Game Theoretic Hardware Trojan Testing under Cost Considerations

Swastik Brahma[1], Laurent Njilla[2], and Satyaki Nan[1]

[1]Department of Computer Science, Tennessee State University, Nashville, TN, USA
{sbrahma,snan}@tnstate.edu
[2]Cyber Assurance Branch, Air Force Research Laboratory, Rome, NY, USA
laurent.njilla@us.af.mil

Abstract. In this paper, we consider the problem of testing integrated circuits (ICs) to check for the presence of hardware Trojans from a game theoretic perspective. Under consideration of complex cost structures involved in the testing process, the paper analytically characterizes the Nash Equilibrium (NE) strategy of a malicious manufacturer for inserting a hardware Trojan into a manufactured IC and that of a defender for testing the acquired IC to check for the presence of Trojans. The paper first considers the defender, who incurs testing costs, to be capable of testing one Trojan type and analytically characterizes the NE of such a scenario. The paper also considers the scenario where the defender can test an IC to check for the presence of multiple types of Trojans under a cost budget constraint and analytically characterizes the NE of such a game. Numerous numerical results are presented in the paper that provide important insights into the game theoretic strategies presented.

Keywords: Game Theory · Hardware Trojans · Security.

1 Introduction

A hardware Trojan is a malicious alteration of the circuitry of an integrated circuit (IC) [7]. The presence of hardware Trojans in ICs can lead to disastrous consequences [7,12,15], including leakage of confidential information from a system, derangement of system operation, and even complete system failure. For example, the failure of a Syrian radar system to warn about an incoming assault has been largely attributed to the presence of malicious circuitry in the system's components [1]. Such attacks have become a serious threat to the semiconductor industry and to modern cyber systems with the outsourcing trends of manufacturing processes in today's economy exacerbating integrity concerns regarding manufactured ICs.

The primary technique that past work [2–5, 7, 8, 16–18] has focused on for mitigating threats from hardware Trojans is the development of testing strategies that can check for the presence of Trojans in acquired ICs. For example, in [2], the authors have used random sequences of test patterns that can generate

This work was supported in part by the NSF under Award Number HRD 1912414 and in part by the Air Force under PIA FA8750-19-3-1000.

DISTRIBUTION A. Approved for public release. Distribution unlimited. Case Number AFRL-2021-3034. Dated 08 Sep 2021.

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, behavior and the size of the inserted Trojans. In [3], the authors propose a method that seeks to detect and estimate the locations of hardware Trojans in ICs using region-based partitioning. Again, in [8], the authors propose a technique, referred to as MERO (Multiple Excitation of Rare Occurence), that maximizes the probability of detecting an inserted Trojan using statistical methods. Since exhaustive testing of all possible Trojan types can be prohibitive, the works in [6,10,11,13,14] develop game theoretic [9] hardware Trojan testing strategies that can intelligently determine which Trojan types should an IC be tested for against a strategic malicious manufacturer. Specifically, the work in [10] presents a two-person Trojan detection game, but limits investigation of the equilibrium to an example scenario of the model. The works in [11,13,14] limit themselves to the use of software-based techniques for analyzing game theoretic testing strategies. In [6], the authors characterize equilibrium strategies for performing testing while, however, ignoring the costs incurred in the testing process.

In contrast to the aforementioned works on developing testing strategies using game theory, in this paper, we investigate game theoretic hardware Trojan testing under consideration of the costs incurred in the testing process and analytically characterize the Nash Equilibrium (NE) strategies as closed-form expressions. It should be noted that, to the best of our knowledge, analytical characterization of NE strategies in closed-forms under testing cost considerations remains an unsolved problem in past work. Specifically, the main contributions of the paper are as follows:

- We present game theoretic models that consider the costs incurred by a defender (i.e., the buyer of an IC) to perform testing and analytically characterize the NE strategies for Trojan insertion (from the perspective of a malicious manufacturer) and testing (from the perspective of the defender) in closed-forms.
- We first consider the scenario where the defender, who incurs costs for performing testing, is capable of testing the acquired IC for one Trojan type and analytically characterize the NE of such a game, which provide important insights into the impact of testing costs on the equilibrium solution.
- We also consider the general scenario where the defender can choose to test the acquired IC against multiple Trojan types under a cost budget constraint and analytically characterize the NE of such a Trojan insertion-testing game under consideration of the availability of various amounts of the defender's cost budget.
- Numerous numerical results are presented to gain important insights into the game theoretic strategies presented in the paper.

The rest of the paper is organized as follows. Section 2 presents our game theoretic model and results where the defender is capable of testing one Trojan type under consideration of the costs incurred for performing testing. Section 3

presents our game theoretic model and results where the defender can select multiple Trojan types for testing under a cost budget constraint. Section 4 presents numerical results that provide important insights into the game theoretic strategies presented. Finally, Section 5 concludes the paper.

2 Game Theoretic Trojan Testing Under Cost Considerations

In this section, we consider the problem of performing game theoretic hardware Trojan testing where a defender D (who corresponds to the buyer of an IC) can test the acquired IC to check for the presence of one Trojan type and a malicious manufacturer (referred to as the attacker (A)) can insert a single Trojan type into a manufactured IC. We investigate the game where the defender can choose to test an IC against multiple types of Trojans under a cost budget constraint in Section 3.

Consider that there are N types of Trojans, viz. $\{1, \cdots, N\}$. Also, consider that the attacker (A) chooses to insert Trojan type $i \in \{1, \cdots, N\}$ with a probability q_i into a manufactured IC (such that $0 \leq \sum_{i=1}^{N} q_i \leq 1$) and that the defender D tests the IC to check for the presence of Trojan type i with a probability p_i (such that $0 \leq \sum_{i=1}^{N} p_i \leq 1$). 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 the defender obtains a benefit B^S from putting the IC to desired use. Also, note that we allow the defender in our model to not test the acquired IC to check for the presence of any Trojan with a probability $p_0 = 1 - \sum_{i=1}^{N} p_i$. Further, we consider that the defender incurs a cost c_i to test the IC for the presence of Trojan type $i \in \{1, \cdots, N\}$ and that if the defender tests the IC against the inserted Trojan type, 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 type which was not inserted by the attacker, or chooses not to test the IC, the inserted Trojan (if the attacker chose to insert one) remains undetected and we consider that an undetected Trojan of type $i \in \{1, \cdots, N\}$ causes the defender to incur damage V_i (and provides a benefit V_i to the attacker). The strategic interactions between the defender and the attacker, in this paper, is modeled as a zero-sum game. Note that in our model we consider the testing costs incurred by the defender to positively impact the attacker's utility reflecting the 'satisfaction' the attacker derives from making the defender incur costs for defending against attacks.

For illustration, the payoff matrix of the game when $N = 2$ is shown in Table 1. As can be seen from the table, 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 NE of the game, as can be seen from the table, depends on the relationships among cost structures of the game. Specifically, the game can have pure strategy Nash equilibria which corresponds to the attacker inserting any Trojan type $i \in \{1, \cdots, N\}$ for which $V_i = \max_{j \in \{1, \cdots, N\}} V_j$ and $F \leq c_i - V_i$, and the defender choosing not to test the IC for the presence of any Trojan. It is easy to show that there does not exist any profitable unilateral deviation from such a

Defender\Attacker	*Don't insert Trojan*	*Insert Trojan type* 1	*Insert Trojan type* 2
Don't test IC	B^S, $-B^S$	$-V_1$, V_1	$-V_2$, V_2
Test Trojan type 1	$B^S - c_1$, $c_1 - B^S$	$F - c_1$, $c_1 - F$	$-V_2 - c_1$, $V_2 + c_1$
Test Trojan type 2	$B^S - c_2$, $c_2 - B^S$	$-V_1 - c_2$, $V_1 + c_2$	$F - c_2$, $c_2 - F$

Table 1. Payoff matrix of the game when $N = 2$.

strategy profile. However, with the attacker's strategy of not inserting a Trojan being strictly dominated (and therefore never adopted by the attacker) as noted above, if $F > c_i - V_i \ \forall i \in \{1, \cdots, N\}$, the defender's strategy of not testing the IC becomes strictly dominated and the game no longer has a pure strategy NE. We provide the mixed strategy NE in this scenario in the next theorem.

Theorem 1. *At NE,*

- *the defender, for any chosen $i \in \{1, \cdots, N\}$, tests the acquired IC to check for the presence of Trojan type i with a probability $p_i = \frac{1-\sum_{j=1, j \neq i}^{N} \frac{(V_j - V_i)}{F+V_j}}{1+\sum_{j=1, j \neq i}^{N} \frac{F+V_i}{F+V_j}}$ and tests the IC for the presence of Trojan type j with a probability $p_j = \frac{(V_j - V_i)}{F+V_j} + \frac{p_i(F+V_i)}{F+V_j}$, $\forall j \in \{1, \cdots, N\}, j \neq i$, and*
- *the attacker, for any chosen $i \in \{1, \cdots, N\}$, inserts Trojan type i into the manufactured IC with a probability $q_i = \frac{1-\sum_{j=1, j \neq i}^{N} \frac{c_i - c_j}{F+V_i}}{1+\sum_{j=1, j \neq i}^{N} \frac{F+V_j}{F+V_i}}$ and inserts Trojan type j with a probability $q_j = \frac{q_i(F+V_i)}{F+V_j} + \frac{c_j - c_i}{F+V_j}, \forall j \in \{1, \cdots, N\}, j \neq i$.*

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

$$E_D^i = (F - c_i)q_i + \sum_{j=1, j \neq i}^{N} (-V_j - c_i)q_j \tag{1}$$

At the mixed strategy NE, since the defender must become indifferent over its undominated strategy space, we must have $E_D^1 = E_D^2 = \cdots = E_D^N$. Now, for $i, j \in \{1, \cdots, N\}, i \neq j$, equating $E_D^i = E_D^j$, after some simplifications, we get

$$q_j = \frac{q_i(F + V_i)}{F + V_j} + \frac{c_j - c_i}{F + V_j} \tag{2}$$

Further, since not inserting a Trojan is a strictly dominated strategy for the attacker, in the attacker's adopted strategy, for any $i \in \{1, \cdots, N\}$, we have

$$q_i + \sum_{j=1, j \neq i}^{N} q_j = 1 \tag{3}$$

$$\Rightarrow q_i + \sum_{j=1, j \neq i}^{N} \frac{q_i(F + V_i)}{F + V_j} + \frac{c_j - c_i}{F + V_j} = 1 \quad \text{(using (2))}$$

$$\Rightarrow q_i = \frac{1 - \sum_{j=1, j \neq i}^{N} \frac{c_i - c_j}{F+V_i}}{1 + \sum_{j=1, j \neq i}^{N} \frac{F+V_j}{F+V_i}} \tag{4}$$

Clearly, from the above, if the attacker, for any chosen $i \in \{1, \cdots, N\}$, chooses q_i as given in (4) and q_j, $\forall j \in \{1, \cdots, N\}, j \neq i$, as given in (2), the defender becomes indifferent over its undominated strategy space making any strategy of defender (such that $\sum_{i=1}^{N} p_i = 1$) to become a best response against the attacker's strategy (as well as it is ensured that $\sum_{i=1}^{N} q_i = 1$, which is needed since the attacker's strategy of not inserting any Trojan is a strictly dominated strategy).

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

$$E_A^i = (c_i - F)p_i + \sum_{j=1, j \neq i}^{N} (V_i + c_j)p_j \tag{5}$$

At the mixed strategy NE, since the attacker must also become indifferent over its undominated strategy space, we must have $E_A^1 = E_A^2 = \cdots = E_A^N$. Now, for $i, j \in \{1, \cdots, N\}, i \neq j$, equating $E_A^i = E_A^j$, after some simplifications, we get

$$p_j = \frac{(V_j - V_i)}{F + V_j} + \frac{p_i(F + V_i)}{F + V_j} \tag{6}$$

Further, when $F > c_i - V_i \ \forall i \in \{1, \cdots, N\}$, since the defender's strategy of not testing the IC becomes strictly dominated, in the defender's adopted strategy, for any $i \in \{1, \cdots, N\}$, we have

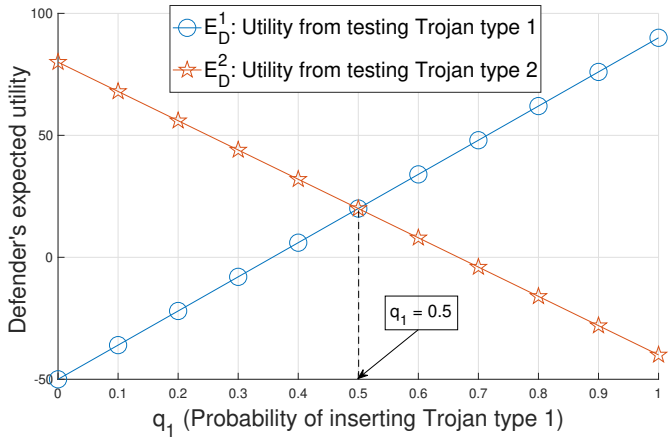
$$p_i + \sum_{j=1, j \neq i}^{N} p_j = 1 \tag{7}$$

$$\Rightarrow p_i + \sum_{j=1, j \neq i}^{N} \frac{(V_j - V_i)}{F + V_j} + \frac{p_i(F + V_i)}{F + V_j} = 1 \quad \text{(using (6))}$$

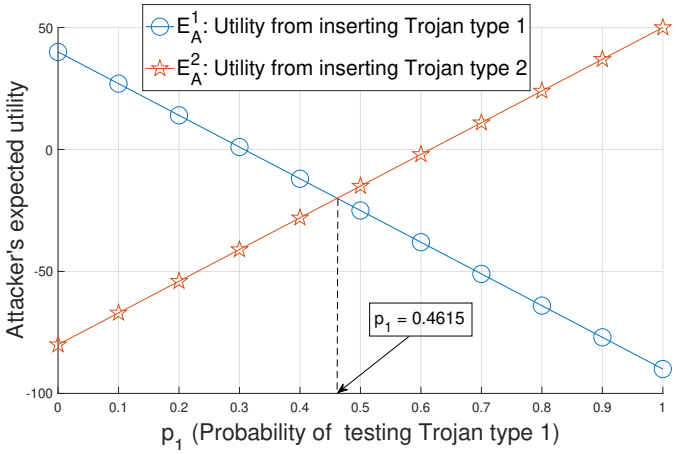
$$\Rightarrow p_i = \frac{1 - \sum_{j=1, j \neq i}^{N} \frac{(V_j - V_i)}{F + V_j}}{1 + \sum_{j=1, j \neq i}^{N} \frac{F + V_i}{F + V_j}} \tag{8}$$

Clearly, if the defender, for any chosen $i \in \{1, \cdots, N\}$, chooses p_i as given in (8) and p_j, $\forall j \in \{1, \cdots, N\}$, $j \neq i$ as given in (6), the attacker becomes indifferent over its undominated strategy space making any strategy of the attacker (such that $\sum_{i=1}^{N} q_i = 1$) to become a best response against the defender's strategy (as well as it is ensured that $\sum_{i=1}^{N} p_i = 1$, which is needed since the defender's strategy of not testing the IC is a strictly dominated strategy).

Thus, in summary, if the attacker, for any chosen $i \in \{1, \cdots, N\}$, chooses q_i as given in (4) and q_j, $\forall j \in \{1, \cdots, N\}$, $j \neq i$ as given in (2) and if the defender, for any chosen $i \in \{1, \cdots, N\}$, chooses p_i as given in (8) and p_j, $\forall j \in \{1, \cdots, N\}$, $j \neq i$ as given in (6), both the defender and the attacker would be playing their best responses against each other. This proves the theorem.



(a) Defender's expected utility versus the attacker's strategies.



(b) Attacker's expected utility versus the defender's strategies.

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

We now provide numerical results to corroborate Theorem 1. In Fig. 1, we show the expected utilities of the defender and the attacker versus their opponents' strategies. For the figure, we consider two Trojan types, viz. $\{1, 2\}$, with $V_1 = 20$, $V_2 = 40$, $F = 100$, $c_1 = 10$, and $c_2 = 20$. In Fig. 1(a), we show the defender's expected utility versus the probability (q_1) with which the attacker inserts Trojan type 1 into the manufactured IC (considering $q_2 = 1 - q_1$). Using (1), the blue line represents the defender's expected utility (E_D^1) from testing the acquired IC to check for the presence of Trojan type 1 and the red line represents the defender's expected utility (E_D^2) from testing the IC to check for the presence of Trojan type 2. The point where the two lines intersect makes the defender's expected utility obtained from testing the IC against Trojan type 1 to be equal to that obtained from testing against Trojan type 2 (as needed at the mixed strategy NE), which, as can be seen from the figure, occurs at $q_1 = 0.5$

(with $q_2 = 1 - 0.5 = 0.5$). It can be verified that the mixed strategy NE of the attacker obtained from Theorem 1 is also $q_1 = 0.5$ and $q_2 = 0.5$.

In Fig. 1(b), we show the attacker's expected utility versus the probability (p_1) with which the defender tests an acquired IC to check for the presence of Trojan type 1 (considering $p_2 = 1 - p_1$). Using (5), the blue line represents the attacker's expected utility (E_A^1) from inserting Trojan type 1 into the manufactured IC and the red line represents the attacker's expected utility (E_A^2) from inserting Trojan type 2. The point where the two lines intersect makes the attacker's expected utility obtained from inserting Trojan type 1 to be equal to that obtained from inserting Trojan type 2 (as needed at the mixed strategy NE), which, as can be seen from the figure, occurs at $p_1 = 0.4615$ (with $p_2 = 1 - 0.4615 = 0.5385$). It can be verified that the mixed strategy NE of the defender obtained from Theorem 1 is also $p_1 = 0.4615$ and $p_2 = 0.5385$. This corroborates Theorem 1.

3 Game Theoretic Trojan Testing under a Cost Budget Constraint

In this section, we consider that the defender can test for the presence of multiple types of Trojans under a cost budget constraint in a game theoretic context. Similar to Section 2, we consider that there are N types of Trojans, viz., $\{1, \cdots, N\}$, with the attacker's strategy denoted as $\mathbf{q} = (q_1, \cdots, q_N)$, where q_i is the probability of the attacker inserting Trojan type $i \in \{1, \cdots, N\}$ into the manufactured IC such that $0 \leq \sum_{i=1}^{N} q_i \leq 1$. We denote the defender's strategy as $\mathbf{p} = (p_1, \cdots, p_N)$, where p_i is the probability with which the defender tests the acquired IC to check for the presence of Trojan type $i \in \{1, \cdots, N\}$, and consider that the defender incurs a cost c_i for testing the IC against Trojan type i. In this section, we allow the defender to test the IC against multiple Trojan types without exceeding a cost budget C such that $\sum_{i=1}^{N} p_i c_i \leq C$. If the set of Trojan types tested by the defender contains the Trojan type inserted by the attacker, the inserted Trojan is considered to be detected and the malicious manufacturer in such a case is imposed a fine F. However, if the set of Trojan types tested by the defender does not contain the Trojan type inserted by the attacker, the inserted Trojan remains undetected and we consider that an undetected Trojan of type $i \in \{1, \cdots, N\}$ makes the defender incur damage V_i. In case the attacker does not insert any Trojan into the manufactured IC, the defender is considered to obtain a benefit B^S from putting the IC to desired use. The expected utility of the defender in such a game is

$$E_D(\mathbf{p}, \mathbf{q}) = B^S\big(1 - \sum_{i=1}^{N} q_i\big) + \sum_{i=1}^{N} \big[p_i q_i F - (1 - p_i) q_i V_i\big] \tag{9}$$

Denoting

$$\gamma_i(p_i) = p_i F - (1 - p_i) V_i - B^S \tag{10}$$

we can rewrite (9) as

$$E_D(\mathbf{p}, \mathbf{q}) = B^S + \sum_{i=1}^{N} \gamma_i(p_i) q_i \tag{11}$$

From a game theoretic perspective, the goal of the *defender* is to choose $\mathbf{p} = (p_1, \cdots, p_N)$ such that (11) is maximized (under consideration of the attacker optimizing against the defender's strategy) and that of the *attacker* is to choose $\mathbf{q} = (q_1, \cdots, q_N)$ such that (11) is minimized (under consideration of the defender optimizing against the attacker's strategy). The game is clearly a zero-sum game. Therefore, the NE of the game (which would coincide with its saddle point) corresponds to choosing $(\mathbf{p}, \mathbf{q})$ that solves the following optimization problem:

$$\max_{\mathbf{p}} \min_{\mathbf{q}} E_D(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{q}} \max_{\mathbf{p}} E_D(\mathbf{p}, \mathbf{q}) \tag{P1}$$

$$\text{subject to: } \sum_{i=1}^{N} p_i c_i \leq C$$

$$\sum_{i=1}^{N} q_i \leq 1$$

To characterize $(\mathbf{p}, \mathbf{q})$ in the above game, we first prove some properties of $\gamma_i(p_i)$ (10).

Lemma 1. *$\gamma_i(p_i)$ as defined in* (10) *is a strictly increasing function of p_i having the slope $F + V_i$, with $\gamma_i(p_i) = 0$ when $p_i = \frac{B^S + V_i}{F + V_i}$.*

Proof. Clearly, $\frac{d(\gamma_i(p_i))}{dp_i} = F + V_i > 0$. Again, equating $\gamma_i(p_i) = 0$ yields $p_i = \frac{B^S + V_i}{F + V_i}$.

In the following, we characterize $(\mathbf{p}, \mathbf{q})$ at NE in the game described above by considering three possible cases in terms of the available cost budget (C) of the defender, viz., $C > \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$, $C = \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$, and $C < \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$.

Now, note that, in the case where $C > \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$, there exists $p_i > \frac{B^S + V_i}{F + V_i}$ $\forall i \in \{1, \cdots, N\}$ such that $\sum_{i=1}^{N} p_i c_i \leq C$. Thus, the NE in this case corresponds to the defender testing the acquired IC to check for the presence of every Trojan type $i \in \{1, \cdots, N\}$ with a probability $p_i > \frac{B^S + V_i}{F + V_i}$ (while satisfying the cost budget constraint) and the attacker not inserting any Trojan into the manufactured IC (i.e., choosing $q_i = 0, \forall i \in \{1, \cdots, N\}$). Clearly, against the defender's strategy of choosing every $p_i > \frac{B^S + V_i}{F + V_i}$ (which makes every $\gamma_i(p_i) > 0$ following Lemma 1), the attacker's best response becomes choosing every $q_i = 0$ (since the attacker seeks to minimize (11)). Again, against the attacker's strategy of choosing every $q_i = 0$, clearly there does not exist any profitable unilateral deviation for the defender from its aforementioned strategy, which proves the above NE.

Next, we characterize the NE for the case where $C = \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$, in which case we say that the defender has a *sufficient cost budget*.

3.1 NE under Sufficient Cost Budget of the Defender

As mentioned above, we say that the defender has a sufficient cost budget when $C = \sum_{i=1}^{N} \frac{B^S+V_i}{F+V_i} c_i$. In the next lemma, we provide the property that characterizes the defender's strategy at NE in this case.

Lemma 2. *When $C = \sum_{i=1}^{N} \frac{B^S+V_i}{F+V_i} c_i$, the defender's strategy $\mathbf{p} = (p_1, \cdots, p_N)$ at NE is such that*

$$\gamma_i(p_i) = 0, \forall i \in \{1, \cdots, N\} \tag{12}$$

Proof. Consider a strategy profile $\mathbf{p} = (p_1, \cdots, p_N)$ of the defender such that $\sum_{i=1}^{N} p_i c_i \leq C$. Denote $\underline{g} = \min_{i \in \{1,\cdots,N\}} \gamma_i(p_i)$, $\overline{g} = \max_{i \in \{1,\cdots,N\}} \gamma_i(p_i)$, where $\gamma_i(p_i)$ is defined in (10), and suppose that $\underline{g} < \overline{g}$[4]. Moreover, define the set $\underline{G} = \{i | i \in \{1, \cdots, N\} \text{ and } \gamma_i(p_i) = \underline{g}\}$, the set $\overline{G} = \{i | i \in \{1, \cdots, N\} \text{ and } \gamma_i(p_i) = \overline{g}\}$, and the set $\underline{G}' = \{1, \cdots, N\} - \underline{G}$ (it can be noted that $|\underline{G}'| > 0$ necessarily holds when $\underline{g} < \overline{g}$). In such a scenario, to satisfy the cost budget constraint, it can be noted that we must have $\underline{g} < 0$. This is because, otherwise (if $\underline{g} \geq 0$), following Lemma 1, $\forall j \in \overline{G}$ we would have $p_j > \frac{B^S+V_j}{F+V_j}$ (i.e., $\gamma_j(p_j) > 0$) and $\forall i \in \{1, \cdots, N\} - \overline{G}$ we would have $p_i \geq \frac{B^S+V_i}{F+V_i}$ (i.e., $\gamma_i(p_i) \geq 0$), which would imply that $\sum_{i=1}^{N} p_i c_i > C$ (i.e., would violate the cost budget constraint).

Now, having noted that $\underline{g} < 0$, it should be further noted that, since the attacker aims to minimize (11), the best response of the attacker against the strategy $\mathbf{p}$ of the defender defined above is to adopt a strategy $\mathbf{q} = (q_1, \cdots, q_N)$ such that $\sum_{i \in \underline{G}} q_i = 1$. Consider now the following two possible cases.

- Case-I ($\overline{g} \leq 0$): In this case, following Lemma 1, $\forall i \in \underline{G}$ we have $p_i < \frac{B^S+V_i}{F+V_i}$ and $\forall j \in \underline{G}'$ we have $p_j \leq \frac{B^S+V_j}{F+V_j}$, which implies that $\sum_{i=1}^{N} p_i c_i < C$. Consider now $w \in \underline{G}$ for which $q_w > 0$ (as follows from the aforementioned attacker's best response $\mathbf{q}$, such a w is guaranteed to exist) and consider changing the strategy of the defender from $\mathbf{p} = (p_w, p_{-w})$ to $\mathbf{p}' = (p_w + \delta, p_{-w})$, where p_{-w} denotes the vector of probabilities used by the defender to test all Trojan types except Trojan type w and $\delta \in (0, \frac{B^S+V_w}{F+V_w} - p_w]$. Clearly, $(p_w + \delta)c_w + \sum_{i \in \{1,\cdots,N\}, i \neq w} p_i c_i \leq C$ (i.e., $\mathbf{p}'$ satisfies the cost budget constraint). Moreover, we have $\gamma_w(p_w + \delta) > \gamma_w(p_w)$ (which follows from Lemma 1) 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 profile $(\mathbf{p}, \mathbf{q})$ (where $\mathbf{q}$, as described earlier, forms a best response of the attacker against $\mathbf{p}$).
- Case-II ($\overline{g} > 0$): In this case, following Lemma 1, $\forall i \in \underline{G}$ we have $p_i < \frac{B^S+V_i}{F+V_i}$ and $\forall j \in \overline{G}$ we have $p_j > \frac{B^S+V_j}{F+V_j}$ (with $\sum_{i=1}^{N} p_i c_i \leq C$). Consider now $w \in \underline{G}$ for which $q_w > 0$ and any $z \in \overline{G}$ (note, as follows from the

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

aforementioned attacker's best response strategy $\mathbf{q}$ against $\mathbf{p}$, $q_z = 0, \forall z \in \overline{G}$) and consider changing the strategy of the defender from $\mathbf{p} = (p_w, p_z, p_{-wz})$ to $\mathbf{p}' = (p_w + \delta_w, p_z - \delta_z, p_{-wz})$, where p_{-wz} denotes the vector of probabilities with which the defender tests all Trojan types except Trojan types w and z, while ensuring[5] $\delta_w c_w \leq \delta_z c_z$ to have the strategy $\mathbf{p}'$ satisfy the cost budget constraint. Now, we have $\gamma_w(p_w + \delta_w) > \gamma_w(p_w)$ (which follows from Lemma 1) implying that $E_D(\mathbf{p}', \mathbf{q}) > E_D(\mathbf{p}, \mathbf{q})$ showing that there again exists a profitable unilateral deviation for the defender from the strategy profile $(\mathbf{p}, \mathbf{q})$ (where $\mathbf{q}$, as described earlier, forms a best response of the attacker against $\mathbf{p}$).

From the above, clearly there always exists a profitable unilateral deviation for the defender from a strategy profile $(\mathbf{p}, \mathbf{q})$ where $\mathbf{p}$ is such that $\underline{g} \neq \overline{g}$ and $\mathbf{q}$ forms a best response of the attacker against $\mathbf{p}$. Thus, at NE, we must have $\underline{g} = \overline{g}$, which implies that $\gamma_i(p_i) = \alpha$ at NE $\forall i \in \{1, \cdots, N\}$, where α is a constant.

We next prove that $\alpha = 0$ at NE. When $\alpha < 0$, the best response of the attacker becomes adopting a strategy $\mathbf{q} = (q_1, \cdots, q_N)$ such that $\sum_{i \in \{1, \cdots, N\}} q_i = 1$ (which would make the utility of the defender (11) to be $E_D(\mathbf{p}, \mathbf{q}) < B^S$). Against such a strategy of the attacker, using arguments similar to Case-I in this proof, it can be shown that there exists profitable unilateral deviations for the defender from any strategy $\mathbf{p}$ for which $\alpha < 0$. Again, $\alpha > 0$ would require $p_i > \frac{B^S + V_i}{F + V_i}, \forall i \in \{1, \cdots, N\}$, which would violate the cost budget constraint. From the above, clearly at NE we must have $\alpha = 0$, which proves the lemma.

Next, using Lemma 1 and Lemma 2, we characterize the NE for the sufficient cost budget case in Theorem 2.

Theorem 2. *When $C = \sum_{i=1}^{N} \frac{B^S + V_i}{F + V_i} c_i$, at NE, the defender's strategy corresponds to testing the acquired IC to check for the presence of every Trojan type $i \in \{1, \cdots, N\}$ with a probability $p_i = \frac{B^S + V_i}{F + V_i}$ and the attacker's strategy corresponds to, for any chosen $i \in \{1, \cdots, N\}$, inserting Trojan type i into the manufactured IC with a probability $q_i = \frac{k}{\frac{F + V_i}{c_i} \sum_{j=1}^{N} \frac{c_j}{F + V_j}}, k \in [0, 1]$, and inserting Trojan type j with a probability $q_j = q_i \frac{c_j}{c_i} \frac{F + V_i}{F + V_j}$, $\forall j \in \{1, \cdots, N\}$, $j \neq i$.*

Proof. Using Lemma 2, at NE, the defender's strategy $\mathbf{p} = (p_1, \cdots, p_N)$ must be such that $\gamma_i(p_i) = 0, \forall i \in \{1, \cdots, N\}$, implying that $p_i = \frac{B^S + V_i}{F + V_i}, \forall i \in \{1, \cdots, N\}$ at NE (using Lemma 1). Against such a strategy of the defender, it should be noted that any strategy $\mathbf{q} = (q_1, \cdots, q_N)$ (such that $0 \leq \sum_{i=1}^{N} q_i \leq 1$) forms a best response for the attacker. 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 $\mathbf{p}$ defined above. Consider now the deviation of the defender from the strategy $\mathbf{p} = (p_i, p_j, p_{-ij})$ at NE defined above

[5] Note that in the strategy $\mathbf{p}'$, $\delta_w c_w$ is the additional cost incurred by the defender due to the increase in the probability of testing Trojan type w and $\delta_z c_z$ is the decrease in cost incurred due to the decrease in the probability of testing Trojan type z.

to a strategy $\mathbf{p}' = (p_i + \delta_i, p_j - \delta_j, p_{-ij})$, where p_{-ij} denotes the vector of probabilities with which the defender tests all Trojan types except Trojan types i and j. To have $\mathbf{p}'$ satisfy the cost budget constraint, we must have[6]

$$\left(\frac{B^S + V_i}{F + V_i} + \delta_i\right) c_i + \left(\frac{B^S + V_j}{F + V_j} - \delta_j\right) c_j + \sum_{z=1, z\neq i, z\neq j}^{N} \frac{B^S + V_z}{F + V_z} c_z = C \quad (13)$$

which implies $\delta_i c_i - \delta_j c_j = 0$ in (13), which yields

$$\delta_j = \delta_i \frac{c_i}{c_j} \quad (14)$$

Now, recalling from Lemma 1 that $\frac{d(\gamma_x(p_x))}{dp_x} = F + V_x$ and from Lemma 2 that $\gamma_x(p_x) = 0\ \forall x \in \{1, \cdots, N\}$ in the strategy $\mathbf{p}$ of the defender at NE defined earlier, in the strategy $\mathbf{p}'$ defined above we have $\gamma_i(p_i + \delta_i) = (F + V_i)\delta_i$ and $\gamma_j(p_j - \delta_j) = -(F + V_j)\delta_j$. Thus, to prevent a profitable unilateral deviation of the defender from the strategy $\mathbf{p}$ to the strategy $\mathbf{p}'$, the strategy $\mathbf{q} = (q_1, \cdots, q_N)$ of the attacker must be such that, $\forall i, j \in \{1, \cdots, N\}, i \neq j$, $|\gamma_i(p_i + \delta_i)|q_i = |\gamma_j(p_j - \delta_j)|q_j$, i.e.,

$$(F + V_i)\delta_i q_i = (F + V_j)\delta_j q_j \quad (15)$$

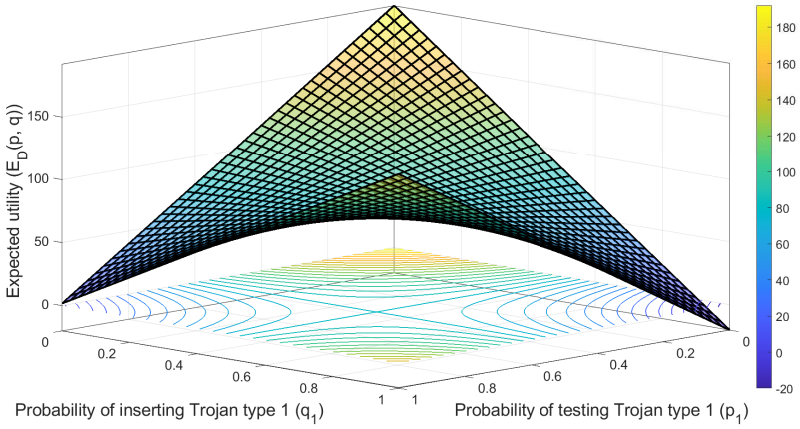
which implies, using (14),

$$q_j = q_i \frac{c_j}{c_i} \frac{F + V_i}{F + V_j},\ \forall i, j \in \{1, \cdots, N\}, i \neq j \quad (16)$$

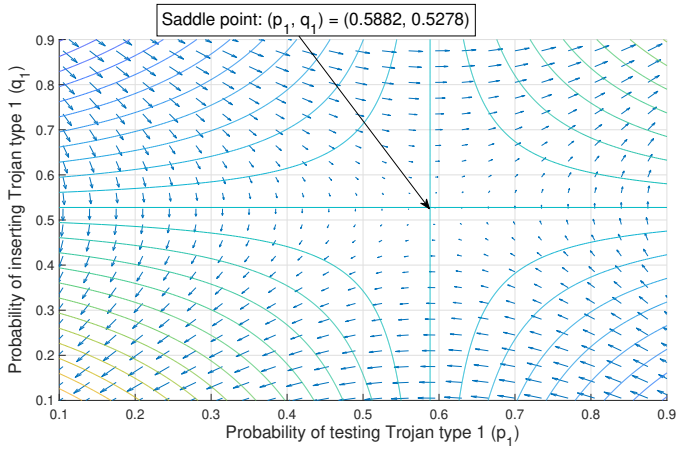
Now, it is easy to show that, for any chosen $i \in \{1, \cdots, N\}$, having $q_j = q_i \frac{c_j}{c_i} \frac{F+V_i}{F+V_j}, \forall j \in \{1, \cdots, N\}, j \neq i$, implies $q_j = q_i \frac{c_j}{c_i} \frac{F+V_i}{F+V_j}, \forall i, j \in \{1, \cdots, N\}, j \neq i$. Moreover, as noted earlier, any strategy $\mathbf{q}$ of the attacker forms a best response against the strategy $\mathbf{p}$ of the defender at NE defined earlier. Thus, we have $\sum_{i=1}^{N} q_i = k$ at NE (where k can be any value in $[0, 1]$), which can be expressed for any chosen $i \in \{1, \cdots, N\}$ as

$$\begin{aligned}
& q_i + \sum_{j=1, j\neq i}^{N} q_j = k \\
\Rightarrow\ & q_i + \sum_{j=1, j\neq i}^{N} q_i \frac{c_j}{c_i} \frac{F + V_i}{F + V_j} = k \qquad \text{(using (16))} \\
\Rightarrow\ & \sum_{j=1}^{N} q_i \frac{c_j}{c_i} \frac{F + V_i}{F + V_j} = k \\
\Rightarrow\ & q_i = \frac{k}{\frac{F+V_i}{c_i} \sum_{j=1}^{N} \frac{c_j}{F+V_j}}
\end{aligned} \quad (17)$$

[6]Note, it is easy to show that, for strategies $\grave{\mathbf{p}} = (\grave{p}_1, \cdots, \grave{p}_N)$ (where $\sum_{i=1}^{N} \grave{p}_i c_i < C$) and $\acute{\mathbf{p}} = (\acute{p}_1, \cdots, \acute{p}_N)$ (where $\sum_{i=1}^{N} \acute{p}_i c_i = C$) of the defender, it always holds true that $E_D(\acute{\mathbf{p}}, \mathbf{q}) \geq E_D(\grave{\mathbf{p}}, \mathbf{q})$ for any attacker's strategy $\mathbf{q}$, implying that $\acute{\mathbf{p}}$ dominates $\grave{\mathbf{p}}$.



(a) Expected utility ($E_D(\mathbf{p}, \mathbf{q})$) versus the defender's strategy (p_1) and the attacker's strategy (q_1).



(b) Contour plot.

Fig. 2. Expected utility ($E_D(\mathbf{p}, \mathbf{q})$) versus the defender's and the attacker's strategies for the sufficient cost budget case.

Clearly, from the above, if the attacker, for any chosen $i \in \{1, \cdots, N\}$ and any $k \in [0, 1]$, chooses q_i as given in (17) and q_j, $\forall j \in \{1, \cdots, N\}, j \neq i$, as given in (16), the attacker would be playing its best response against the strategy $\mathbf{p}$ of the defender defined earlier (which, recall, comprises of $p_i = \frac{B^S + V_i}{F + V_i}$ $\forall i \in \{1, \cdots, N\}$) without the defender having any profitable unilateral deviations from the strategy $\mathbf{p}$. This proves the theorem.

Next, we provide numerical results in Fig. 2 to corroborate Theorem 2 considering two Trojan types, viz. $\{1, 2\}$, with $B^S = 80$, $V_1 = 20$, $V_2 = 40$, $F = 150$, $c_1 = c_2 = 30$, and $C = \frac{B^S + V_1}{F + V_1} c_1 + \frac{B^S + V_2}{F + V_2} c_2 = 36.5944$. In Fig. 2(a), considering $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$, we present a 3-D plot of $E_D(\mathbf{p}, \mathbf{q})$ (11) versus the probability (p_1) with which the defender tests an acquired IC to check for the presence of Trojan type 1 (with $p_2 = \frac{C - p_1 c_1}{c_2}$ so that $p_1 c_1 + p_2 c_2 = C$) and the probability (q_1) with which the attacker inserts Trojan type 1 into the manufactured IC (with $q_2 = 1 - q_1$). In Fig. 2(b), we depict the contours and the gradient plot for $E_D(\mathbf{p}, \mathbf{q})$ in the $p_1 - q_1$ plane. From the figures, we observe

that there exists a saddle point whose coordinates are $(p_1, q_1) = (0.5882, 0.5278)$. Specifically, from Fig. 2(b), it can be seen that the gradient arrows point toward the point $(0.5882, 0.5278)$ in one direction and point outward from the point $(0.5882, 0.5278)$ in the perpendicular direction, implying that $(p_1, q_1) = (0.5882, 0.5278)$, with $(p_2, q_2) = (\frac{C-p_1c_1}{c_2}, 1-q_1) = (0.6316, 0.4722)$, is a saddle point (and hence the NE). It can be verified that the NE obtained from Theorem 2, considering $k = 1$ for the attacker, is also $(p_1, p_2) = (0.5882, 0.6316)$ and $(q_1, q_2) = (0.5278, 0.4722)$, which corroborates the theorem.

3.2 NE under Insufficient Cost Budget of the Defender

We now consider the scenario when $C < \sum_{i=1}^{N} \frac{B^S+V_i}{F+V_i} c_i$, which we refer to as the defender having an *insufficient cost budget*. Without loss of generality, consider $V_1 \leq V_2 \leq \cdots \leq V_N$ for the analysis of this scenario. In the next lemma, we provide the property that characterizes the defender's NE strategy in this case.

Lemma 3. *When $C < \sum_{i=1}^{N} \frac{B^S+V_i}{F+V_i} c_i$, the defender's strategy $\mathbf{p} = (p_1, \cdots, p_N)$ at NE is such that*

$$\gamma_j(p_j) = \min_{i \in \{1,\cdots,N\}} \gamma_i(p_i), \text{ if } p_j > 0 \tag{18}$$

Proof. Consider a strategy profile $\mathbf{p} = (p_1, \cdots, p_N)$ of the defender such that $\sum_{i=1}^{N} p_i c_i \leq C$ and denote $\underline{g} = \min_{i \in \{1,\cdots,N\}} \gamma_i(p_i)$. Clearly, $\underline{g} < 0$ (since, if $\underline{g} \geq 0$, we must have $p_i \geq \frac{B^S+V_i}{F+V_i}, \forall i \in \{1, \cdots, N\}$, which would violate the cost budget constraint). Define the set $\underline{G} = \{i | i \in \{1, \cdots, N\} \text{ and } \gamma_i(p_i) = \underline{g}\}$ and the set $\underline{G}' = \{1, \cdots, N\} - \underline{G}$. Since the attacker aims to minimize (11), the best response of the attacker against the strategy $\mathbf{p}$ is to adopt a strategy $\mathbf{q} = (q_1, \cdots, q_N)$ such that $\sum_{i \in \underline{G}} q_i = 1$. Suppose now that there exists $j \in \underline{G}'$ for which $p_j > 0$ (clearly, $q_j = 0$ in the aforementioned attacker's best response strategy $\mathbf{q}$ against $\mathbf{p}$). Consider now $w \in \underline{G}$ for which $q_w > 0$ (as follows from the aforementioned attacker's best response $\mathbf{q}$, such a w is guaranteed to exist) and consider changing the defender's strategy from $\mathbf{p} = (p_w, p_j, p_{-wj})$ to $\mathbf{p}' = (p_w + \delta_w, p_j - \delta_j, p_{-wj})$, where p_{-wj} denotes the vector of probabilities with which the defender tests all Trojan types except Trojan types w and j, while ensuring $\delta_w c_w \leq \delta_j c_j$ (to ensure that $\mathbf{p}'$ satisfies the cost budget constraint). Now, we have $\gamma_w(p_w + \delta_w) > \gamma_w(p_w)$ (which follows from Lemma 1) implying that $E_D(\mathbf{p}', \mathbf{q}) > E_D(\mathbf{p}, \mathbf{q})$, where $\mathbf{q}$ (as described earlier) forms a best response of the attacker against $\mathbf{p}$, showing that, with the attacker playing its best response, there exists a profitable unilateral deviation for the defender from any strategy $\mathbf{p}$ where there exists $j \in \underline{G}'$ for which $p_j > 0$, which proves the lemma.

We next present two important remarks based on Lemma 3.

Remark 1. Since, as discussed in the proof of Lemma 3, in the defender's strategy $\mathbf{p} = (p_1, \cdots, p_N)$ at NE, we must have $p_i = 0 \ \forall i \in \underline{G}'$ (and $\underline{g} < 0$), it follows using Lemma 1 that at NE $\gamma_j(p_j) < 0, \forall j \in \{1, \cdots, N\}$.

Remark 2. Since $\frac{d}{dV_i}\frac{B^S+V_i}{F+V_i} \geq 0$ *for* $B^S \leq F$ *(i.e.,* $\frac{B^S+V_i}{F+V_i}$ *is a non-decreasing function of* V_i *when* $B^S \leq F$*) and since* $\frac{d(\gamma_i(p_i))}{dp_i} \leq \frac{d(\gamma_j(p_j))}{dp_j}$ *if* $V_i \leq V_j$ *(which follows from Lemma 1), it can be noted that it follows from Lemma 3, considering* $V_1 \leq V_2 \leq \cdots \leq V_N$ *without loss of generality, that the NE strategy of the defender has the form* $\mathbf{p} = (p_1 = 0, \cdots, p_{k-1} = 0, p_k > 0, p_{k+1} > 0, \cdots, p_N > 0)$*, where* $k \in [1, N]$*.*

Next, using Lemma 3, Remark 1, and Remark 2, we present the NE of the insufficient cost budget case in Theorem 3.

Theorem 3. *When* $C < \sum_{i=1}^{N} \frac{B^S+V_i}{F+V_i} c_i$*, at NE,*

- *the defender's strategy corresponds to* $\mathbf{p} = (p_1 = 0, \cdots, p_{k-1} = 0, p_k = \frac{B^S+V_k}{F+V_k} - \delta_k, p_{k+1} = \frac{B^S+V_{k+1}}{F+V_{k+1}} - \delta_{k+1} \cdots, p_N = \frac{B^S+V_N}{F+V_N} - \delta_N)$*, where,* $\delta_k = \frac{\sum_{i=k}^{N}\left(\frac{B^S+V_i}{F+V_i}\right)c_i - C}{(F+V_k)\sum_{i=k}^{N}\frac{c_i}{F+V_i}}$*,* $\delta_i = \frac{F+V_k}{F+V_i}\delta_k, \forall i \in [k+1, N]$*, and* $k \in [1, N]$ *is the least value that satisfies* $\delta_k = \frac{\sum_{i=k}^{N}\left(\frac{B^S+V_i}{F+V_i}\right)c_i - C}{(F+V_k)\sum_{i=k}^{N}\frac{c_i}{F+V_i}} < \frac{B^S+V_k}{F+V_k}$*, and*
- *the attacker's strategy corresponds to, for any chosen* $i \in \underline{G}$*, where* $\underline{G} = \{i | i \in \{1, \cdots, N\}$ *and* $\gamma_i(p_i) = \min_{i \in \{1,\cdots,N\}} \gamma_i(p_i)\}$ *with* p_i *being the defender's strategy of testing against Trojan type* i *at NE, choosing* $q_i = \frac{1}{\frac{F+V_i}{c_i}\sum_{j=1}^{N}\frac{c_j}{F+V_j}}$ *and* $q_j = q_i \frac{c_j}{c_i}\frac{F+V_i}{F+V_j}$*,* $\forall j \in \underline{G}$*,* $j \neq i$*.*

Proof. Using Remark 1 and Remark 2, let us represent the NE strategy of the defender as $\mathbf{p} = (p_1 = 0, \cdots, p_{k-1} = 0, p_k = \frac{B^S+V_k}{F+V_k} - \delta_k, p_{k+1} = \frac{B^S+V_{k+1}}{F+V_{k+1}} - \delta_{k+1}, \cdots, p_N = \frac{B^S+V_N}{F+V_N} - \delta_N)$ for $k \in [1, N]$ and $\delta_i \in (0, \frac{B^S+V_i}{F+V_i})$ for $i \in [k, N]$. Now, to have $\mathbf{p}$ satisfy the cost budget constraint, we must have

$$\sum_{i=1}^{k-1}(0 \cdot c_i) + \sum_{i=k}^{N}\left(\frac{B^S+V_i}{F+V_i} - \delta_i\right) \cdot c_i = C$$
$$\Rightarrow \sum_{i=k}^{N}\delta_i c_i = \sum_{i=k}^{N}\left(\frac{B^S+V_i}{F+V_i}\right)c_i - C \tag{19}$$

Now, recalling from Lemma 1 that $\frac{d\gamma_x(p_x)}{dp_x} = F + V_x$ and that $\gamma_x(\frac{B^S+V_x}{F+V_x}) = 0$, $x \in \{1, \cdots, N\}$, and noting that Lemma 3 implies that in the defender's NE strategy we have $\gamma_k(p_k) = \gamma_{k+1}(p_{k+1}) = \cdots = \gamma_N(p_N)$, we conclude that in the strategy $\mathbf{p}$ of the defender at NE we have

$$(F + V_k)\delta_k = (F + V_{k+1})\delta_{k+1} = \cdots = (F + V_N)\delta_N \tag{20}$$

which implies that

$$\delta_i = \frac{F + V_k}{F + V_i}\delta_k, \forall i \in [k+1, N] \tag{21}$$

Substituting (21) into (19), we get

$$\sum_{i=k}^{N} \delta_k \frac{F+V_k}{F+V_i} c_i = \sum_{i=k}^{N} \left(\frac{B^S+V_i}{F+V_i} \right) c_i - C$$

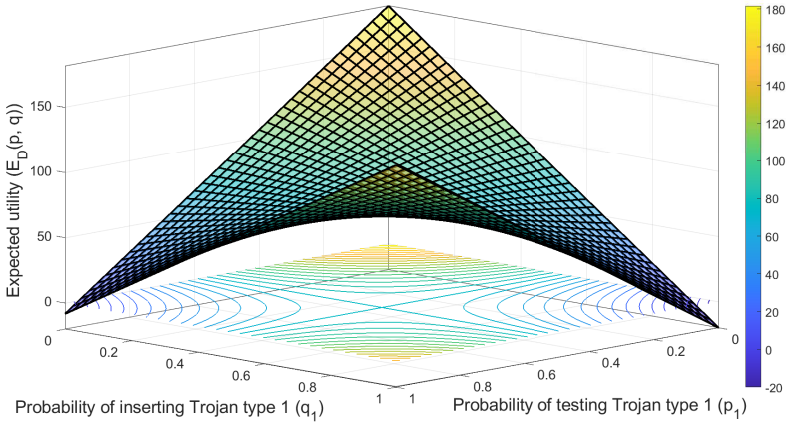
$$\Rightarrow \ \delta_k = \frac{\sum_{i=k}^{N} \left(\frac{B^S+V_i}{F+V_i} \right) c_i - C}{(F+V_k) \sum_{i=k}^{N} \frac{c_i}{F+V_i}} \tag{22}$$

Now, it is easy to show that higher the value of k chosen (while having $\gamma_k(p_k) = \cdots = \gamma_N(p_N)$ and satisfying the cost budget constraint), lower would be $\min_{i \in \{1,\cdots,N\}} \gamma_i(p_i)$, which implies that lower would be the expected utility of the defender against a strategic attacker. Thus, in the strategy $\mathbf{p} = (p_1 = 0, \cdots, p_{k-1} = 0, p_k = \frac{B^S+V_k}{F+V_k} - \delta_k, p_{k+1} = \frac{B^S+V_{k+1}}{F+V_{k+1}} - \delta_{k+1}, \cdots, p_N = \frac{B^S+V_N}{F+V_N} - \delta_N)$ of the defender at NE, to have $\frac{B^S+V_i}{F+V_i} - \delta_i > 0$, $i \in [k, N]$, the defender must choose the least value of $k \in [1, N]$ such that δ_k (22) satisfies

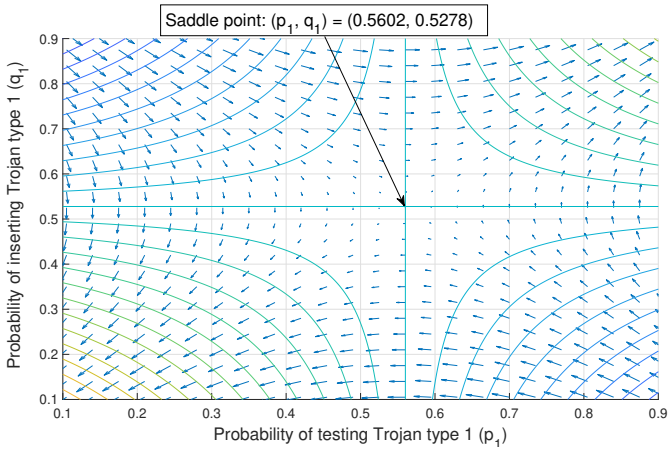
$$\delta_k = \frac{\sum_{i=k}^{N} \left(\frac{B^S+V_i}{F+V_i} \right) c_i - C}{(F+V_k) \sum_{i=k}^{N} \frac{c_i}{F+V_i}} < \frac{B^S+V_k}{F+V_k} \tag{23}$$

with δ_i, $\forall i \in [k+1, N]$, chosen as given in (21), which proves the defender's NE strategy as given in the theorem. Against such a strategy $\mathbf{p}$ of the defender, since the attacker seeks to minimize (11), the best response of the attacker becomes adopting a strategy $\mathbf{q} = (q_1, \cdots, q_N)$ such that $\sum_{i \in \underline{G}} q_i = 1$, where $\underline{G} = \{i | i \in \{1, \cdots, N\}$ and $\gamma_i(p_i) = \min_{i \in \{1,\cdots,N\}} \gamma_i(p_i)\}$. However, not all such strategies of the attacker comprise a NE. It can be shown, using an approach similar to the proof of Theorem 2, which we omit for brevity, that the NE strategy of the attacker consists of, for any chosen $i \in \underline{G}$, inserting Trojan type i into the manufactured IC with a probability $q_i = \frac{1}{\frac{F+V_i}{c_i} \sum_{j=1}^{N} \frac{c_j}{F+V_j}}$ and inserting Trojan type j with a probability $q_j = q_i \frac{c_j}{c_i} \frac{F+V_i}{F+V_j}$, $\forall j \in \underline{G}$, $j \neq i$. This proves the theorem.

Next, we provide numerical results in Fig. 3 to corroborate Theorem 3 considering two Trojan types, viz. $\{1, 2\}$, with $B^S = 80$, $V_1 = 20$, $V_2 = 40$, $F = 150$, $c_1 = c_2 = 30$, and $C = 35$. In Fig. 3(a), considering $\mathbf{p} = (p_1, p_2)$ and $\mathbf{q} = (q_1, q_2)$, we present a 3-D plot of $E_D(\mathbf{p}, \mathbf{q})$ (11) versus the probability (p_1) with which the defender tests an acquired IC to check for the presence of Trojan type 1 (with $p_2 = \frac{C - p_1 c_1}{c_2}$ so that $p_1 c_1 + p_2 c_2 = C$) and the probability (q_1) with which the attacker inserts Trojan type 1 into the manufactured IC (with $q_2 = 1 - q_1$). In Figure 3(b), we depict the contours and the gradient plot for $E_D(\mathbf{p}, \mathbf{q})$ in the $p_1 - q_1$ plane. From the figures, we observe that there exists a saddle point whose coordinates are $(p_1, q_1) = (0.5602, 0.5278)$. Specifically, from Fig. 3(b), it can be seen that the gradient arrows point toward the point $(0.5602, 0.5278)$ in one direction and point outward from the point $(0.5602, 0.5278)$ in the perpendicular direction, implying that $(p_1, q_1) = (0.5602, 0.5278)$, with $(p_2, q_2) =$



(a) Expected utility ($E_D(\mathbf{p}, \mathbf{q})$) versus the defender's strategy (p_1) and the attacker's strategy (q_1).



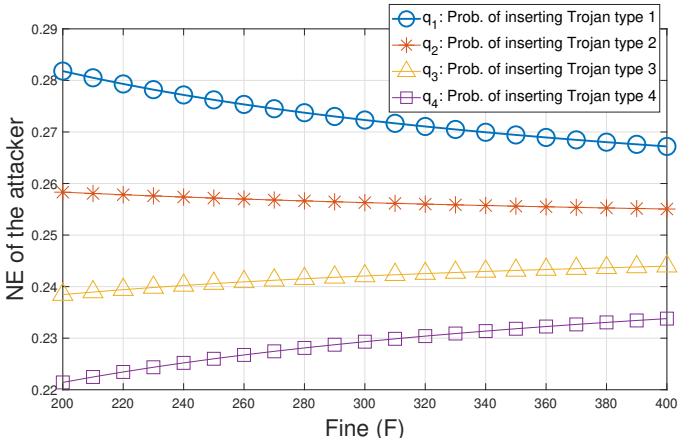
(b) Contour plot.

Fig. 3. Expected utility ($E_D(\mathbf{p}, \mathbf{q})$) versus the defender's and the attacker's strategies for the insufficient cost budget case.

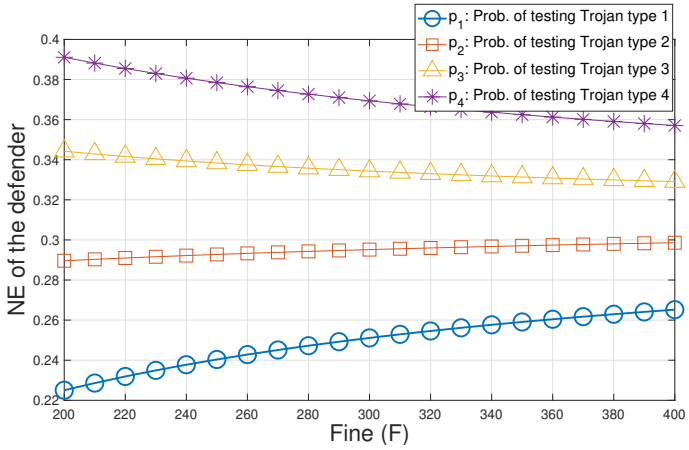
$(\frac{C-p_1 c_1}{c_2}, 1-q_1) = (0.6065, 0.4722)$, is a saddle point (and hence the NE). It can be verified that the NE obtained from Theorem 3 is also $(p_1, p_2) = (0.5602, 0.6065)$ and $(q_1, q_2) = (0.5278, 0.4722)$, which corroborates the theorem.

4 Numerical Results

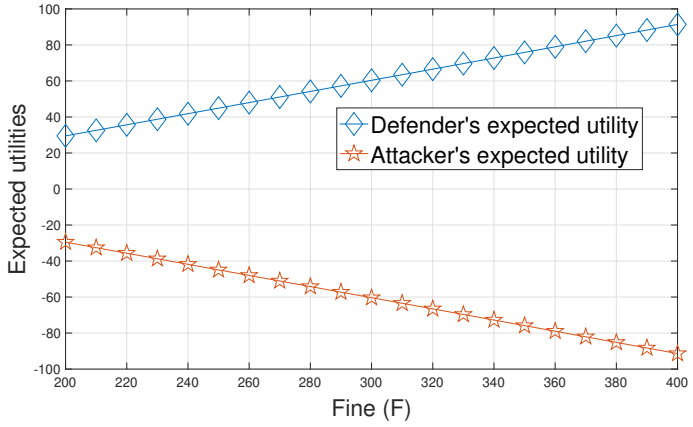
In this section, we provide numerical results to provide important insights into our developed game theoretic Trojan testing strategies. In Fig. 4, we show the impact of the fine (F) on the strategies of the attacker and the defender and on their expected utilities at NE for the insufficient cost budget case. For the figure, we consider four types of Trojans, viz. $\{1, 2, 3, 4\}$, with $B^S = 100$, $V_1 = 20$, $V_2 = 40$, $V_3 = 60$, $V_4 = 80$, $c_1 = c_2 = c_3 = c_4 = 40$, and $C = 50$. The NE strategies for the figure have been computed using Theorem 3. As can be seen from Fig. 4(a),



(a) NE strategy of the attacker versus fine (F).



(b) NE strategy of the defender versus fine (F).



(c) Expected utilities of the defender and attacker versus fine (F) at NE.

Fig. 4. Impact of fine (F) on the defender's and the attacker's strategies and on their expected utilities at NE.

as F increases, at NE, the attacker increases its probability of inserting a Trojan which is relatively more damaging in nature (which corresponds to Trojan types 3 and 4 having $V_3 = 60$ and $V_4 = 80$, respectively) while decreasing its probability of inserting a Trojan which is relatively less damaging (which corresponds to Trojan types 1 and 2 having $V_1 = 20$ and $V_2 = 40$, respectively). This can be attributed to the fact that, since F negatively impacts the attacker's utility, increasing the probability of inserting a more damaging Trojan as F increases helps the attacker counteract the negative impact of having to pay a heftier fine upon the defender correctly detecting an inserted Trojan. For the defender's strategy at NE, with increasing F, it can be noted from Fig. 4(b) that the defender increases its probabilities of testing the acquired IC to check for the presence of relatively less damaging Trojans (which correspond to Trojan types 1 and 2) and decreases its probabilities of testing the IC against more damaging Trojans (which correspond to Trojan types 3 and 4).

Further, from Fig. 4(b) it can be observed that, as is intuitive, for any given F, at NE, the defender tests an acquired IC against a more damaging Trojan with a higher probability than that of testing against a less damaging one while, as can be seen from Fig. 4(a), the attacker exhibits the reverse trend. As can be seen from Fig. 4(c), the expected utility of the defender at NE increases with F, and accordingly the attacker's expected utility decreases with F, indicating the advantage that charging a higher fine has on enhancing the defender's utility against strategic insertion of hardware Trojans by malicious manufacturers.

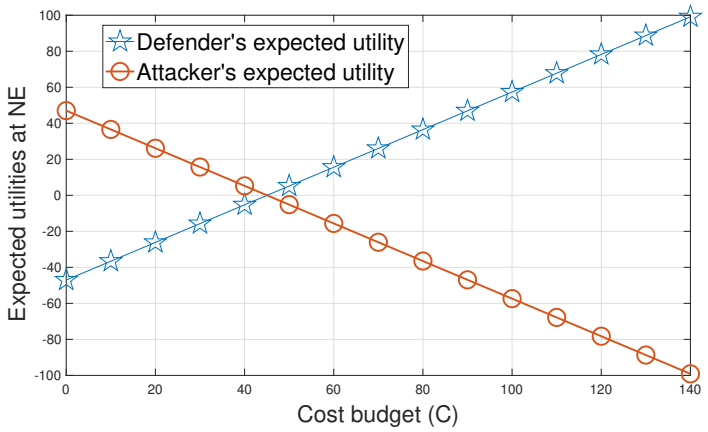


Fig. 5. Expected utilities of the defender and the attacker at NE versus the cost budget (C).

In Fig. 5, we show the expected utilities of the defender and the attacker at NE (computed using Theorem 3) versus the cost budget (C) available for performing testing (considering C taken in the figure to satisfy the insufficient cost budget case). For the figure, we consider four Trojan types, viz. $\{1, 2, 3, 4\}$, with $B^S = 100$, $F = 120$, $V_1 = 20$, $V_2 = 40$, $V_3 = 60$, $V_4 = 80$, and $c_1 = c_2 = c_3 = c_4 = 40$. As can be seen from the figure, the expected utility of the defender

at NE increases as the cost budget (C) of the defender for performing testing increases. This is because, with increasing C, the capability of the defender to test more types of Trojans increases which enhances its ability to correctly determine whether the acquired IC contains a Trojan which, in turn, enhances the defender's capability to avoid the damage caused by an inserted Trojan and impose a fine on the malicious manufacturer (both of which positively impact the defender's utility). As can also be seen from the figure, as expected, the attacker's expected utility decreases with increasing C. Such trends in the expected utilities show the ability of our characterized NE strategies to tactfully exploit the cost budget available for performing testing to enhance the utility of the defender against a strategic attacker.

5 Conclusion

This paper investigated the problem of game theoretic hardware Trojan testing and analytically characterized NE strategies for inserting a Trojan (from the perspective of a malicious manufacturer) and testing a Trojan (from the perspective of a defender) in closed-forms under consideration of testing costs incurred by the defender. The paper first characterized the NE for the case where the defender, who incurs costs for performing testing, is capable of testing an acquired IC against one Trojan type. The paper also characterized the NE for the case where the defender can test the acquired IC against multiple types of Trojans under a cost budget constraint. Numerical results were presented to gain important insights into the NE strategies characterized in the paper.

References

1. Adee, S.: The hunt for the kill switch. IEEE Spectrum **45**(5), 34–39 (2008)
2. Agrawal, D., Baktir, S., Karakoyunlu, D., Rohatgi, P., Sunar, B.: Trojan detection using ic fingerprinting. In: 2007 IEEE Symposium on Security and Privacy (SP'07). pp. 296–310. IEEE (2007)
3. Banga, M., Hsiao, M.S.: A region based approach for the identification of hardware trojans. In: 2008 IEEE International Workshop on Hardware-Oriented Security and Trust. pp. 40–47. IEEE (2008)
4. Bhasin, S., Regazzoni, F.: A survey on hardware trojan detection techniques. In: 2015 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 2021–2024. IEEE (2015)
5. Bhunia, S., Hsiao, M.S., Banga, M., Narasimhan, S.: Hardware trojan attacks: Threat analysis and countermeasures. Proceedings of the IEEE **102**(8), 1229–1247 (2014)
6. Brahma, S., Nan, S., Njilla, L.: Strategic hardware trojan testing with hierarchical trojan types. In: 2021 55th Annual Conference on Information Sciences and Systems (CISS). pp. 1–6 (2021)
7. Chakraborty, R.S., Narasimhan, S., Bhunia, S.: Hardware trojan: Threats and emerging solutions. In: 2009 IEEE International high level design validation and test workshop. pp. 166–171. IEEE (2009)

8. Chakraborty, R.S., Wolff, F., Paul, S., Papachristou, C., Bhunia, S.: Mero: A statistical approach for hardware trojan detection. In: International Workshop on Cryptographic Hardware and Embedded Systems. pp. 396–410. Springer (2009)
9. Fudenberg, D., Tirole, J.: Game Theory. MIT Press, Cambridge, MA (1991)
10. Graf, J.: Trust games: How game theory can guide the development of hardware trojan detection methods. In: 2016 IEEE International Symposium on Hardware Oriented Security and Trust (HOST). pp. 91–96. IEEE (2016)
11. Graf, J., Batchelor, W., Harper, S., Marlow, R., Carlisle, E., Athanas, P.: A practical application of game theory to optimize selection of hardware trojan detection strategies. Journal of Hardware and Systems Security **4**(2), 98–119 (2020)
12. Hu, N., Ye, M., Wei, S.: Surviving information leakage hardware trojan attacks using hardware isolation. IEEE Transactions on Emerging Topics in Computing **7**(2), 253–261 (2017)
13. Kamhoua, C.A., Zhao, H., Rodriguez, M., Kwiat, K.A.: A game-theoretic approach for testing for hardware trojans. IEEE Transactions on Multi-Scale Computing Systems **2**(3), 199–210 (2016)
14. Kwiat, K., Born, F.: Strategically managing the risk of hardware trojans through augmented testing. In: 13th Annual Symposium on Information Assurance (ASIA). pp. 20–24 (2018)
15. Nagarajan, K., De, A., Khan, M.N.I., Ghosh, S.: Trapped: Dram trojan designs for information leakage and fault injection attacks. arXiv preprint arXiv:2001.00856 (2020)
16. Rajendran, J., Gavas, E., Jimenez, J., Padman, V., Karri, R.: Towards a comprehensive and systematic classification of hardware trojans. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. pp. 1871–1874. IEEE (2010)
17. Salmani, H., Tehranipoor, M., Plusquellic, J.: New design strategy for improving hardware trojan detection and reducing trojan activation time. In: 2009 IEEE International Workshop on Hardware-Oriented Security and Trust. pp. 66–73. IEEE (2009)
18. Schulze, T.E., Kwiat, K., Kamhoua, C., Chang, S.C., Shi, Y.: 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. IEEE (2016)