A recursive greedy algorithm for walks in directed graphs," in *Annu. Symp. Foundations of Computer Science (FOCS)*, 2005, pp. 245–253.

Farinaz Koushanfar (S'99–M'06) received the Ph.D. degree in electrical engineering and computer science and the M.A. degree in statistics, both from University of California Berkeley, in 2005, and the M.S. degree in electrical engineering from the University of California Los Angeles.

She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, where she directs the Texas Instruments DSP Leadership University Program. Her research interests include adaptive and low power embedded systems design, hardware security, and design intellectual property protection.

Prof. Koushanfar is a recipient of the Presidential Early Career Award for Scientists and Engineers (the highest honor bestowed by the U.S. government on early career outstanding scientists and engineers), the Office of Naval Research (ONR) Young Investigator Program Award, the Defense Advanced Project Research Agency (DARPA) Young Faculty Award, the National Science Foundation CAREER Award, MIT Technology Review TR-35, an Intel Open Collaborative Research fellowship, and a best paper award at Mobicom.

Azalia Mirhoseini (S'09) received the B.S. degree in electrical engineering from Sharif University of Technology in 2009. She is currently working toward the M.S. degree in the Department of Electrical and Computer Engineering, Rice University, Houston, TX.

Her research interests are in modeling and optimization of hybrid energy supply systems, low power embedded systems, and modeling for detection of hardware malware.

Ms. Mirhoseini is a recipient of Microsoft Women Graduate Student Scholarship (2010), and a National Gold Medal winner in the Iran Mathematics Olympiad (2004).