Hardware Trojan Testing with Hierarchical Trojan Types Under Cognitive Biases

Satyaki Nan
Department of Computing
Georgia Southwestern State University
Americus, GA, USA
satyaki.nan@gsw.edu

Swastik Brahma
Department of Computer Science
University of Cincinnati
Cincinnati, OH, USA
brahmask@ucmail.uc.edu

2024 58th Annual Conference on Information Sciences and Systems (CISS) | 979-8-3503-6929-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISS59072.2024.10480192

***Abstract*— In this paper, we consider the problem of testing Integrated Circuits (ICs) to check for the presence of hardware Trojans while diligently accounting for the hierarchical classification structure of Trojans, the error-prone nature of testing processes, and the strategic mindsets and behavioral irrationalities (cognitive biases) of buyers and manufacturers of ICs. As shown in the paper, such factors greatly impact the design of Trojan insertion and testing strategies. Under a hierarchy of Trojan types and testing imperfections, the paper first analytically characterizes Trojan insertion-testing strategies at Nash Equilibrium (NE) considering a buyer (defender) and malicious manufacturer (attacker) to be strategic and rational in nature. Then, the paper analytically characterizes such strategies when the involved entities are strategic but irrational in nature. Among others, results presented in the paper emphasize the asymmetric nature of the impact of behavioral irrationalities on the defender's and attacker's utilities. The paper also presents numerous simulation results to gain important insights into our analytically characterized Trojan insertion-testing strategies.**

***Index Terms*—Hardware Trojan, Game Theory, Prospect Theory.**

I. Introduction

The detection of hardware Trojans in Integrated Circuits (ICs) is an important problem and has received attention in the past [1]–[7]. For example, in [4], the authors propose a region-based partitioning and excitation approach to accurately detect the location of Trojans in ICs. In [7], the authors generate test patterns that can distinguish between the power profile of a genuine IC and the Trojan counterpart, but their effectiveness is limited in terms of the manufacturing processes, behaviors, and sizes of the Trojans. Again, in [6], the authors propose a methodology, referred to as MERO (Multiple Excitation of Rare Occurrence), for statistical test generation that maximizes the probability of detecting inserted Trojans. Since exhaustive testing of all possible Trojan types can be cost ineffective, the works in [8]–[14] model the detection of hardware Trojans using Game Theory [15] to determine which Trojan types to test against a strategic malicious manufacturer. For example, [8] uses software-based techniques to analyze Nash Equilibrium (NE)-based Trojan insertion-testing strategies while [12]–[14] analytically characterize such strategies.

It should be noted that the categorization of Trojans follows a *hierarchical structure* that consists of multiple Trojan classes with each class containing multiple Trojan types. For example, a class of Trojans corresponds to one that leaks information from cyber systems with the class containing various types of such Trojans. We refer the reader to [1] for a treatise on the classification of Trojans. Further, it should be noted that testing of ICs is an *error-prone* process which can fail to detect the presence of Trojans in ICs that contain them. Moreover, it should be noted that Trojan insertion and IC testing processes can naturally become impacted both by *strategic mindsets* as well as by *behavioral irrationalities* (cognitive biases) of buyers and manufacturers of ICs, who are ultimately human decision-makers. *While the aforementioned factors* ***together*** *influence Trojan insertion-testing strategies, there is, however, a lack of literature that considers them in concert.* We aim to fill this void in this paper.

Specifically, in this paper, using Game Theory to model the strategic mindsets of a buyer and a manufacturer, and Prospect Theory [16] to model their cognitive biases, we analytically characterize NE-based Trojan insertion-testing strategies under consideration of the hierarchical classification structure of Trojans and the error-prone nature of testing processes. The main contributions of the paper are as follows:

- Under a hierarchy of Trojan types and testing imperfections, we first employ Game Theory to analytically characterize Trojan insertion-testing strategies at NE considering a buyer (defender) and malicious manufacturer (attacker) to be strategic and rational in nature.
- Further, under a hierarchy of Trojan types and testing imperfections, we then employ both Game Theory and Prospect Theory to analytically characterize Trojan insertion-testing strategies at NE considering the defender and attacker to be strategic but irrational in nature.
- Extensive simulation results are provided to gain important insights into our analytically characterized strategies.

The rest of the paper is organized as follows. Section II presents our characterized game theoretic Trojan insertion-testing strategies considering the defender and attacker to be strategic and rational in nature while Section III presents such strategies when the involved entities are strategic but irrational in nature. Section IV presents simulation results to provide insights into analytically characterized strategies. Finally, section V concludes the paper.

This work was supported by the U.S. National Science Foundation (NSF) under Award Number CCF-2302197 and by the University of Cincinnati (UC).

II. Game Theoretic 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. E.g., Class 1 can contain *information leaking* Trojans (e.g., [17], [18]) while Class 2 can contain those that increase the *power consumption* of a device (e.g., [19]).

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. Further, consider that the buyer of the IC, whom we refer to as the defender (D), tests the IC for the presence of a Trojan from Class 1 with a probability p_1 (and tests the IC for the presence of a Trojan from Class 2 with a probability $(1 - p_1)$). For simplicity of exposition, consider the defender and the attacker to uniformly pick a Trojan type for testing and insertion, respectively, from their chosen Trojan classes. To model imperfections of the testing process, consider that when the defender tests the acquired IC against a Trojan type that was inserted by the attacker, the Trojan is detected with a probability P_d. Also consider that when the acquired IC tests positive for the presence of a Trojan, 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 fails to detect the inserted Trojan, i.e., either tests the IC against a Trojan type which was not inserted by the attacker or tests the IC against the inserted Trojan type but the error-prone nature of the conducted test fails to detect it, the buyer installs the acquired IC resulting in the undetected Trojan of class i, $i \in \{1, 2\}$, to provide 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 acquired IC for the presence of a Trojan from Class* 1 *with a probability* $p_1 = \frac{1+\frac{N_2(V_1-V_2)}{P_d(F+V_2)}}{1+\frac{N_2(F+V_1)}{N_1(F+V_2)}}$ *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 from testing the IC for the presence of a Trojan from Class 1 is

$$E_D^1 = \left[\frac{(FP_d - (1-P_d)V_1)}{N_1} - V_1\left(\frac{N_1 - 1}{N_1}\right)\right]q_1 - V_2(1-q_1) \tag{1}$$

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

$$E_D^2 = \left[\frac{(FP_d - (1-P_d)V_2)}{N_2} - V_2\left(\frac{N_2 - 1}{N_2}\right)\right](1-q_1) - V_1 q_1 \tag{2}$$

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

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

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

$$E_A^1 = \left[\frac{(-FP_d + (1-P_d)V_1)}{N_1} + V_1\left(\frac{N_1 - 1}{N_1}\right)\right]p_1 + V_1(1-p_1) \tag{4}$$

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

$$E_A^2 = \left[\frac{(-FP_d + (1-P_d)V_2)}{N_2} + V_2\left(\frac{N_2 - 1}{N_2}\right)\right](1-p_1) + V_2 p_1 \tag{5}$$

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

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

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 class $i \in \{1, \cdots, M\}$ containing N_i types of Trojans. We denote the strategy of the attacker as $\mathbf{q} = (q_1, \cdots, q_M)$ 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. Further, we denote the strategy of the defender as $\mathbf{p} = (p_1, \cdots, p_M)$ 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. As before, we consider P_d to be the probability with which the inserted Trojan gets detected when the defender tests the acquired IC against the inserted Trojan type, F to be the fine that is imposed on the attacker upon detecting a Trojan in its sold IC, and V_i to be the damage that is sustained by the defender upon failing to detect an inserted Trojan of class $i \in \{1, \cdots, M\}$. Next, we characterize the mixed strategy Nash Equilibrium (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* $p_i = \frac{1+\sum_{j=1}^{M} \frac{N_j(V_i-V_j)}{P_d(F+V_j)}}{1+\sum_{j=1, j \neq i}^{M} \frac{N_j(F+V_i)}{N_i(F+V_j)}}$ *and tests the IC for the*

presence of a Trojan from class j with a probability $p_j = \frac{\frac{P_d(F+V_i)}{N_i}p_i+(V_j-V_i)}{\frac{P_d(F+V_j)}{N_j}}$, $\forall j \in \{1,\cdots,M\}, j \neq i$,

- *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 a 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 (E_D^i) of D from testing the IC for the presence of a Trojan from class i, $i \in \{1,\cdots,M\}$, is

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

At equilibrium, 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

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

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

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

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

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

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).

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

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

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

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

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

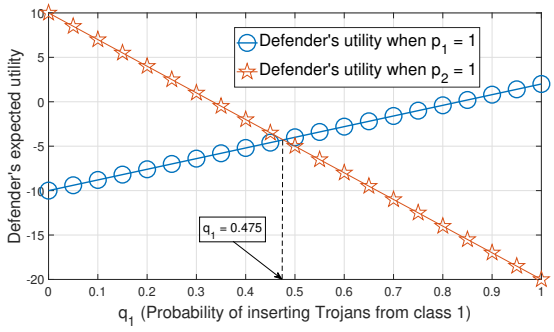
$$p_i + \sum_{j=1,j\neq i}^{M} p_j = 1 \tag{13}$$

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

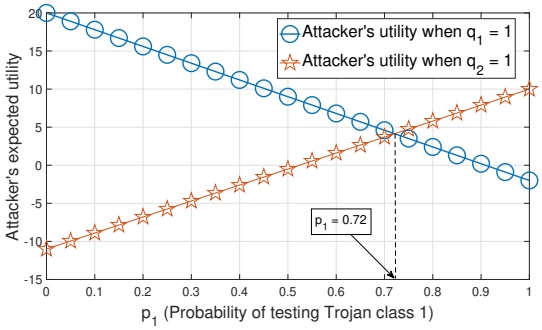
$$\Rightarrow p_i = \frac{1+\sum_{j=1}^{M}\frac{N_j(V_i-V_j)}{P_d(F+V_j)}}{1+\sum_{j=1,j\neq i}^{M}\frac{N_j(F+V_i)}{N_i(F+V_j)}} \tag{14}$$

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. □



(a)



(b)

Fig. 1. Expected utilities of the defender and the attacker versus their opponents' strategies

Next, we provide numerical results to corroborate the above results considering two Trojan classes, viz. Class 1 and Class 2. In Fig. 1(a), we 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 = 5, N_2 = 5, V_1 = 20, V_2 = 10, and F = 200, P_d = 0.5 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 $q_1 = 0.475$ and which can be shown to tally with the attacker's NE strategy found from Theorem 1.

In Fig.1(b), we show the attacker's expected utility versus the probability (p_1) of testing a Trojan from Class 1. For the figure, we consider N_1 = 5, N_2 = 5, V_1 = 20, V_2 = 10, and F = 200, P_d = 0.5 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 $p_1 = 0.72$ and which can be shown to tally with the defender's NE strategy found from Theorem 1. This corroborates Theorem 1.

III. Hardware Trojan Testing with Cognitively Biased Defender and Attacker

In this section, we consider Trojan testing when the defender and the attacker, in addition to acting in a strategic manner, are cognitively biased in nature. To address such a scenario, we have developed a game *and* prospect theoretic Trojan insertion-testing model. We first provide a brief overview of Prospect Theory [16], which provides a descriptive model of human cognitive biases, before describing our model.

A. Prospect Theory

In prospect theory [16], humans, due to their *cognitive biases*, do not weigh outcomes by their objective probabilities but rather by transformed distorted probabilities subjectively. The transformation of probabilities is computed using a weighting function $w(.)$ whose argument is an objective probability. In this paper, to model the over-weighting/under-weighting of objective probabilities, we use the Prelec function [20], which is known to be a well-accepted model of human behavior having empirical evidence. Specifically, for an objective probability p, the Prelec function is defined as

$$w(p) = exp(-(-\log p)^{\alpha}), \quad 0 < \alpha \leq 1 \tag{15}$$

where α is a parameter that models how a human subjectively distorts an objective probability. For illustration, Fig. 2 plots $w(p)$ against p for different values of α.

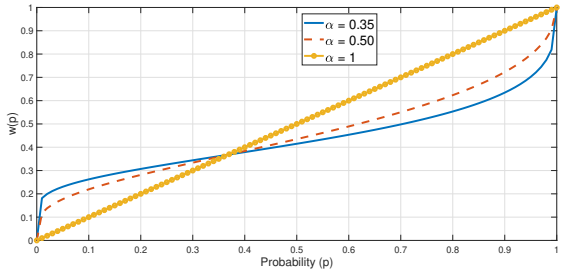


Fig. 2. Behavior of Prelec function

Based on the subjective distortion of probabilities, a cognitively biased human agent's *prospect theoretic* utility from a gamble that can lead to outcomes having valuations $x_1, x_2, \cdots, x_N$ with probabilities $p_1, p_2, \cdots, p_N$, respectively, is $\sum_{i=1}^{N} x_i w(p_i)$, which clearly deviates from norms followed by conventional *expected* utility theoretic models. In the following, we account for such deviations in our analysis of Trojan insertion-testing under strategic considerations while accounting for the hierarchical classification structure of Trojans and the error-prone nature of testing processes.

B. Prospect Theoretic Trojan Testing

We consider a similar game model as described in Section II-A, with the attacker and the defender, however, considered cognitively biased who subjectively perceive objective probabilities (using (15)) to obtain prospect theoretic utilities from their chosen strategies. In such a scenario, in the next theorem, we characterize the mixed strategy NE of the game over the defender's and the attacker's strategy spaces.

THEOREM 2. *In the presence of M classes of Trojans with class $i \in \{1, \cdots, M\}$ containing N_i Trojan types, when the defender and attacker are cognitively biased in nature,*

- *the defender's strategy $(p_1, \cdots, p_M)$ at NE corresponds to, for any chosen class $i \in \{1, \cdots, M\}$, the roots of the following M equations solved simultaneously:*

$$\left(\frac{F}{N_j}\right) w(P_d p_j) - \left(\frac{F}{N_i}\right) w(P_d p_i) + \left(\frac{V_i}{N_i}\right) w\big(p_i(1-P_d)\big) - \left(\frac{V_j}{N_j}\right) w\big(p_j(1-P_d)\big) + \left(\frac{V_i(N_i-1)}{N_i}\right) w(p_i) - \left(\frac{V_j(N_j-1)}{N_j}\right) w(p_j) + V_i w(1-p_i) - V_j w(1-p_j) = 0 \tag{16a}$$

$$\forall j \in \{1, \cdots, i-1, i+1, \cdots, M\}$$

$$p_i + \sum_{j=1, j\neq i}^{M} p_j = 1 \tag{16b}$$

- *the attacker's strategy $(q_1, \cdots, q_M)$ at NE corresponds to, for any chosen class $i \in \{1, \cdots, M\}$, the roots of the following M equations solved simultaneously:*

$$\left(\frac{F}{N_i}\right) w(P_d q_i) - \left(\frac{F}{N_j}\right) w(P_d q_j) - \left(\frac{V_i}{N_i}\right) w\big(q_i(1-P_d)\big) + \left(\frac{V_j}{N_j}\right) w\big(q_j(1-P_d)\big) - \left(\frac{V_i(N_i-1)}{N_i}\right) w(q_i) + \left(\frac{V_j(N_j-1)}{N_j}\right) w(q_j) + V_i w(q_i) - V_j w(q_j) = 0 \tag{17a}$$

$$\forall j \in \{1, \cdots, i-1, i+1, \cdots, M\}$$

$$q_i + \sum_{j=1, j\neq i}^{M} q_j = 1 \tag{17b}$$

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

$$PT_D^i = \left(\frac{F}{N_i}\right) w(P_d q_i) - \left(\frac{V_i}{N_i}\right) w\big(q_i(1-P_d)\big) - \left(\frac{V_i(N_i-1)}{N_i}\right) w(q_i) + \sum_{j=1, j\neq i}^{M} (-V_j) w(q_j) \tag{18}$$

At the mixed strategy NE, we must have $PT_D^1 = PT_D^2 = \cdots = PT_D^M$, for equating $PT_D^i = PT_D^j$ can be shown to yield (17a). Clearly, for any chosen $i \in \{1, \cdots, M\}$, (17a) must hold $\forall j \in \{1, \cdots, i-1, i+1, \cdots, M\}$ to make the defender indifferent over its entire undominated strategy space while ensuring (17b) to ensure the feasibility of the attacker's strategy.

Now, the prospect theoretic utility (say, PT_A^i) of A from inserting a Trojan from class i, $i \in \{1, \cdots, M\}$, is

$$PT_A^i = -\left(\frac{F}{N_i}\right) w(P_d p_i) + \left(\frac{V_i}{N_i}\right) w\big(p_i(1-P_d)\big) + \left(\frac{V_i(N_i-1)}{N_i}\right) w(p_i) + V_i w(1-p_i) \quad (19)$$

At the mixed strategy NE, since we must have $PT_A^1 = PT_A^2 = \cdots = PT_A^M$, for $i, j \in \{1, \cdots, M\}$, $i \neq j$, equating $PT_A^i = PT_A^j$ can be shown to yield (16a). Clearly, for any chosen $i \in \{1, \cdots, M\}$, (16a) must hold $\forall j \in \{1, \cdots, i-1, i+1, \cdots, M\}$ to make the attacker indifferent over its entire undominated strategy space while ensuring (16b) to ensure the feasibility of the defender's strategy. This proves the theorem. □

It can be noted that Matlab's `fzero` tookit [21] can be used to simultaneously solve the equations in Theorem 2 in a computationally efficient manner. In the following remark, we provide a closed-form characterization of the NE presented in Theorem 2 for a special case.

REMARK 1. *Consider $V_1 = V_2 = \cdots = V_M$ and $N_1 = N_2 = \cdots = N_M$. In such a scenario, the NE presented in Theorem 2 simplifies to $p_i = \frac{1}{M}$ and $q_i = \frac{1}{M}$, $\forall i \in \{1, \cdots, M\}$.*

LEMMA 2. *The NE strategies characterized in Theorem 2 exist.*

Proof. Let us denote (17a) as

$$f(q_i) = \left(\frac{F}{N_i}\right) w(P_d q_i) - \left(\frac{F}{N_j}\right) w(P_d q_j) - \left(\frac{V_i}{N_i}\right) w\big(q_i(1-P_d)\big) + \left(\frac{V_j}{N_j}\right) w\big(q_j(1-P_d)\big) - \left(\frac{V_i(N_i-1)}{N_i}\right) w(q_i) + \left(\frac{V_j(N_j-1)}{N_j}\right) w(q_j) + V_i w(q_i) - V_j w(q_j) = 0 \quad (20)$$

It can be shown that $df(q_i)/dq_i \geq 0$ which implies that $f(q_i)$ is a monotonically increasing function of q_i. Further, we have $\lim_{q_i \to 0} f(q_i) < 0$ and $\lim_{q_i \to 1} f(q_i) > 0$. Thus, for $i, j \in \{1, \cdots, M\}$, $i \neq j$, we can conclude that for any given q_j, there exists a value of q_i at which $f(q_i) = 0$ i.e., which satisfies (17a). Similarly, it can be shown that, for $i, j \in \{1, \cdots, M\}$, $i \neq j$, for any given p_j, there exists a value of p_i which satisfies (16a). This proves the lemma. □

In Fig. 3, considering $M = 2$, we show the nature of $f(q_i)$ (20) w.r.t. q_i considering F = 200, V_1 = 50, V_2 = 80, N_1 = 5, N_2 = 4, α = 0.4, and P_d = 0.5. The figure corroborates the aforementioned nature of $f(q_i)$ (20) w.r.t. q_i and that there exists q_i such that $f(q_i)$ = 0. This corroborates Lemma 2.

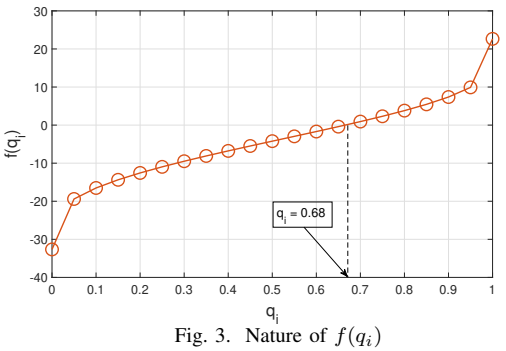


Fig. 3. Nature of $f(q_i)$

IV. SIMULATION RESULTS

In this section, we provide simulation results to provide important insights into our developed game and prospect theoretic Trojan testing techniques. In Fig. 4, 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. For the figure, we consider that N_2 = 5, V_1 = 40, V_2 = 45, P_d = 0.3, and F = 100. The NE strategies in the figure have been calculated using Theorem 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.

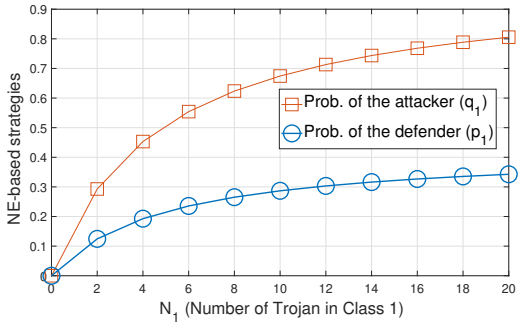


Fig. 4. NE-based strategy of the attacker and the defender versus Fine (F).

In Fig. 5, we show the prospect theoretic utilities of the defender and attacker at NE against a varying probability of detection (P_d). For the figure, we consider N_1 = 5, N_2 = 4, V_1 = 20, V_2 = 10, α = 0.5, and F = 200. The NE strategies in the figure have been calculated using Theorem 2. As can be seen from the figure, and as is also intuitive, as P_d increases, i.e., as the tests conducted by the defender becomes more accurate, the defender's utility increases, and that of the attacker decreases.

In Fig. 6, we show the prospect theoretic utilities of the defender and attacker versus the probability distortion coefficient (α) in the Prelec function (15). For the figure, we consider

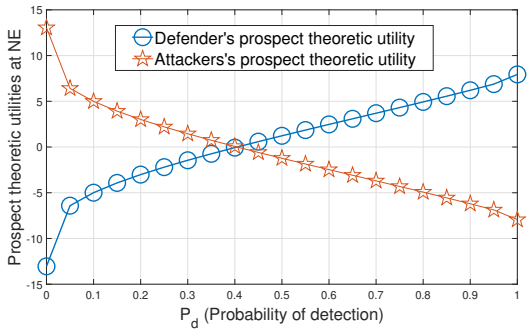


Fig. 5. Prospect theoretic Utility Versus P_d.

$N_1 = 5$, $N_2 = 4$, $V_1 = 20$, $V_2 = 10$, $P_d = 0.5$, and $F = 200$. The NE strategies in the figure have been calculated using Theorem 2. As can be seen from the figure, as α increases, i.e., as the defender and the attacker become less cognitive biased (more rational), the defender's prospect theoretic utility decreases while that of the attacker increases. This shows that a higher degree of rationality can be better exploited by the attacker to optimize the attack than it can be employed by the defender to adopt its best response defense.

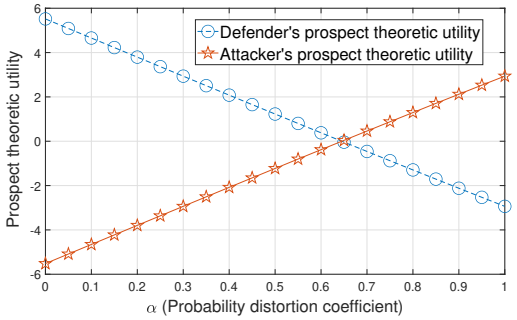


Fig. 6. Prospect theoretic utility versus α (Probability distortion coefficient).

V. Conclusion

This paper considered the problem of hardware Trojan testing under consideration of the hierarchical classification structure of Trojans and the error-prone nature of testing processes while accounting for the strategic mindsets and behavioral irrationalities of buyers and manufacturers of ICs. In such a scenario, the paper first analytically characterized NE-based Trojan insertion-testing strategies considering a buyer and malicious manufacturer to be strategic and rational in nature. Then, the paper analytically characterized such strategies considering the buyer and malicious manufacturer to be strategic but irrational in nature. Numerous simulation results have been presented to provide important insights into our analytically characterized strategies.

References

[1] S. Bhunia, M. S. Hsiao, M. Banga, and S. Narasimhan, "Hardware trojan attacks: Threat analysis and countermeasures," *Proceedings of the IEEE*, vol. 102, no. 8, pp. 1229–1247, 2014.

[2] R. S. Chakraborty, S. Narasimhan, and S. Bhunia, "Hardware trojan: Threats and emerging solutions," in *IEEE International High Level Design Validation and Test Workshop*, 2009, pp. 166–171.

[3] T. E. Schulze, K. Kwiat, C. Kamhoua, S.-C. Chang, and Y. Shi, "Record: Temporarily randomized encoding of combinational logic for resistance to data leakage from hardware trojan," in *2016 IEEE Asian Hardware-Oriented Security and Trust (AsianHOST)*, pp. 1–6.

[4] M. Banga and M. S. Hsiao, "A region based approach for the identification of hardware trojans," in *IEEE International Workshop on Hardware-Oriented Security and Trust*, 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," in *IEEE International Workshop on Hardware-Oriented Security and Trust*, 2009, pp. 66–73.

[6] R. S. Chakraborty, F. Wolff, S. Paul, C. Papachristou, and S. Bhunia, "Mero: A statistical approach for hardware trojan detection," in *Intl. Workshop on Cryptographic Hardware and Embedded Sys.*, 2009.

[7] D. Agrawal, S. Baktir, D. Karakoyunlu, P. Rohatgi, and B. Sunar, "Trojan detection using ic fingerprinting," in *IEEE Symposium on Security and Privacy (SP '07)*, 2007, pp. 296–310.

[8] C. A. Kamhoua, H. Zhao, M. Rodriguez, and K. A. Kwiat, "A game-theoretic approach for testing for hardware trojans," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 3, pp. 199–210, 2016.

[9] J. Graf, W. Batchelor, S. Harper, R. Marlow, E. Carlisle, and P. Athanas, "A practical application of game theory to optimize selection of hardware trojan detection strategies," *Journal of Hardware and Sys. Sec.*, 2020.

[10] J. Graf, "Trust games: How game theory can guide the development of hardware trojan detection methods," in *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2016, pp. 91–96.

[11] K. Kwiat and F. Born, "Strategically managing the risk of hardware trojans through augmented testing," in *13th Annual Symposium on Information Assurance (ASIA)*, 2018, pp. 20–24.

[12] S. Brahma, L. Njilla, and S. Nan, "Game theoretic hardware trojan testing under cost considerations," in *International Conference on Decision and Game Theory for Security*. Springer, 2021, pp. 251–270.

[13] S. Nan, L. Njilla, S. Brahma, and C. A. Kamhoua, "Game and prospect theoretic hardware trojan testing," in *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, 2023, pp. 1–6.

[14] S. Brahma, S. Nan, and L. Njilla, "Strategic hardware trojan testing with hierarchical trojan types," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021, pp. 1–6.

[15] D. Fudenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[16] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 99–127.

[17] N. Hu, M. Ye, and S. Wei, "Surviving information leakage hardware trojan attacks using hardware isolation," *IEEE Transactions on Emerging Topics in Computing*, vol. 7, no. 2, pp. 253–261, 2017.

[18] Y. Zhao, X. Hu, S. Li, J. Ye, L. Deng, Y. Ji, J. Xu, D. Wu, and Y. Xie, "Memory trojan attack on neural network accelerators," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2019, pp. 1415–1420.

[19] X. Wang, "Hardware trojan attacks: Threat analysis and low-cost countermeasures through golden-free detection and secure design," Ph.D. dissertation, Case Western Reserve University, 2014.

[20] D. Prelec, "The probability weighting function," *Econometrica*, 1998.

[21] https://www.mathworks.com/help/matlab/ref/fzero.html.