Game and Prospect Theoretic Hardware Trojan Testing

Satyaki Nan
Dept. of Comp. Science
Tennessee State University
Nashville, TN, USA
snan@tnstate.edu

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

Swastik Brahma
Dept. of Comp. Science
University of Cincinnati
Cincinnati, OH, USA
brahmask@ucmail.uc.edu

Charles A. Kamhoua
Network Security Branch
US Army Research Lab.
Adelphi, MD, USA
charles.a.kamhoua.civ@mail.mil

Abstract— **In this paper, we address the problem of hardware Trojan testing with the buyer of an Integrated Circuit (IC), who is referred to as the defender, and the malicious manufacturer of the IC, who is referred to as the attacker, strategically acting against each other. Our developed model accounts for both imperfections in the testing process as well as costs incurred for performing testing. First, we analytically characterize Nash Equilibrium (NE) strategies for Trojan insertion and testing from the attacker's and the defender's perspectives, respectively, considering them to be fully rational in nature. Further, we also characterize NE-based Trojan insertion-testing strategies considering the attacker and the defender to have cognitive biases which make them exhibit irrationalities in their behaviors. Numerous simulation results are presented throughout the paper to provide important insights.**

I. Introduction

The presence of hardware Trojans in Integrated Circuits (IC) pose a serious threat to the semiconductor industry. In such threats, a malicious manufacturer alters the design of an IC to attain malign objectives such as leakage of sensitive system information, degradation of system performance, and even complete disruption of system operation [1]. Mitigation of hardware Trojan threats requires *testing* acquired ICs to check for the presence of Trojans in them [1]–[5]. For example, the authors in [2] have designed sequences of test patterns that can generate noticeable differences between the power profile of a genuine IC and its Trojan counterpart for the detection of Trojans, but the effectiveness of the proposed scheme is limited in terms of the manufacturing processes, behaviors, and the sizes of the inserted Trojans. Again, a region-based partitioning scheme for detecting Trojans has been proposed in [3]. Further, in [5], the authors propose a methodology, referred to as MERO, to optimize the probability of detecting inserted Trojans using statistical methods.

Since a malicious manufacturer can act in a strategic manner while inserting Trojans, the authors in [6]–[11] design game theoretic [12] hardware Trojan testing techniques to account for strategic interactions. However, while such works provide useful insights, nevertheless, game theoretic models and analytical approaches developed by past work suffer from various limitations. For example, the work in [8] analyzes a two-player Trojan detection game, but limits investigation of the equilibrium to an example scenario of the model. Again, the approaches in [7], [9] rely on the use of software-based techniques for analyzing game theoretic testing strategies. Analytical characterizations of testing strategies at Nash Equilibrium (NE) can be found in [10], which, however, ignores the costs incurred in the testing process. Moreover, all the aforementioned works, including [11], overlook *imperfections* of the testing process as well as ignore the *human factors* involved in performing Trojan insertion-testing, both of which can greatly impact attack-defense strategies. *We aim to overcome such limitations in this paper.*

Specifically, in this paper, we analytically characterize NE-based strategies for Trojan insertion-testing under considerations of both costs incurred for performing testing and imperfections of the testing process. Further, we also address human *cognitive biases* of the defender and the attacker in our characterization of Trojan insertion-testing strategies. The main contributions of the paper are as follows:

- We analytically characterize NE-based strategies for Trojan insertion-testing under consideration of both testing costs and testing imperfections.
- Further, in addition to adopting a game theoretic perspective, we employ Prospect Theory [13] to model cognitive biases of the defender and the attacker (manufacturer) and characterize NE-based strategies for Trojan insertion-testing under the resulting behavioral irrationalities.
- We present numerous simulation results to gain important insights into the developed strategies.

The rest of the paper is organized as follows. Section II analyzes game theoretic Trojan insertion-testing strategies under considerations of testing costs and testing imperfections while considering the defender and the attacker to be fully rational. Section III analyzes Trojan insertion-testing strategies when the defender and the attacker exhibit irrationalities. Section IV presents simulation results to provide insights. Finally, Section V concludes the paper.

II. Game Theoretic Trojan Testing

Consider that there are N types of Trojans, viz. $\{1, \cdots, N\}$. Also, consider that there is a malicious IC manufacturer (referred to as the attacker A) who inserts a Trojan of type $i \in \{1, \cdots, N\}$ with a probability q_i into a manufactured IC, where $0 \leq \sum_{i=1}^{N} q_i \leq 1$. Further, consider that there is a buyer (referred to as the defender D) who acquires an IC from A and tests the acquired IC to check for the presence of Trojan type $i \in \{1, \cdots, N\}$ with a probability p_i, where $0 \leq \sum_{i=1}^{N} p_i \leq 1$. In this section, we consider that D and A are fully *rational* in nature. Suppose that the cost incurred

This work was supported in part by the NSF under Award Number CCF-2302197 and in part by the University of Cincinnati.

Defender \ Attacker	*Trojan not inserted*	*Insert Trojan type* 1	*Insert Trojan type* 2	*Insert Trojan type* 3
IC not tested	B^S, $-B^S$	$-V_1$, V_1	$-V_2$, V_2	$-V_3$, V_3
Test Trojan type 1	$B^S - c_1$, $c_1 - B^S$	$P_dF - (1-P_d)V_1 - c_1$, $-P_dF + (1-P_d)V_1 + c_1$	$-V_2 - c_1$, $V_2 + c_1$	$-V_3 - c_1$, $V_3 + c_1$
Test Trojan type 2	$B^S - c_2$, $c_2 - B^S$	$-V_1 - c_2$, $V_1 + c_2$	$P_dF - (1-P_d)V_2 - c_2$, $-P_dF + (1-P_d)V_2 + c_2$	$-V_3 - c_2$, $V_3 + c_2$
Test Trojan type 3	$B^S - c_3$, $c_3 - B^S$	$-V_1 - c_3$, $V_1 + c_3$	$-V_2 - c_3$, $V_2 + c_3$	$P_dF - (1-P_d)V_3 - c_3$, $-P_dF + (1-P_d)V_3 + c_3$

TABLE I
PAYOFF MATRIX FOR DEFENDER (D) AND ATTACKER (A) WITH THREE TROJAN TYPES ($N = 3$)

by the defender D to test the IC against Trojan type i is c_i. Moreover, to model the *imperfections* of the testing process, consider that, given that the attacker A has inserted a Trojan of type i and that the defender D has conducted a test to check for the presence of Trojan type i, the conducted test detects the inserted Trojan with a probability P_d.

Note that we consider that the attacker does not insert any Trojan with a probability $q_0 = 1 - \sum_{i=1}^{N} q_i$, in which case we say that the defender obtains a benefit B^S from putting the IC to desired use. Also, note that we allow the defender to not test the acquired IC with a probability $p_0 = 1 - \sum_{i=1}^{N} p_i$. In such a scenario, given that the attacker A has inserted a Trojan into the sold IC, if the defender D chooses not to test the IC, or 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 conducted test fails to detect it, then the inserted Trojan remains undetected and we consider that an undetected Trojan of type $i \in \{1, \cdots, N\}$ causes D to incur damage V_i (while providing a benefit V_i to A). However, if the defender D is able to successfully detect the presence of the Trojan that was inserted by the attacker, then we consider D to impose a fine F on the malicious manufacturer (with D refraining from using the acquired IC in such a case). Note that, we consider the testing costs incurred by the defender to positively impact the attacker's utility, which reflects the 'satisfaction' that the attacker obtains from making the defender incur costs for defending against attacks. For illustration, the payoff matrix of the game for $N = 3$ is shown in Table I.

First, we investigate the *pure strategy* NE of the aforementioned game. As can be seen from Table I, the strategy of the attacker not inserting any Trojan is a strictly dominated strategy (i.e., we have $\sum_{i=1}^{N} q_i = 1$ at NE). The existence and characteristics of pure strategy NE of the game depends on P_d. If there exists $i \in \{1, \cdots, N\}$ such that $V_i = \max_{k \in \{1,\cdots,N\}} V_k$ and $P_d \le \frac{c_i}{F+V_i}$, the game permits pure strategy Nash equilibria which corresponds to the attacker A inserting any such Trojan type i (which would be a maximally damaging Trojan type) with the defender choosing not to test the IC. The equilibrium follows from the fact that, given that A inserts a Trojan type i that is maximally damaging in nature, the payoff that the defender obtains from choosing not to test the IC is at least as much as that of the other strategies (under $P_d \le \frac{c_i}{F+V_i}$). Again, given that D chooses not to test the IC, the best response of the attacker is clearly to insert a Trojan type i that is maximally damaging in nature.

Now, when $P_d > \frac{c_i}{F+V_i}$ $\forall i \in \{1, \cdots, N\}$, in which case the strategy of the defender choosing not to test the IC becomes strictly dominated, a pure strategy NE exists only if there exists $i \in \{1, \cdots, N\}$ such that $V_i = \max_{k \in \{1,\cdots,N\}} V_k$ and $P_d \le \frac{V_i - V_j}{V_i + F}$ where $V_j = \max_{k \in \{1,\cdots,i-1,i+1,\cdots,N\}} V_k$, in which case, the game's pure strategy Nash equilibria correspond to both the defender and the attacker choosing to test and insert, respectively, such a Trojan type i (which would be a maximally damaging Trojan type). It can be verified that there does not exist any profitable unilateral deviation for D and A from such a strategy under the aforementioned conditions. However, when $P_d > \frac{c_i}{F+V_i}$ $\forall i \in \{1, \cdots, N\}$, if there does not exist such a maximally damaging Trojan type i as just described, the game does not have any pure strategy NE. Next, we characterize the *mixed strategy* NE in such a scenario.

In the rest of the paper, suppose that, w.l.o.g, $V_1 \ge V_2 \ge \cdots \ge V_N$. Also, suppose that there are M maximally damaging Trojan types, i.e., $M = |\{i | V_i = \max_{j \in \{1,\cdots,N\}} V_j\}|$.

A. Presence of multiple maximally damaging Trojan types

We first consider the scenario where M out of the N Trojan types are maximally damaging in nature, i.e., $V_1 = V_2 = \cdots = V_M > V_{M+1} \ge V_{M+2} \ge \cdots \ge V_N$ under $M > 1$. In such a scenario, it can be noted that the strategy of the attacker inserting Trojan type $i \in \{M+1, \cdots, N\}$ becomes strictly dominated[2]. Subsequently, the strategy of the defender testing the IC against Trojan type $i \in \{M+1, \cdots, N\}$ also becomes strictly dominated. For notational simplicity, let us define $V = V_1 = V_2 = \cdots = V_M$. In such a scenario, the mixed strategy NE of the game is provided in the next theorem.

THEOREM 1. *Given that, for $M > 1$, M out of the N Trojan types are maximally damaging in nature with $V = V_1 = V_2 = \cdots = V_M > V_{M+1} \ge V_{M+2} \ge \cdots \ge V_N$, at NE,*

- *the defender tests the IC for the presence of Trojan type i with a probability $p_i = \frac{1}{M}$, $\forall i \in \{1, \cdots, M\}$, and*
- *the attacker, for any chosen $i \in \{1, \cdots, M\}$, inserts Trojan type i into the manufactured IC with a prob. $q_i = \frac{1 - \sum_{k=1,k\neq i}^{M} \frac{c_k - c_i}{P_d(F+V)}}{M}$ and inserts Trojan type j with a prob. $q_j = q_i + \left[\frac{c_j - c_i}{P_d(F+V)}\right], \forall j \in \{1, \cdots, M\}, j \neq i$.*

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

$$E_D^i = [P_dF - (1-P_d)V]q_i - V \sum_{k=1,k\,\neq\,i}^{M} q_k - c_i \qquad (1)$$

At the mixed strategy NE, since D must be indifferent over its undominated strategy space, for $i, j \in \{1, \cdots, M\}, i \neq j$, equating $E_D^i = E_D^j$, after some manipulations yield

$$q_j = q_i + \left[\frac{c_j - c_i}{P_d(F+V)}\right] \qquad (2)$$

[2] For example, from the payoff matrix shown in Table I for $N = 3$, if $V_1 = V_2 > V_3$, it can be noted that strategy of the attacker inserting Trojan type 3 becomes strictly dominated (regardless of the strategy adopted by D).

Further, over the undominated strategy space of the attacker, for any chosen $i \in \{1, \cdots, M\}$, we have $q_i + \sum_{k=1,k\neq i}^{M} q_k = 1$, which implies, using (2), that

$$q_i + \sum_{k=1,k\neq i}^{M} \left[q_i + \left(\frac{c_k - c_i}{P_d(F+V)} \right) \right] = 1 \tag{3}$$

$$\Rightarrow q_i = \frac{1 - \sum_{k=1,k\neq i}^{M} \frac{c_k - c_i}{P_d(F+V)}}{M} \tag{4}$$

Clearly, from the above, if the attacker, for any chosen $i \in \{1, \cdots, M\}$, chooses q_i as given in (4) and q_j, $\forall j \in \{1, \cdots, M\}, j \neq i$, as given in (2), the defender becomes indifferent over its undominated strategy space making any strategy of defender (such that $\sum_{i=1}^{M} p_i = 1$) to become a best response against the attacker's strategy.

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

$$E_A^i = [(1 - P_d)V - P_d F + c_i]p_i + \sum_{k=1,k\,\neq\,i}^{M} (V + c_k)p_k \tag{5}$$

At the mixed strategy NE, since the attacker must also become indifferent over its undominated strategy space, for $i, j \in \{1, \cdots, M\}, i \neq j$, equating $E_A^i = E_A^j$, after some manipulations yield $p_i = p_j$. Further, over the undominated strategy space of the defender, for any $i \in \{1, \cdots, M\}$, we have $p_i + \sum_{k=1,k\,\neq\,i}^{M} p_k = 1$, which implies, using the fact that, for $i \neq j$, $E_A^i = E_A^j$ when $p_i = p_j$, that

$$p_i = \frac{1}{M} \quad \forall i \in \{1, \cdots, M\} \tag{6}$$

Clearly, $\forall i \in \{1, \cdots, M\}$, if the defender chooses p_i as given in (6), the attacker would become indifferent over its undominated strategy space making any strategy of the attacker (such that $\sum_{i=1}^{M} q_i = 1$) to become a best response against the defender's strategy.

Thus, if the attacker A, for any chosen $i \in \{1, \cdots, M\}$, chooses q_i as given in (4) and q_j, $\forall j \in \{1, \cdots, M\}$, $j \neq i$, as given in (2), and if the defender D, $\forall i \in \{1, \cdots, M\}$, chooses p_i as given in (6), both D and A would be playing their best responses against each other. This proves the theorem. □

B. Presence of a unique maximally damaging Trojan type

We now consider $M = 1$, i.e., there exists a unique maximally damaging Trojan type with L Trojan types having the *second-highest* damaging factor. Thus, we have $V_1 > V_2 = V_3 = \cdots = V_{L+1} > V_{L+2} \geq V_{L+3} \geq \cdots \geq V_N$. In such a case, the strategy of the attacker and the defender inserting and testing, respectively, a Trojan type $i \in \{L+2, \cdots, N\}$ become strictly dominated. For notational simplicity, define $V = V_2 = \cdots = V_{L+1}$. In the following theorem, we characterize the mixed strategy NE of the defender and the attacker over their undominated strategy spaces.

THEOREM 2. *Given that $M = 1$, i.e., there exists a unique maximally damaging Trojan type, and that L Trojan types have the second-highest damaging factor, with $V = V_2 = \cdots = V_{L+1}$, at NE,*

- *the defender chooses $p_1 = \frac{1-L\left[\frac{V-V_1}{P_d(F+V)}\right]}{1+L\left(\frac{F+V_1}{F+V}\right)}$ and $p_i = p_1\left(\frac{V_1+F}{V+F}\right) + \frac{V-V_1}{P_d(F+V)}$, $\forall i \in \{2, 3, \cdots, L+1\}$, and*
- *the attacker chooses $q_1 = \frac{1-\frac{1}{P_d(F+V)}\sum_{k=2}^{L+1}(c_k-c_1)}{1+L\left(\frac{F+V_1}{F+V}\right)}$ and $q_i = q_1\frac{(F+V_1)}{(F+V)} + \frac{c_i-c_1}{P_d(F+V)}$, $\forall i \in \{2, 3, \cdots, L+1\}$.*

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

$$E_D^1 = [P_d F - (1 - P_d)V_1]q_1 - V \sum_{k=2}^{L+1} q_k - c_1 \tag{7}$$

Further, the expected utility (say, E_D^i) of D from testing the IC against Trojan type $i \in \{2, 3, \cdots, L+1\}$ is

$$E_D^i = [P_d F - (1 - P_d)V]q_i - V \sum_{k=2,k\neq i}^{L+1} q_k - V_1 q_1 - c_i \tag{8}$$

Equating (7) and (8) to make the defender indifferent between testing against Trojan type 1 and type $i \in \{2, 3, \cdots, L+1\}$, as required at the mixed strategy NE, we get

$$q_i = q_1 \frac{(F + V_1)}{(F + V)} + \frac{c_i - c_1}{P_d(F + V)} \tag{9}$$

Now, since the strategy of the attacker inserting Trojan type $i \in \{1, 2, \cdots, L+1\}$ remains undominated, we have $q_1 + \sum_{k=2}^{L+1} q_k = 1$, which implies, using (9), that

$$q_1 + \sum_{k=2}^{L+1} \left[q_1 \frac{(F + V_1)}{(F + V)} + \frac{c_k - c_1}{P_d(F + V)} \right] = 1 \tag{10}$$

$$\Rightarrow q_1 = \frac{1 - \frac{1}{P_d(F+V)} \sum_{k=2}^{L+1} (c_k - c_1)}{1 + L\left(\frac{F+V_1}{F+V}\right)} \tag{11}$$

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

$$E_A^1 = [(1 - P_d)V_1 - P_d F + c_1]p_1 + \sum_{k=2}^{L+1} (V_1 + c_k)p_k \tag{12}$$

Further, the expected utility of A from inserting Trojan type $i \in \{2, 3, \cdots, L+1\}$ is

$$E_A^i = [(1 - P_d)V - P_d F + c_i]p_i + \sum_{k=1,k\neq i}^{L+1} (V + c_k)p_k \tag{13}$$

Equation (12) and (13) to make the attacker indifferent between inserting Trojan type 1 and type $i \in \{2, 3, \cdots, L+1\}$, as required at the mixed strategy NE, we get

$$p_i = p_1 \left(\frac{V_1 + F}{V + F} \right) + \frac{V - V_1}{P_d(F + V)} \tag{14}$$

Now, since the strategy of the defender testing Trojan type $i \in \{1, 2, \cdots, L+1\}$ remains undominated, we have $p_1 + \sum_{i=2}^{L+1} p_i = 1$, which implies, using (14), that

$$p_1 + \sum_{i=2}^{L+1} \left[p_1 \left(\frac{V_1 + F}{V + F} \right) + \frac{V - V_1}{P_d(F + V)} \right] = 1 \tag{15}$$

$$\Rightarrow p_1 = \frac{1 - L\left[\frac{V - V_1}{P_d(F+V)}\right]}{1 + L\left(\frac{F+V_1}{F+V}\right)} \tag{16}$$

In summary, if the attacker A chooses q_1 as given in (11), and $q_i, \forall i \in \{2, \cdots, L+1\}$, as given in (9), and if the defender D chooses p_1 as given in (16), and $p_i, \forall i \in \{2, \cdots, L+1\}$, as given in (14), then both D and A would be playing their best responses against each other. This proves the theorem. □

III. Trojan Testing under Cognitive biases

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 thereby exhibiting behavioral irrationalities. 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 [13], which provides a descriptive model of human cognitive biases, before describing our model.

A. Prospect Theory

In prospect theory [13], humans, due to their *cognitive biases*, do not weight outcomes by their objective probabilities, but rather by transformed distorted probabilities in a subjective manner. 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 well accepted Prelec function [14], which is defined as

$$w(p) = exp(-(-\log p)^{\alpha}), \quad 0 < \alpha \leq 1 \tag{17}$$

where α is the parameter which models how a human subjectively distorts an objective probability. For illustration, Fig. 1 plots $w(p)$ against p for different values of α.

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.

B. Prospect Theoretic Trojan Testing

We consider a similar Trojan insertion-testing model as described in Section II, with the attacker and the defender, however, considered cognitively biased in nature who subjectively perceive objective probabilities (using (17)) to obtain prospect theoretic utilities from their chosen strategies. For illustration, in Table II, we show the prospect theoretic payoff matrix of the game for $N = 3$.

As can be noted from Table II, the strategy of the attacker not inserting a Trojan is a strictly dominated strategy. In such scenario, it can be shown that, if there exists $i \in \{1, \cdots, N\}$ such that $V_i = \max_{k \in \{1,\cdots,N\}} V_k$ and $w(P_d)F - w(1-P_d)V_i \leq c_i - V_i$, then the strategy of the attacker inserting such a Trojan type i (which would be a maximally damaging Trojan type) with the defender choosing not to test the IC comprise a pure strategy NE. However, if $w(P_d)F - w(1-P_d)V_i > c_i - V_i \; \forall i \in \{1, \cdots, N\}$, a pure strategy NE exists only if there exists $i \in \{1, \cdots, N\}$ such that $V_i = \max_{k \in \{1,\cdots,N\}} V_k$ and $w(1-P_d)V_i - w(P_d)F \geq V_j$ where $V_j = \max_{k \in \{1,\cdots,i-1,i+1,\cdots,N\}} V_k$, in which case pure strategy Nash equilibria of the game corresponds to both the defender and the attacker testing and inserting, respectively, such a Trojan type i (which would be a maximally damaging Trojan type). In case there does not exist such a Trojan type i as just described when $w(P_d)F - w(1-P_d)V_i > c_i - V_i$ $\forall i \in \{1, \cdots, N\}$, then the game does not have a pure strategy NE. Next, we explore the mixed strategy NE in such a scenario considering that M out of the N Trojan types are maximally damaging in nature, i.e., $M = |\{i | V_i = \max_{j \in \{1,\cdots,N\}} V_j\}|$.

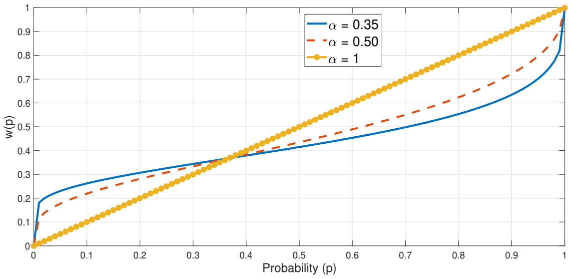

Fig. 1. Behavior of Prelec function

C. Presence of multiple maximally damaging Trojan types

We first consider $M > 1$, i.e., $V = V_1 = V_2 = \cdots = V_M > V_{M+1} \geq V_{M+2} \geq \cdots \geq V_N$. In such a scenario, the strategy of the defender and the attacker testing and inserting, respectively, a Trojan type $i \in \{M+1, \cdots, N\}$ become strictly dominated. In the next theorem, we characterize the mixed strategy NE of the game over the defender's and the attacker's undominated strategy spaces.

Theorem 3. *When the defender and the attacker are cognitively biased, for $M > 1$, if M out of the N Trojan types are maximally damaging in nature with $V = V_1 = V_2 = \cdots = V_M > V_{M+1} \geq V_{M+2} \geq \cdots \geq V_N$,*

- *the defender's strategy $(p_1, \cdots, p_M)$ at NE corresponds to $p_i = \frac{1}{M}$, $\forall i \in \{1, \cdots, M\}$, and*
- *the attacker's strategy $(q_1, \cdots, q_M)$ at NE corresponds to, for any chosen $i \in \{1, \cdots, M\}$, the roots of the following M equations solved simultaneously:*

$$F[w(P_d q_i) - w(P_d q_j)] + V[w((1-P_d)q_j) - w((1-P_d) q_i)] + V[w(q_i) - w(q_j)] + c_j - c_i = 0 \tag{18a}$$
$$\forall j \in \{1, \cdots, i-1, i+1, \cdots, M\}$$

$$q_i + \sum_{j=1, j \neq i}^{M} q_j = 1 \tag{18b}$$

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

$$PT_D^i = [Fw(P_d q_i) - Vw((1-P_d)q_i)] - V \sum_{k=1, k \neq i}^{M} w(q_k) - c_i \tag{19}$$

At the mixed strategy NE, since the defender must become indifferent between testing Trojan types i and j, $i, j \in \{1, \cdots, M\}$, $i \neq j$, equating $PT_D^i = PT_D^j$ yields (18a), which, for any chosen $i \in \{1, \cdots, M\}$ 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 (18b) to ensure feasibility of the attacker's strategy.

Now, the prospect theoretic utility (say, PT_A^i) of attacker A from choosing to insert Trojan type $i \in \{1, \cdots, M\}$ is

$$PT_A^i = Vw((1-P_d)p_i) - Fw(P_d p_i) + c_i w(p_i) + \sum_{k=1, k \neq i}^{M} (V + c_k) w(p_k) \tag{20}$$

Defender \ Attacker	*Trojan not inserted*	*Insert Trojan type* 1	*Insert Trojan type* 2	*Insert Trojan type* 3
IC not tested	$B^S, -B^S$	$-V_1, V_1$	$-V_2, V_2$	$-V_3, V_3$
Test Trojan type 1	$B^S - c_1, c_1 - B^S$	$w(P_d)F - w(1-P_d)V_1 - c_1,$ $-w(P_d)F + w(1-P_d)V_1 + c_1$	$-V_2 - c_1, V_2 + c_1$	$-V_3 - c_1, V_3 + c_1$
Test Trojan type 2	$B^S - c_2, c_2 - B^S$	$-V_1 - c_2, V_1 + c_2$	$w(P_d)F - w(1-P_d)V_2 - c_2,$ $-w(P_d)F + w(1-P_d)V_2 + c_2$	$-V_3 - c_2, V_3 + c_2$
Test Trojan type 3	$B^S - c_3, c_3 - B^S$	$-V_1 - c_3, V_1 + c_3$	$-V_2 - c_3, V_2 + c_3$	$w(P_d)F - w(1-P_d)V_3 - c_3,$ $-w(P_d)F + w(1-P_d)V_3 + c_3$

TABLE II
PROSPECT THEORETIC PAYOFF MATRIX FOR DEFENDER (D) AND ATTACKER (A) WITH THREE TROJAN TYPES ($N = 3$)

At the mixed strategy NE, since the attacker must become indifferent between inserting Trojan types i and j, $i, j \in \{1, \cdots, M\}$, $i \neq j$, equating $PT_A^i = PT_A^j$ yields

$$F[w(P_d p_j) - w(P_d p_i)] + V[w((1-P_d)p_i) - w((1-P_d)p_j)] + V[w(p_j) - w(p_i)] = 0 \tag{21}$$

which holds when $p_i = p_j$. Thus, to make the attacker indifferent over its entire undominated strategy space, we must have $p_1 = p_2 = \cdots = p_M$, which implies, since $p_i + \sum_{j=1, j\neq i}^{M} p_j = 1$, that $p_i = 1/M$, $\forall i \in \{1, \cdots, M\}$.

Thus, if the defender chooses $p_i = \frac{1}{M}$, $\forall i \in \{1, \cdots, M\}$, and if the attacker, for any chosen $i \in \{1, \cdots, M\}$, adopts its strategy by solving (18a) and (18b) simultaneously, both the defender and the attacker would be playing their best responses against each other. This proves the theorem. □

It can be noted that Matlab's `fzero` tookit [15] can be used to simultaneous solve (18a) and (18b) in a computationally efficient manner. In the following remark, we provide a closed form solution for (18a) and (18b) for a special case.

REMARK 1. *Note that, if $c_i = c_j$, $i, j \in \{1, \cdots, M\}$, $i \neq j$, (18a) holds when $q_i = q_j$. Thus, under uniform testing costs, i.e., when $c_1 = c_2 = \cdots = c_M$, the attacker's strategy at NE when $M > 1$ (i.e., under the case considered in Theorem 3) becomes $q_1 = q_2 = \cdots = q_M$, which implies, to satisfy (18b), that $q_i = \frac{1}{M}$, $\forall i \in \{1, \cdots, M\}$.*

D. Presence of a unique maximally damaging Trojan type

We now consider $M = 1$, i.e., there exists a unique maximally damaging Trojan type. For analytical tractability, we consider that a single Trojan type has the second-highest damaging factor, i.e., $V_1 > V_2 > V_3 \geq V_4 \geq \cdots \geq V_N$. In such a case, it can be shown that the strategy of the defender and attacker testing and inserting, respectively, a Trojan type $i \in \{3, \cdots, N\}$ become strictly dominated (i.e., we have $p_1 + p_2 = 1$ and $q_1 + q_2 = 1$ at NE). Next, we characterize the mixed strategy NE of the defender and the attacker over their undominated strategy spaces.

THEOREM 4. *When the defender and the attacker are cognitively biased, given that the Trojan types that have the highest and the second-highest damaging factors are unique, which would be Trojan types* 1 *and* 2*, respectively, at NE,*

- *the defender tests against Trojan type* 1 *with a probability p_1 (with $p_2 = 1 - p_1$) that corresponds to the root of the following equation:*

$$F[w(P_d(1-p_1)) - w(P_d p_1)] + V_1[w((1-P_d)p_1) + w(1-p_1)] - V_2[w((1-P_d)(1-p_1)) + w(p_1)] = 0 \tag{22}$$

- *the attacker inserts Trojan type 1 with a probability q_1 (with $q_2 = 1 - q_1$) that corresponds to the root of the following equation:*

$$F[w(P_d q_1) - w(P_d(1-q_1))] + V_1[w(q_1) - w((1-P_d)q_1)] + V_2[w((1-P_d)(1-q_1)) - w(1-q_1)] + c_2 - c_1 = 0 \tag{23}$$

Proof. For $i, j \in \{1, 2\}$, $i \neq j$, the prospect theoretic utility (say, PT_D^i) of defender D from testing the IC for the presence of Trojan type i is

$$PT_D^i = [Fw(P_d q_i) - V_i w((1-P_d)q_i)] - V_j w(q_j) - c_i \tag{24}$$

Since D must be indifferent between testing against Trojan types 1 and 2 at the mixed strategy NE, equating $PT_D^1 = PT_D^2$, and simplifying it using $q_1 + q_2 = 1$, we get (23).

Now, for $i, j \in \{1, 2\}$, $i \neq j$, the prospect theoretic utility (PT_A^i) of attacker A from choosing to insert Trojan type i is

$$PT_A^i = V_i w((1-P_d)p_i) - Fw(P_d p_i) + c_i w(p_i) + (V_i + c_j)w(p_j) \tag{25}$$

Since A must be indifferent between inserting Trojan types 1 and 2 at the mixed strategy NE, equating $PT_A^1 = PT_A^2$, and simplifying it using $p_1 + p_2 = 1$, we get (22).

Thus, if the attacker chooses q_1 such that it solves (23), with $q_2 = 1 - q_1$, and if the defender chooses p_1 such that it solves (22), with $p_2 = 1 - p_1$, both would be playing their best responses against each other. This proves the theorem. □

Theorem 4 uses (22) and (23) to characterize the strategies at NE. We now prove that the NE strategies exist by showing that the equations have solutions in [0,1].

LEMMA 1. *The NE strategies characterized in Theorem 4 exist.*

Proof. Let us denote (23) as

$$f(q_1) = F[w(P_d q_1) - w(P_d(1-q_1))] + V_1[w(q_1) - w((1-P_d)q_1)] + V_2[w((1-P_d)(1-q_1)) - w(1-q_1)] + c_2 - c_1 = 0 \tag{26}$$

It can be shown from Table II that, for $i, j \in \{1, 2\}$, $i \neq j$, if $c_i - c_j > Fw(P_d) + V_i[1 - w(1 - P_d)]$, then the strategy of the defender testing against Trojan type i becomes strictly dominated and the defender no longer plays a mixed strategy at NE. Hence, we explore the existence of $q_1 \in [0, 1]$ that solves (26) when the aforementioned condition is not satisfied. In such a case, it can be shown that $df(q_1)/dq_1 \geq 0$, which implies that $f(q_1)$ is a monotonically increasing function of q_1. Further, we have $\lim_{q_1 \to 0} f(q_1) < 0$ and $\lim_{q_1 \to 1} f(q_1) > 0$. Thus, we can conclude that there exists a value of $q_1 \in [0, 1]$ at which $f(q_1) = 0$. This proves the lemma. □

In Fig. 2, we show the nature of $f(q_1)$ (26) w.r.t. q_1 considering c_1 = 40, c_2 = 30, F = 150, P_d = 0.5, α = 0.3, V_1 = 50, and V_2 = 20. The figure verifies the aforementioned nature of $f(q_1)$ (26) w.r.t. q_1 and that there exists $q_1 \in$ [0, 1]

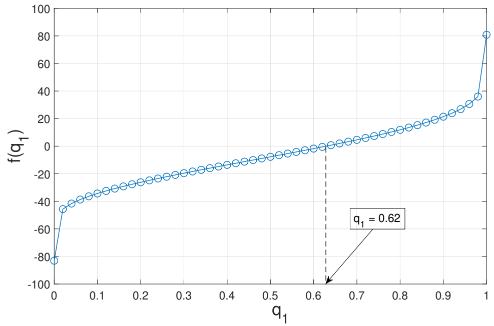

Fig. 2. Nature of $f(q_1)$ in (26).

such that $f(q_1) = 0$. This corroborates Lemma 1. Similarly, a solution of (22) can be shown to exist.

IV. Simulation Results

In this section, we provide simulation results to provide important insights into our developed techniques. In Fig. 3, we show the expected utilities of the defender and attacker at NE versus the probability of detection (P_d). For the figure, we consider the case studied in Section II-A (where $M > 1$) with $c_1 = c_2 = c_3 = 10$, $V = V_1 = V_2 = V_3 = 50$, and $F = 300$. The NE strategies of the defender and the attacker were calculated using Theorem 1. As can be seen from the figure, as expected, for any given M, the expected utility of the defender increases and that of the attacker decreases with P_d. Moreover, as can be seen, a lower value of M positively impacts the defender's utility (and negatively impacts the attacker's utility), since a lower value of M reduces the degree of uncertainty that the defender has regarding the inserted Trojan type.

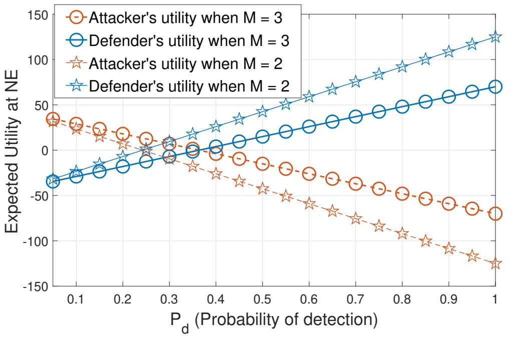

Fig. 3. Expected utilities of the defender and attacker at NE versus P_d

In Fig. 4, we show the defender's prospect theoretic utility at NE against P_d for the scenario explored in Section III-D. For the figure, we consider that $c_1 = 60$, $c_2 = 50$, $V_1 = 80$, $V_2 = 60$, and $F = 500$. The NE strategies of the defender and the attacker were calculated using Theorem 4. As can be seen from the figure, as expected, for any given probability distortion parameter α in (17), the prospect theoretic utility of the defender shows a non-decreasing trend with P_d. Further, interestingly, as can be seen, when P_d is below a certain threshold, a lower value of α, which corresponds to being more cognitively biased (i.e., being more irrational), results in higher utilities for the defender (i.e., is beneficial for the defender), while the trend reverses when P_d exceeds the threshold.

V. Conclusion

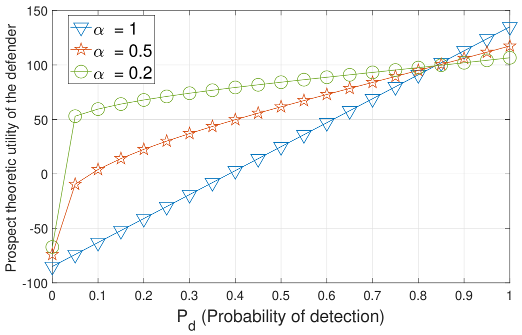

Fig. 4. Defender's prospect theoretic utility versus P_d.

This paper considered the problem of hardware Trojan testing under considerations of imperfections in the testing process and testing costs incurred. The paper first addressed the problem from a game theoretic perspective and analytically characterized NE-based Trojan insertion-testing strategies considering the defender (buyer of an IC) and the attacker (manufacturer of the IC) to be rational entities. Further, the paper employed Prospect Theory to model cognitive biases of the defender and attacker and analytically characterized NE-based Trojan insertion-testing strategies under the resulting behavioral irrationalities. Our results show that such irrationalities can lead to enhancement of the defender's utility depending on the degree of testing imperfections. Numerous simulation results were presented to gain important insights.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

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

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