Strategic Hardware Trojan Testing with Hierarchical Trojan Types

Swastik Brahma and Satyaki Nan
Department of Computer Science
Tennessee State University
Nashville, TN 37209
{sbrahma,snan}@tnstate.edu

Laurent Njilla
Cyber Assurance Branch
US Air Force Research Laboratory
Rome, NY
laurent.njilla@us.af.mil

2021 55th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-1268-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISS50987.2021.9400297

***Abstract*— In this paper, we consider the problem of detecting hardware Trojans in an Integrated Circuit (IC) from a game theoretic standpoint. The paper considers the presence of multiple classes of Trojans, with each class containing multiple Trojan types, and characterizes the Nash Equilibrium (NE) strategy for inserting a Trojan (from the perspective of a malicious entity) and detecting a Trojan (from the perspective of a defender) under consideration of the impact that an undetected Trojan has on the defender's system. The paper also models a sequential hardware Trojan testing game, where the defender tests for the presence of Trojans over time, and characterizes the NE strategy of such a game. Numerous simulation results are presented to gain insights into the game theoretic hardware Trojan testing techniques presented in the paper.**

***Index Terms*—Hardware Trojans, Game Theory, Security.**

I. Introduction

The detection of hardware Trojans in Integrated Circuits (ICs) is an important problem and has received attention in the past [1]–[6], [11]–[15], [17]. For example, in [4] the authors propose a region-based partitioning and excitation approach for circuit designs that can accurately estimate the location of Trojans in ICs. In [12], the authors have used random sequences of test patterns that can generate noticeable differences between the power profile of a genuine IC and the Trojan counterpart, but their effectiveness is limited in terms of the manufacturing processes, behavior and size of the Trojans. Again, in [13], the authors propose a methodology, referred to as MERO (Multiple Excitation of Rare Occurence), for statistical test generation that maximizes the probability of detecting inserted Trojans. Since exhaustive testing of all possible Trojan types can be prohibitive, the works in [7]–[9], [18] model the detection of hardware Trojans using Game Theory [10] to determine which Trojan type to test against a strategic malicious manufacturer. Specifically, in [7], the authors develop a game theoretic strategy for testing Trojans in an IC, but rely on software-based techniques to find the equilibrium solution. [8] investigates metrics for analyzing defense measures against strategic insertion of hardware Trojans. [9] presents a two-person Trojan detection game, but limits investigation of the equilibrium to a specific instance of the model. The work in [18] analyzes game theoretic testing strategies via simulations.

It should be noted that past work on game theoretic Trojan testing (e.g., [7], [9], [18]) relies on software-based techniques for determining the equilibrium solution. Moreover, to the best of our knowledge, past work has not accounted for the hierarchical classification structure exhibited by Trojans in analyzing attack-defense strategies. Furthermore, to the best of our knowledge, past work has also not modeled and analyzed from a game theoretic perspective sequential Trojan testing where the defender can test for the presence of Trojans over time. In this paper, we aim to develop game theoretic hardware Trojan testing techniques while overcoming the aforementioned limitations of past work. Specifically, we present a game theoretic framework for testing hardware Trojans in an IC considering Trojans to exhibit a hierarchical structure consisting of multiple Trojan classes and analytically characterize the Nash Equilibrium (NE) based testing strategies. The main contributions of the paper are as follows.

- We consider the presence of multiple classes of Trojans, with each class containing multiple Trojan types, and analytically characterize the NE strategy for inserting a Trojan (from the perspective of a malicious entity) and detecting a Trojan (from the perspective of a defender).
- We first consider the scenario where an IC can be tested once for determining the presence of a hardware Trojan and characterize the NE-based Trojan insertion and detection strategies.
- We also model and analyze a sequential hardware Trojan testing game where an IC can be tested repeatedly for determining the presence of a Trojan and characterize the NE-based Trojan insertion and detection strategies.
- Extensive simulation results are provided to gain insights into the game theoretic techniques presented in the paper.

The rest of the paper is organized as follows. Section II presents our game theoretic results for testing Trojans when the defender can test an IC once for the presence of a Trojan. Section III presents our results for sequential hardware Trojan testing. Section IV presents simulation results that provide insights into the game theoretic techniques presented. Finally, Section V concludes the paper.

This research was supported in part by the ARO under Grant Number W911NF-18-1-0152 and in part by the NSF under Grant Number HRD 1912414.

DISTRIBUTION A. Approved for public release. Distribution unlimited. Case Number AFRL-2021-0123. Dated 20 Jan 2021.

II. Game Theory-Based Hardware Trojan Testing With Hierarchical Trojan Types

In practice, Trojans exhibit a hierarchical structure consisting of multiple classes with each class containing multiple Trojan types. To illustrate our model and result, we first consider two classes of Trojans, viz. Class 1 and Class 2, with class $i \in \{1,2\}$, containing N_i Trojan types. For e.g., Class 1 can contain *information leaking* type of Trojans (e.g., [14], [15]) while Class 2 can contain those that increase the *power consumption* of a device (e.g., [16], [17]). We refer the reader to [1] for a treatise on the classification of Trojans.

We consider that a malicious manufacturer (referred to as the attacker (A)) chooses to insert a Trojan from Class 1 with a probability q_1 (and a Trojan from Class 2 with a probability $1-q_1$) into the manufactured IC. Again, we consider the buyer of the IC, whom we refer to as the defender (D), to test the IC once for the presence of a Trojan from Class 1 with a probability p_1 (and to test the IC for the presence of a Trojan from Class 2 with a probability $1-p_1$). For simplicity of exposition, we consider the defender and the attacker to uniformly pick a Trojan type for testing and insertion, respectively, from their chosen Trojan classes (i.e., to pick a Trojan type from chosen class i with a probability $1/N_i$). We consider that if the defender tests the IC against the inserted Trojan, the Trojan is detected, and the malicious manufacturer is imposed a fine F (which negatively impacts the attacker's utility and positively impacts the defender's utility). However, if the defender tests the IC for the presence of a Trojan which was not inserted by the attacker, the Trojan remains undetected and we consider that an undetected Trojan of class i, where $i \in \{1,2\}$, provides a benefit V_i to the attacker (which positively impacts the attacker's utility and negatively impacts the defender's utility). The strategic interaction between the defender and the attacker, in this paper, is modeled as a zero-sum game. Next, we characterize the mixed strategy NE of the game in terms of the Trojan detection and insertion strategies of the defender and the attacker, respectively.

LEMMA 1. *Given two classes of Trojans, viz.Class* 1 *and Class* 2*, with class* i *containing* N_i *types of Trojans, at NE, the defender tests the IC for the presence of a Trojan from Class* 1 *with a probability* $p_1 = \frac{\frac{F+V_2}{N_2}+(V_1-V_2)}{1+\frac{N_2(F+V_1)}{N_1(F+V_2)}}$ *and the attacker inserts a Trojan from Class* 1 *with a probability* $q_1 = \frac{1}{1+\frac{N_2}{N_1}\left(\frac{F+V_1}{F+V_2}\right)}$.

Proof. The expected utility (say, E_D^1) of D (defender) from testing the IC for the presence of a Trojan from Class 1 is,

$$E_D^1 = \left[\frac{Fq_1}{N_1} + (-V_1)q_1\left(\frac{N_1-1}{N_1}\right) + (-V_2)(1-q_1)\right] \quad (1)$$

Similarly, the expected utility (say, E_D^2) of D from testing the IC for the presence of a Trojan from Class 2 is,

$$E_D^2 = -V_1q_1 + \left[\frac{F(1-q_1)}{N_2} + (V_2q_1 - V_2)\left(\frac{N_2-1}{N_2}\right)\right] \quad (2)$$

Equating (1) and (2) to make the defender indifferent between choosing a Trojan from Class 1 and Class 2 at the mixed strategy NE yields,

$$q_1 = \frac{1}{1+\frac{N_2}{N_1}\left(\frac{F+V_1}{F+V_2}\right)} \quad (3)$$

Now, the expected utility (say, E_A^1) of A (attacker) from choosing to insert a Trojan from Class 1 is,

$$E_A^1 = \frac{(-FP_1)}{N_1} + \frac{(V_1P_1)(N_1-1)}{N_1} + V_1(1-P_1) \quad (4)$$

Similarly, the expected utility (say, E_A^2) of A from choosing to insert a Trojan from Class 2 is,

$$E_A^2 = V_2P_1 + \frac{(-F)(1-P_1)}{N_2} + \frac{(V_2)(1-P_1)(N_2-1)}{N_2} \quad (5)$$

Equating (4) and (5) to make the attacker indifferent between choosing to insert a Trojan from Class 1 and Class 2 at the mixed strategy NE yields,

$$p_1 = \frac{\frac{F+V_2}{N_2} + (V_1-V_2)}{1+\frac{N_2(F+V_1)}{N_1(F+V_2)}} \quad (6)$$

This proves the lemma. □

Next, we generalize the aforementioned game considering M classes of Trojans to be present.

A. Game Theoretic Trojan Testing with M Trojan Classes

We now generalize the aforementioned game model by considering that there are M classes of Trojans, viz. $\{1,\cdots,M\}$, with each class i containing N_i types of Trojans. We consider that the strategy of the attacker is to adopt $\mathbf{q} = (q_1,\cdots,q_M)$ such that $\sum_{i=1}^{M} q_i = 1$, where q_i is the probability of the attacker inserting a Trojan from class i with the attacker considered to uniformly choose a type of Trojan from its chosen class. Again, the strategy of the defender is to adopt $\mathbf{p} = (p_1,\cdots,p_M)$ such that $\sum_{i=1}^{M} p_i = 1$, where p_i is the probability with which the defender tests the IC for the presence of a Trojan from class i with the defender considered to uniformly choose a type of Trojan from its chosen class. We consider that if the defender tests the IC against the inserted Trojan, the Trojan is detected, and the malicious manufacturer is imposed a fine F. However, if the defender tests the IC for the presence of a Trojan which was not inserted by the attacker, the Trojan remains undetected and we consider that an undetected Trojan of class i provides a benefit V_i to the attacker. Next, we characterize the mixed strategy NE of the aforementioned game where M classes of Trojans are present.

THEOREM 1. *At NE,*

- *The defender, for any chosen* $i \in \{1,\cdots,M\}$*, tests the IC for the presence of a Trojan from class* i *with a probability* $p_i = \frac{1+\sum_{j=1}^{M}\frac{N_j(V_i-V_j)}{F+V_j}}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j(F+V_i)}{N_i(F+V_j)}}$ *and tests the IC for the presence of a Trojan from class* j *with the probability* $p_j = \frac{\frac{F+V_i}{N_i}p_i+(V_j-V_i)}{\frac{F+V_j}{N_j}}$, $\forall j \in \{1,\cdots,M\}, j \neq i$.
- *The attacker, for any chosen* $i \in \{1,\cdots,M\}$*, inserts a Trojan from class* i *with a probability* $q_i = \frac{1}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j(F+V_i)}{N_i(F+V_j)}}$ *and inserts a Trojan from class*

j with the probability $q_j = \frac{N_j(F+V_i)}{N_i(F+V_j)} q_i, \forall j \in \{1, \cdots, M\}, j \neq i$.

Proof. The expected utility (say, E_D^i) of D from testing the IC for the presence of a Trojan from class i, where $i \in \{1, \cdots, M\}$, is,

$$E_D^i = \left[\frac{F q_i}{N_i} + (-V_i) q_i \left(\frac{N_i - 1}{N_i} \right) + \sum_{j=1, j \neq i}^{M} (-V_j)(q_j) \right] \quad (7)$$

At the mixed strategy NE, we must have $E_D^1 = E_D^2 = \cdots = E_D^M$. Now, for $i, j \in \{1, \cdots, M\}, i \neq j$, equating $E_D^i = E_D^j$, after some manipulations yield,

$$q_j = q_i \frac{N_j(F + V_i)}{N_i(F + V_j)} \quad (8)$$

Now, for $\mathbf{q} = (q_1, \cdots, q_M)$ to be a feasible strategy, for any chosen $i \in \{1, \cdots, M\}$, we must have

$$q_i + \sum_{j=1, j \neq i}^{M} q_j = 1 \quad (9)$$

$$\Rightarrow q_i + \sum_{j=1, j \neq i}^{M} q_i \frac{N_j(F + V_i)}{N_i(F + V_j)} = 1 \quad (\text{using}(8))$$

$$\Rightarrow q_i = \frac{1}{1 + \sum_{j=1, j \neq i}^{M} \frac{N_j(F+V_i)}{N_i(F+V_j)}} \quad (10)$$

Clearly, from the above, if the attacker, for any chosen $i \in \{1, \cdots, M\}$, chooses q_i as given in (10) and q_j, $\forall j \in \{1, \cdots, M\}, j \neq i$, as given in (8), any strategy of defender becomes a best response against the attacker's strategy since the defender becomes indifferent between choosing a Trojan class for testing (as well as $\mathbf{q} = (q_1, \cdots, q_M)$ is ensured to be a feasible strategy).

Now, the expected utility (say, E_A^i) of A from choosing to insert a Trojan from class i, where $i \in \{1, \cdots, M\}$, is,

$$E_A^i = \left[\frac{(-F) p_i}{N_i} + (V_i) p_i \left(\frac{N_i - 1}{N_i} \right) + (V_i)(1 - p_i) \right] \quad (11)$$

At the mixed strategy NE, we must have $E_A^1 = E_A^2 = \cdots = E_A^M$. Now, for $i, j \in \{1, \cdots, M\}, i \neq j$, equating $E_A^i = E_A^j$, after some manipulations yield,

$$p_j = \frac{\frac{F+V_i}{N_i} p_i + (V_j - V_i)}{\frac{F+V_j}{N_j}} \quad (12)$$

Now, for $\mathbf{p} = (p_1, \cdots, p_M)$ to be a feasible strategy, for any chosen $i \in \{1, \cdots, M\}$, we must have

$$p_i + \sum_{j=1, j \neq i}^{M} p_j = 1 \quad (13)$$

$$\Rightarrow p_i + \sum_{j=1, j \neq i}^{M} \frac{\frac{F+V_i}{N_i} p_i + (V_j - V_i)}{\frac{F+V_j}{N_j}} = 1 \ (\text{using}(12))$$

$$\Rightarrow p_i = \frac{1 + \sum_{j=1}^{M} \frac{N_j(V_i - V_j)}{F+V_j}}{1 + \sum_{j=1, j \neq i}^{M} \frac{N_j(F+V_i)}{N_i(F+V_j)}} \quad (14)$$

Clearly, from the above, if the defender, for any chosen $i \in \{1, \cdots, M\}$, chooses p_i as given in (14) and p_j, $\forall j \in \{1, \cdots, M\}, j \neq i$, as given in (12), any strategy of attacker becomes a best response against the defender's strategy (as well as $\mathbf{p} = (p_1, \cdots, p_M)$ is ensured to be a feasible strategy).

Thus, if the attacker chooses q_i, for any chosen $i \in \{1, \cdots, M\}$, as given in (10) and q_j, $\forall j \neq i$, as given in (8) while the defender chooses p_i, for any chosen $i \in \{1, \cdots, M\}$, as given in (14) and p_j, $\forall j \neq i$, as given in (12), both would be playing their best responses against each other. This proves the theorem. □

Next, we provide numerical results to corroborate the above results. In Figure 1, we consider that there are two Trojan classes, viz. Class 1 and Class 2, and show the defender's expected utility versus the probability (q_1) of inserting a Trojan from Class 1. For the figure, we consider N_1 = 3, N_2 = 3, V_1 = 3, V_2 = 2, and F = 25, and plot the defender's utilities from always testing a Trojan from Class 1 (i.e., from choosing $p_1 = 1$) and from always testing a Trojan from Class 2 (i.e., from choosing $p_2 = 1$). The point where the two utilities intersect implies that the expected utility of the defender from testing a Trojan from Class 1 equals that of the defender from testing a Trojan from Class 2 (as needed at the mixed strategy NE), which as can be seen from the figure, occurs at $q_1 = 0.49$ and which can be shown to tally with the attacker's NE strategy found from Theorem 1.

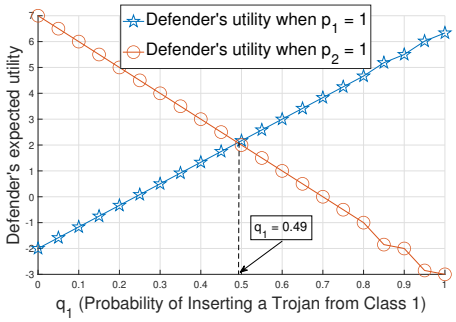

Fig. 1. Defender's expected utility versus the attacker's strategy.

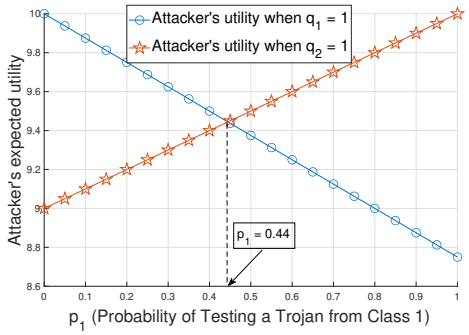

Fig. 2. Attacker's expected utility versus the defender's strategy.

In Figure 2, we show the attacker's expected utility versus the probability (p_1) of testing a Trojan from Class 1 when two Trojan classes (viz. Class 1 and Class 2) are present. For the figure, we consider V_1 = 10, V_2 = 10, F = 15, N_1 = 20, and N_2 = 25, and plot the attacker's utilities from always inserting a Trojan from Class 1 (i.e., from choosing $q_1 = 1$)

and from always inserting a Trojan from Class 2 (i.e., from choosing $q_2 = 1$). The point where the two utilities intersect implies that the expected utility of the attacker from inserting a Trojan from Class 1 equals that of the attacker from inserting a Trojan from Class 2 (as needed at the mixed strategy NE), which as can be seen from the figure, occurs at $p_1 = 0.44$ and which can be shown to tally with the defender's NE strategy found from Theorem 1. This corroborates Theorem 1.

Next, we present our game theoretic model and analysis when the defender can test an IC to check for the presence of a Trojan over multiple time slots.

III. Sequential Hardware Trojan Testing

In this section, we again consider the presence of M Trojan classes with each class i, $i \in \{1, \cdots, M\}$, containing N_i Trojan types and the attacker inserting a Trojan from class i into the IC with a probability q_i. We consider the defender to sequentially choose a Trojan from the available classes for testing with the defender being able to perform testing for at most K time slots. Specifically, we consider the defender to test the IC for the presence of a Trojan from class i with a probability p_i in every time slot until the IC tests positive for the presence of a Trojan or until the maximum number of time slots (K) available for testing is reached. We also consider that, as can be considered practical, if the IC tests negative for the presence of a specific Trojan type in a class in a certain time slot, the defender does not test the IC for the presence of that Trojan type in any subsequent time slot. The defender and the attacker is considered to uniformly pick a Trojan type for testing and insertion, respectively, from their chosen Trojan classes for simplicity of exposition. We also consider that if the defender tests the IC against the inserted Trojan while performing the sequential testing, the Trojan is detected, and the malicious manufacturer is imposed a fine F, and that an undetected Trojan of class i provides a benefit V_i to the attacker.

In the above game, denoting the strategy of the defender and the attacker as $\mathbf{p} = (p_1, \cdots, p_M)$ and $\mathbf{q} = (q_1, \cdots, q_M)$, respectively, the expected utility of the defender is,

$$E_D(\mathbf{p}, \mathbf{q}) = F + \sum_{i=1}^{M} (V_i + F)\left(\frac{Kp_i}{N_i} - 1\right) q_i \tag{15}$$

Denoting,

$$\gamma_i(p_i) = (V_i + F)\left(\frac{Kp_i}{N_i} - 1\right) \tag{16}$$

we can express (15) as

$$E_D(\mathbf{p}, \mathbf{q}) = F + \sum_{i=1}^{M} \gamma_i(p_i) q_i \tag{17}$$

The goal of the defender is to choose $\mathbf{p}$ to maximize (17) and that of the attacker is to choose $\mathbf{q}$ to minimize (17). It can be noted that when $\sum_{i=1}^{M} N_i < K$ (i.e., $\sum_{i=1}^{M} \frac{N_i}{K} < 1$), at NE, the defender chooses $p_i > \frac{N_i}{K}$ $\forall i \in \{1, \cdots, M\}$ (such that $\sum_{i=1}^{M} p_i \leq 1$) and the attacker chooses $q_i = 0$ $\forall i \in \{1, \cdots, M\}$. It is easy to verify that there does not exist any profitable unilateral deviation for the defender and the attacker from their aforementioned strategies under the above condition. In the following, we characterize the NE for the more challenging scenario when $\sum_{i=1}^{M} N_i \geq K$. We first prove some properties of $\gamma_i(p_i)$ (16).

Lemma 2. *$\gamma_i(p_i)$ (16) is an increasing function of p_i having the slope $(V_i + F)\frac{K}{N_i}$, with $\gamma_i(p_i) = 0$ when $p_i = \frac{N_i}{K}$.*

Proof. Clearly, $\frac{d(\gamma_i(p_i))}{dp_i} = (V_i + F)\frac{K}{N_i} > 0$. Again, equating $\gamma_i(p_i) = 0$ yields $p_i = \frac{N_i}{K}$. This proves the lemma. □

Next, we prove a condition that must be satisfied for the defender's strategy to form a NE when $\sum_{i=1}^{M} N_i \geq K$.

Lemma 3. *At NE, we must have,*

$$\gamma_i(p_i) = constant, \ \forall i \in \{1, \cdots, M\} \tag{18}$$

Proof. Consider a strategy profile $\mathbf{p} = (p_1, \cdots, p_M)$ such that $\sum_{i=1}^{M} p_i \leq 1$, denote $\underline{g} = \min_{i \in \{1, \cdots, M\}} \gamma_i(p_i)$, $\overline{g} = \max_{i \in \{1, \cdots, M\}} \gamma_i(p_i)$, and suppose that[2] $\underline{g} < \overline{g}$. Moreover, define the set $\underline{G} = \{i | i \in \{1, \cdots, M\} \text{ and } \gamma_i(p_i) = \underline{g}\}$ and the set $\overline{G} = \{i | i \in \{1, \cdots, M\} \text{ and } \gamma_i(p_i) = \overline{g}\}$. Consider also that $\sum_{i=1}^{M} N_i \geq K$ (i.e., $\sum_{i=1}^{M} \frac{N_i}{K} \geq 1$). In such a scenario, to satisfy $\sum_{i=1}^{M} p_i \leq 1$ it can be noted that we must have $\underline{g} < 0$. This is because, otherwise (if $\underline{g} \geq 0$), using Lemma 2, $\forall j \in \overline{G}$ we would have $p_j > \frac{N_j}{K}$ (i.e., $\gamma_j(p_j) > 0$) and $\forall i \in \{1, \cdots, M\} - \overline{G}$ we would have $p_i \geq \frac{N_i}{K}$ (i.e., $\gamma_i(p_i) \geq 0$) which implies $\sum_{i=1}^{M} p_i > 1$ making $\mathbf{p}$ infeasible.

Now, it should be noted that, since the attacker aims to minimize (17), the best response of the attacker against the strategy $\mathbf{p}$ is to adopt the strategy $\mathbf{q} = (q_1, \cdots, q_N)$ such that $\sum_{i \in \underline{G}} q_i = 1$. Consider now that $\overline{g} > 0$. In this case, $\forall i \in \underline{G}$ we have $p_i < \frac{N_i}{K}$ and $\forall j \in \overline{G}$ we have $p_j > \frac{N_j}{K}$ with $\sum_{i=1}^{M} p_i \leq 1$. Consider now $w \in \underline{G}$ for which $q_w > 0$ and $z \in \overline{G}$ and consider changing the strategy of the defender from $\mathbf{p} = (p_w, p_z, p_{-wz})$ to $\mathbf{p}' = (p_w + \delta, p_z - \delta, p_{-wz})$, where $0 < \delta \leq \min(1 - p_w, p_z)$ and p_{-wz} denotes the vector of probabilities with which the defender chooses the classes except for classes w and z. Now, we have $\gamma_w(p_w + \delta) > \gamma_w(p_w)$ (using Lemma 2) implying that $E_D(\mathbf{p}', \mathbf{q}) > E_D(\mathbf{p}, \mathbf{q})$ showing that there exists a profitable unilateral deviation for the defender from the strategy $\mathbf{p}$. Similarly, it can be shown that there exists a profitable unilateral deviation for the defender from the strategy $\mathbf{p}$ when $\overline{g} \leq 0$. From the above, it can be concluded that at NE we must have $\underline{g} = \overline{g}$ (i.e., $\gamma_1(p_1) = \gamma_2(p_2) = \cdots = \gamma_M(p_M)$), which proves the lemma. □

We next characterize the NE.

Theorem 2. *When $\sum_{i=1}^{M} N_i \geq K$, at NE,*

- *The defender's strategy corresponds to $\mathbf{p} = (p_1, \cdots, p_M) = (\frac{N_1}{K} - \delta_1, \cdots, \frac{N_M}{K} - \delta_M)$ where, for any chosen $i \in \{1, \cdots, M\}$, $\delta_i = \frac{\sum_{j=1}^{M} \frac{N_j}{K} - 1}{1 + \sum_{j=1, j \neq i}^{M} \frac{N_j}{N_i} \frac{V_i + F}{V_j + F}}$ and $\delta_j = \frac{N_j}{N_i} \frac{V_i + F}{V_j + F} \delta_i, \forall j \in \{1, \cdots, M\}, j \neq i$.*
- *The attacker's strategy corresponds to $\mathbf{q} = (q_1, \cdots, q_M)$, where, for any chosen $i \in \{1, \cdots, M\}$, $q_i =$*

[2] In other words, for such a strategy profile, there exists $i, j \in \{1, \cdots, M\}$ for which $\gamma_i(p_i) \neq \gamma_j(p_j)$

$\frac{\tau}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j}{N_i}\frac{V_i+F}{V_j+F}}$ *and* $q_j = \frac{V_i+F}{V_j+F}\frac{N_j}{N_i}q_i, \forall j \in \{1,\cdots,M\}, j \neq i$*, with* $\tau \in [0,1]$ *when in the defender's strategy* $\delta_k = 0\ \forall k \in \{1,\cdots,M\}$*, and* $\tau = 1$*, otherwise.*

Proof. It should be noted that Lemma 3 implies that the NE strategy $\mathbf{p}$ of the defender must be of the form $\mathbf{p} = (p_1,\cdots,p_M) = (\frac{N_1}{K}-\delta_1,\cdots,\frac{N_M}{K}-\delta_M)$ (since, otherwise, (18) will be violated). Moreover, for such a strategy to be feasible, we must have,

$$\sum_{i=1}^{M} p_i = \sum_{i=1}^{M}\frac{N_i}{K} - \delta_i = 1 \tag{19}$$

Further, from Lemma 3, at NE, $\forall i,j \in \{1,\cdots,M\}$, $i \neq j$, we must have

$$\gamma_i(p_i) = \gamma_j(p_j)$$

$$\Rightarrow (V_i+F)\frac{K}{N_i}\delta_i = (V_j+F)\frac{K}{N_j}\delta_j \quad \text{(using Lemma 2)}$$

which implies,

$$\frac{\delta_j}{\delta_i} = \frac{N_j}{N_i}\frac{V_i+F}{V_j+F},\ \forall i,j \in \{1,\cdots,M\}, i \neq j \tag{20}$$

Now, for any $i \in \{1,\cdots,M\}$, from (19), we have,

$$\delta_i + \sum_{j=1,j\neq i}^{M}\delta_j = \sum_{j=1}^{M}\frac{N_j}{K} - 1$$

$$\Rightarrow \delta_i + \sum_{j=1,j\neq i}^{M}\frac{N_j}{N_i}\frac{V_i+F}{V_j+F}\delta_i = \sum_{j=1}^{M}\frac{N_j}{K} - 1 \text{ (using (20))}$$

$$\Rightarrow \delta_i = \frac{\sum_{j=1}^{M}\frac{N_j}{K}-1}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j}{N_i}\frac{V_i+F}{V_j+F}} \tag{21}$$

Thus, if the defender, for any $i \in \{1,\cdots,M\}$, chooses δ_i according to (21) and δ_j, $\forall j \in \{1,\cdots,M\}, j \neq i$, according to (20), the strategy $\mathbf{p} = (p_1,\cdots,p_M) = (\frac{N_1}{K}-\delta_1,\cdots,\frac{N_M}{K}-\delta_M)$ both becomes feasible (i.e., satisfies $\sum_{i=1}^{M} p_i = 1$) as well as satisfies Lemma 3.

Now, since the aforementioned strategy of the defender makes $\gamma_i(p_i) = constant,\ \forall i \in \{1,\cdots,M\}$, several strategies of attacker can become a best response against the defender's strategy. However, not all such strategies of the attacker result in a NE since some may allow profitable unilateral deviations to exist for the defender from the strategy defined above. To prevent profitable unilateral deviations of the defender to exist, $\forall i,j \in \{1,\cdots,M\}$, $i \neq j$, we must have

$$(V_i+F)\frac{K}{N_i}q_i = (V_j+F)\frac{K}{N_j}q_j \tag{22}$$

which implies,

$$\frac{q_j}{q_i} = \frac{N_j}{N_i}\frac{V_i+F}{V_j+F},\ \forall i,j \in \{1,\cdots,M\}, i \neq j \tag{23}$$

Now, for the strategy $\mathbf{q} = (q_1,\cdots,q_M)$ to be feasible, we must have $\sum_{i=1}^{M} q_i = \tau$, $\tau \in [0,1]$, which, for any $i \in \{1,\cdots,M\}$, implies that,

$$q_i + \sum_{j=1,j\neq i}^{M} q_j = \tau$$

$$\Rightarrow q_i + \sum_{j=1,j\neq i}^{M}\frac{N_j}{N_i}\frac{V_i+F}{V_j+F}q_i = \tau \quad \text{(using (23))}$$

$$\Rightarrow q_i = \frac{\tau}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j}{N_i}\frac{V_i+F}{V_j+F}} \tag{24}$$

Now, in the defender's strategy, if $\delta_i = 0$ (i.e., $p_i = \frac{N_i}{K}$) $\forall i \in \{1,\cdots,M\}$, we would have $\gamma_i(p_i) = 0\ \forall i$, which implies that any $\tau \in [0,1]$ would form a best response of the attacker against the defender's strategy (since any strategy $\mathbf{q} = (q_1,\cdots,q_N)$, such that $0 \leq \sum_{i=1}^{M} q_i \leq 1$, would make the expected utility in (17) to be F). However, in the defender's strategy, if $\delta_i \neq 0\ \forall i \in \{1,\cdots,M\}$, $\tau = 1$ forms the attacker's best response (i.e., the attacker's best response becomes adopting $\mathbf{q}$ such that $\sum_{i=1}^{M} q_i = 1$). Thus, if the attacker, for any chosen $i \in \{1,\cdots,M\}$, chooses q_i according to (24) (with τ chosen appropriately as described above) and q_j, $\forall j \in \{1,\cdots,M\}$, $j \neq i$, according to (23), the attacker's strategy is feasible, the strategy forms a best response against the defender, and no profitable unilateral deviations for the defender exist. This proves the theorem. □

IV. Simulation Results

In this section, we provide simulation results to provide important insights into our developed game theoretic Trojan testing techniques. In Figure 3, we show the NE-based strategies of the attacker and the defender versus the number of Trojans (N_1) in Class 1 when two Trojan classes (Classes 1 and 2) are present with the defender testing the IC once for the presence of a Trojan. For the figure, we consider that N_2 = 20, V_1 = 10, V_2 = 10, and F = 25. The NE strategies in the figure are calculated using Lemma 1. As can be seen from the figure, as we increase the number of Trojans (N_1) in Class 1, the probabilities with which the attacker inserts a Trojan from Class 1 (q_1) and the defender tests a Trojan from Class 1 (p_1) at NE increase. This is because as N_1 increases, it becomes easier for the attacker to go undetected by inserting a Trojan from Class 1, making the attacker increase its probability (q_1) of inserting a Trojan from Class 1 with increasing N_1 at NE. Accordingly, as a best response, the defender also increases its probability of testing a Trojan from Class 1 at NE with increasing N_1.

In Figure 4, we show the NE-based strategies of the attacker and the defender versus the benefit V_1 acquired by the attacker from an undetected Trojan from Class 1 (in other words, the damage sustained by the defender when a Trojan from Class 1 goes undetected) considering two Trojan classes (Classes 1 and 2) to be present. We again consider the defender to test the IC once for the presence of a Trojan. For the figure, we consider N_1 = 20, N_2 = 20, V_2 = 10, and F = 25. The NE strategies in the figure are calculated using Lemma 1. As can be seen from the figure, and as is also intuitive, with the increase of V_1, i.e., as Trojans in Class 1 become more damaging in nature, the probability (p_1) with which the defender tests a Trojan from

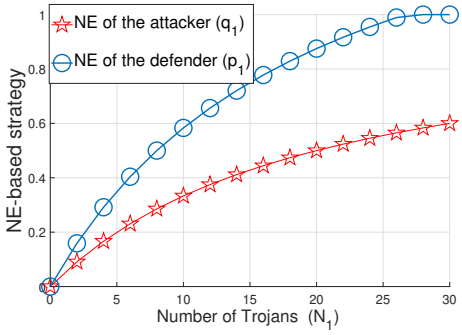

Fig. 3. NE-based strategy of the attacker and the defender versus the number of Trojans (N_1) in Class 1.

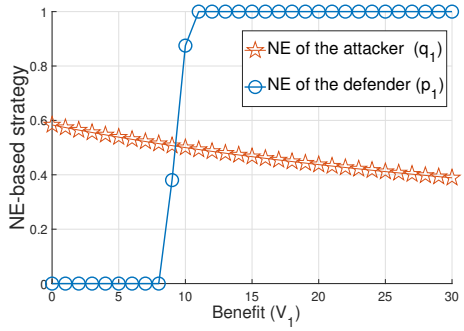

Fig. 4. NE-based strategy of the attacker and the defender versus V_1.

Class 1 at NE shows a non-decreasing trend. Accordingly, the attacker, with increasing V_1, as a best response decreases its probability (q_1) of inserting a Trojan from Class 1 at NE.

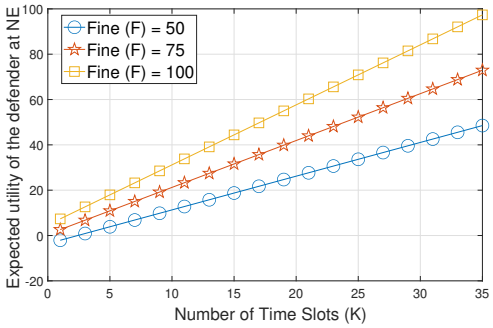

Fig. 5. Expected utility of the defender versus the number of time slots (K).

In Figure 5, we show the expected utility of the defender (15) at NE versus the number of time slots (K) available for sequential hardware Trojan testing considering three classes of Trojans to be present. For the figure, we consider N_1 = 10, N_2 = 12, N_3 = 14, V_1 = 10, V_2 = 15, and V_3 = 20. The NE strategies of the defender and the attacker were computed using Theorem 2. As can be seen from the figure, for any given fine (F), as the number of time slots (K) increases, the expected utility of the defender at NE increases, since with increasing K, the defender's chances of detecting the inserted Trojan increases. Moreover, as can be seen from the figure, and as is also intuitive, for any given K, the expected utility of the defender increases at NE as the fine (F) increases.

V. Conclusion

This paper presented techniques for employing game theory-based hardware Trojan testing. The presented game models considered the presence of multiples classes of Trojans, with each class containing multiple Trojan types. In such a scenario, the paper first characterized the NE strategy for inserting a Trojan (from the perspective of a malicious manufacturer) and testing a Trojan (from the perspective of a defender) when the defender can test the IC for the presence of a Trojan once. Moreover, the paper also presented a sequential hardware Trojan testing game, where the defender sequentially chooses Trojans for testing over time, and characterized its NE. Numerous simulation results were presented to gain insights into the presented game theoretic Trojan testing techniques.

References

[1] S. Bhunia, M. S. Hsiao, M. Banga and S. Narasimhan, "Hardware Trojan Attacks: Threat Analysis and Countermeasures," in Proceedings of the IEEE, 2014, vol. 102, no. 8, pp. 1229-1247

[2] R. S. Chakraborty, S. Narasimhan and S. Bhunia, "Hardware Trojan: Threats and emerging solutions," IEEE International High Level Design Validation and Test Workshop, San Francisco, CA, 2009, pp. 166-171.

[3] T. E. Schulze, K. Kwiat, C. Kamhoua, Shih-Chieh Chang and Yiyu Shi, "RECORD: Temporarily Randomized Encoding of Combinational Logic for Resistance to Data Leakage from hardware Trojan," IEEE Asian Hardware-Oriented Security and Trust (AsianHOST), Yilan, 2016, pp. 1-6.

[4] M. Banga and M. S. Hsiao, "A region based approach for the identification of hardware Trojans," IEEE International Workshop on Hardware-Oriented Security and Trust, Anaheim, CA, 2008, pp. 40-47.

[5] H. Salmani, M. Tehranipoor and J. Plusquellic, "New design strategy for improving hardware Trojan detection and reducing Trojan activation time," IEEE International Workshop on Hardware-Oriented Security and Trust, Francisco, CA, 2009, pp. 66-73.

[6] J. Rajendran, E. Gavas, J. Jimenez, V. Padman and R. Karri, "Towards a comprehensive and systematic classification of hardware Trojans," IEEE Intl. Sym. on Circuits and Systems, Paris, 2010, pp. 1871-1874.

[7] C. A. Kamhoua, H. Zhao, M. Rodriguez and K. A. Kwiat, "A Game-Theoretic Approach for Testing for Hardware Trojans," IEEE Trans. on Multi-Scale Comp. Systems, vol. 2, no. 3, pp. 199-210, 2016.

[8] J. Graf, W. Batchelor, S. Harper, R. Marlow, E. Carlisle IV and P. Athanas, "A practical application of game theory to optimize selection of hardware Trojan detection strategies," in J Hardw Syst Secur, vol. 4, pp. 98-119, 2020.

[9] J. Graf, "Trust games: How game theory can guide the development of hardware Trojan detection methods," 2016 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), McLean, VA, 2016, pp. 91-96.

[10] D. Fudenberg and J. Tirole, "Game Theory," MIT press, Cambridge, MA, 1991.

[11] S. Bhasin and F. Regazzoni, "A survey on hardware trojan detection techniques," IEEE International Symposium on Circuits and Systems (ISCAS), Lisbon, 2015, pp. 2021-2024.

[12] D. Agarwal, S. Baktir, D. Karakoy, P. Rohatgi and B. Sunar, "Trojan Detection using IC Fingerprinting," IBM Research Report, 2006.

[13] R. S. Chakraborty, F. Wolff, S. Paul, C. Papachristou and S. Bhunia, "MERO: A statistical approach for hardware Trojan detection," in Proc. Workshop Cryptograph. Hardware Embedded Syst., 2009, pp. 396-410.

[14] N. Hu, M. Ye and S. Wei, "Surviving Information Leakage Hardware Trojan Attacks Using Hardware Isolation" in IEEE Transactions on Emerging Topics in Computing, vol. 7, no. 02, pp. 253-261, 2019.

[15] Swarup Bhunia and Mark Tehranipoor. "Hardware Security: A Hands-on Learning Approach," Morgan Kaufmann Publishers Inc. (1st edition), CA, 2018, USA.

[16] H. Xue, S. Li and S. Ren, "Power analysis-based Hardware Trojan detection," IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, 2017, pp. 253-257.

[17] R. Shathanaa and N. Ramasubramanian, "Improving Power and Latency Metrics for Hardware Trojan Detection During High Level Synthesis," 9th IEEE Intl. Conf. on Computing, Communication and Networking Technologies (ICCCNT), Bangalore, 2018, pp. 1-7.

[18] K. Kwiat and F. Born, "Strategically Managing the Risk of Hardware Trojans Through Augmented Testing," 13th Annual Symposium on Information Assurance (ASIA 2018).