2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Exploiting Trust for Resilient Hypothesis Testing with Malicious Robots

Matthew Cavorsi*, Orhan Eren Akgün*, Michal Yemini, Andrea J. Goldsmith, and Stephanie Gil

Abstract— **We develop a resilient binary hypothesis testing framework for decision making in adversarial multi-robot crowdsensing tasks. This framework exploits stochastic trust observations between robots to arrive at tractable, resilient decision making at a centralized Fusion Center (FC) even when i) there exist malicious robots in the network and their number may be larger than the number of legitimate robots, and ii) the FC uses one-shot noisy measurements from all robots. We derive two algorithms to achieve this. The first is the Two Stage Approach (2SA) that estimates the legitimacy of robots based on received trust observations, and provably minimizes the probability of detection error in the worst-case malicious attack. Here, the proportion of malicious robots is known but arbitrary. For the case of an unknown proportion of malicious robots, we develop the Adversarial Generalized Likelihood Ratio Test (A-GLRT) that uses both the reported robot measurements and trust observations to estimate the trustworthiness of robots, their reporting strategy, and the correct hypothesis simultaneously. We exploit special problem structure to show that this approach remains computationally tractable despite several unknown problem parameters. We deploy both algorithms in a hardware experiment where a group of robots conducts crowdsensing of traffic conditions on a mock-up road network similar in spirit to Google Maps, subject to a Sybil attack. We extract the trust observations for each robot from actual communication signals which provide statistical information on the uniqueness of the sender. We show that even when the malicious robots are in the majority, the FC can reduce the probability of detection error to 30.5% and 29% for the 2SA and the A-GLRT respectively.**

I. Introduction

We are interested in the problem where robots observe the environment and estimate the presence of an event of interest. Each robot relays their measurement to a *Fusion Center* (FC) that makes an informed binary decision on the occurrence of the event. An unknown subset of the network are malicious robots whose goal is to increase the likelihood that the FC makes a wrong decision [1]–[4]. This problem can be cast as an adversarial binary hypothesis testing problem, with relevance to a broad class of robotics tasks that rely on distributed sensing with possibly malicious or untrustworthy robots. For example, robots might perform coordinated coverage to maximize their ability to sense events of interest [5]–[8], share target information for coordinated tracking [9]–[12], or merge map information to provide a global understanding of the environment [13]–[16]. In crowdsensing tasks such as traffic prediction, a server may use GPS data to estimate if a particular roadway is congested or not [17] (see Fig. 1). Unfortunately, this process is vulnerable to malicious robots [1], [3]. For example, prior works have shown that a Sybil attack can cause crowdsensing applications like Google Maps to incorrectly perceive traffic conditions, resulting in erroneous reporting of traffic flows [18], [19].

The problem of binary adversarial hypothesis testing has been studied within the context of sensor networks [20]–[22]. Many approaches use data, such as a history of measurements and hypothesis outcomes, to assess the trustworthiness of the robots [23]–[26]. For example, if a robot consistently disagrees with the final decision of the FC, then the FC can flag that robot as potentially adversarial. However, the success of these methods often hinges upon a crucial assumption that more than half of the network is legitimate. A growing body of work investigates additionally sensed quantities arising from the *physicality of cyber-physical systems* such as multi-robot networks, to cross-validate and assess the trustworthiness of robots [5], [27]–[29]. This could include using camera feeds, GPS signals, or even the signatures of received wireless communication signals, to acquire additional information regarding *the trustworthiness* of the robots [29]–[31]. Importantly, this class of trust observations can often be obtained from a one-shot observation, independent of the transmitted measurement. The works [32], [33] use trust observations to recover resilient consensus and distributed optimization even in the case where more than half of the network is malicious. In this paper *we wish to derive a framework for adversarial hypothesis testing that allows the FC to reduce its probability of error, even in the one-shot scenario and where legitimate robots do not hold a majority in the network, by exploiting stochastic trust observations that are independent of the transmitted measurements.*

We derive algorithms for achieving resilient hypothesis testing by exploiting stochastic trust observations between the FC and a group of robots participating in event detection. We derive a framework that exploits one-shot trust observations, hereafter called *trust values*, over each link to arrive at *tractable, closed-form solutions* when the majority of the network may be malicious and

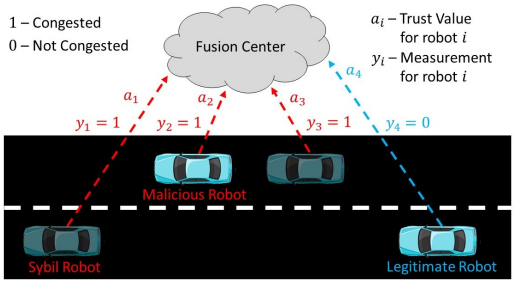

Fig. 1: Malicious robots can perform a Sybil Attack to try to force a FC to incorrectly perceive traffic conditions on a road. The FC can aggregate measurements and trust values from robots to accurately estimate the true traffic condition of the road despite the attack.

(*Co-primary authors). M. Cavorsi, O. E. Akgün, and S. Gil are with the School of Engineering and Applied Sciences, Harvard University, USA: mcavorsi@g.harvard.edu, erenakgun@g.harvard.edu, sgil@seas.harvard.edu. M. Yemini is with the Faculty of Engineering, Bar-Ilan University, Israel: michal.yemini@biu.ac.il. A. J. Goldsmith is with the Department of Electrical and Computer Engineering, Princeton University, USA: goldsmith@princeton.edu.

The authors gratefully acknowledge partial support through AFOSR grant FA9550-22-1-0223 and AFOSR award #002484665.

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160385

the strategy of the malicious robots is unknown – a challenging and otherwise intractable problem to solve in the general case [34].

For the case where an upper limit on the proportion of malicious robots is known, we develop the *Two Stage Approach* (2SA). In the first stage this algorithm uses trust values to determine the most likely set of malicious robots, and then applies a Likelihood Ratio Test (LRT) only over trusted robots in the second stage. We show that this approach minimizes the error probability of the estimated hypothesis at the FC for a worst-case attack scenario. For the case where an upper bound on the proportion of malicious robots is unknown, we develop the *Adversarial Generalized Likelihood Ratio Test* (A-GLRT) algorithm which uses both stochastic trust values and event measurements to jointly estimate the trustworthiness of each robot, the strategy of malicious robots, and the hypothesis of the event. Our A-GLRT algorithm is based upon a common approach for decision making with unknown parameters, the Generalized Likelihood Ratio Test (GLRT), which replaces the unknown parameters with their maximum likelihood estimates (MLE) [35]. We show that the addition of trust values allows us to decouple the trustworthiness estimation from the strategy of the adversaries, allowing us to calculate the exact MLE of unknown parameters in polynomial time, instead of approximating them as in previous works [34], [36]. Our simulation results show that the A-GLRT empirically yields a lower probability of error than the 2SA, but at the expense of higher computational cost.

Finally we conduct a hardware experiment based on crowdsensing traffic conditions using a group of robots under a Sybil Attack. We show that the FC can recover a performance of 30.5% and 29.0% error, for the 2SA and A-GLRT respectively, even in the case where more than half of the robots are malicious.

II. Problem Formulation

We consider a network of N robots, where each robot is indexed by some $i \in \mathcal{N}$ and $\mathcal{N} = \{1,...,N\}$, that are deployed to sense an environment and determine if an event of interest has occurred. The event of interest is captured by the random variable Ξ, where $\Xi = 1$ if the event has occurred and $\Xi = 0$ otherwise. Each robot i uses its sensed information to make a local decision about whether the event has happened or not, captured by the random variable Y_i, where its realization $y_i = 1$ if robot i believes the event has happened and $y_i = 0$ otherwise. We denote the hypothesis that $\Xi = 1$ by $\mathcal{H}_1$ and $\Xi = 0$ by $\mathcal{H}_0$. Each robot forwards its local decision to a centralized fusion center (FC).

We are concerned with the scenario where not all robots are trustworthy, that is, some are *malicious* and may manipulate the data that they send to the FC by flipping their measured bit with the goal of increasing the probability that the FC makes the wrong decision. We denote the set of malicious robots by $\mathcal{M} \subset \mathcal{N}$. The set of robots that are not malicious are termed *legitimate robots*, denoted by $\mathcal{L} \subseteq \mathcal{N}$, where $\mathcal{L} \cup \mathcal{M} = \mathcal{N}$ and $\mathcal{L} \cap \mathcal{M} = \emptyset$. Additionally, we define the true trust vector, $\mathbf{t} \in \{0,1\}^N$, where $t_i = 1$ if $i \in \mathcal{L}$ and $t_i = 0$ if $i \in \mathcal{M}$. We note that the true trust vector is unknown by the FC, but it is defined for analytical purposes.

We assume the following behavioral models for robots:

Definition 1 (Legitimate robot): A *legitimate robot* i measures the event and sends its measurement Y_i to the FC without altering it. We assume for each robot $i \in \mathcal{L}$, the measurement is subject to noise with the following false alarm and missed detection probabilities

$$\begin{aligned} P_{FA,i} &= \Pr(Y_i = 1 | \Xi = 0, t_i = 1) = P_{FA,L}, \\ P_{MD,i} &= \Pr(Y_i = 0 | \Xi = 1, t_i = 1) = P_{MD,L}, \end{aligned} \tag{1}$$

where $P_{FA,L} \in (0, 0.5)$ and $P_{MD,L} \in (0, 0.5)$ without loss of generality. We assume all legitimate robots have the same $P_{FA,L}$ and $P_{MD,L}$ and these are known by the FC.

Definition 2 (Malicious robot): A robot is said to be a *malicious robot* if it can choose to alter its measurements before sending it to the FC. We assume that all malicious robots have the same probabilities of false alarm and missed detection, denoted by $P_{FA,M}$ and $P_{MD,M}$ respectively. These probabilities are ***not*** known by the FC, and can take any value in $[0,1]$.

We assume that all measurements are independent of each other given the true hypothesis. Furthermore, the assumption we use that measurements coming from malicious robots are i.i.d. is common in the literature (see [2], [4], [21], [22]). In addition to the measurements Y_i, we assume that each Y_i is tagged with a *trust value* $\alpha_i \in \mathbb{R}$. Specifically, we consider the class of problems where the FC can leverage the cyber-physical nature of the network to extract an estimation of trust about each communicating robot.

Definition 3 (Trust Value α_i): A *trust value* α_i is a stochastic variable that captures information about the true legitimacy of a robot i. We denote the set of all possible trust values by $\mathcal{A}$ and denote a trust value realization for robot i by a_i.

Assumption 1: We assume that the set $\mathcal{A}$ is finite and that the trust values are i.i.d. given the true legitimacy of the robot, and are independent of the measurements, Y_i, and the true hypothesis. We denote the probability mass function of the trust values by $p_\alpha(a_i|t_i = 1)$ and $p_\alpha(a_i|t_i = 0)$ for legitimate and malicious robots, respectively. We assume the probability mass functions are known or can be estimated by the FC for all i. [1] Finally, to omit trivial or noninformative cases, we assume that $p_\alpha(a_i|t_i = 0) \in (0,1)$, $p_\alpha(a_i|t_i = 1) \in (0,1)$, and $p_\alpha(a_i|t_i = 0) \neq p_\alpha(a_i|t_i = 1)$ for all $a \in \mathcal{A}$ and all i.

A. The objective of the FC

Denote the vector of all measurements with $\boldsymbol{Y} = (Y_1,...,Y_N)$ and its realization $\boldsymbol{y} = (y_1,...,y_N)$, and the vector of stochastic trust values by $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_N)$ and its realization by $\boldsymbol{a} = (a_1,...,a_N)$. Let $\mathcal{D}_0$ and $\mathcal{D}_1$ be the decision regions at the FC. That is, $(\boldsymbol{a},\boldsymbol{y}) \in \mathcal{D}_0$ if the FC chooses hypothesis $\mathcal{H}_0$ whenever it measures the pair $(\boldsymbol{a}, \boldsymbol{y})$, and $(\boldsymbol{a}, \boldsymbol{y}) \in \mathcal{D}_1$ otherwise. To simplify our notations we denote $\mathcal{D} \triangleq \{\mathcal{D}_0,\mathcal{D}_1\}$.

Denote by P_{FA} and P_{MD} the false alarm and missed detection probabilities of the FC. These probabilities are affected by the strategy of the malicious robots, i.e., $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$. We have that:

$$\begin{aligned} P_{\text{FA}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}}) &= \sum_{(\boldsymbol{a},\boldsymbol{y}) \in \mathcal{D}_1} \Pr(\boldsymbol{\alpha} = \boldsymbol{a}, \boldsymbol{Y} = \boldsymbol{y} | \mathcal{H}_0, \boldsymbol{t}, P_{\text{FA,M}}), \\ P_{\text{MD}}(\mathcal{D},\boldsymbol{t},P_{\text{MD,M}}) &= \sum_{(\boldsymbol{a},\boldsymbol{y}) \in \mathcal{D}_0} \Pr(\boldsymbol{\alpha} = \boldsymbol{a}, \boldsymbol{Y} = \boldsymbol{y} | \mathcal{H}_1, \boldsymbol{t}, P_{\text{MD,M}}), \end{aligned}$$

[1] In [30]–[33], for example, the trust values $\alpha_i \in [0,1]$ are stochastic and are determined from physical properties of wireless transmissions. We use these trust values in our hardware experiment in Section IV where we discretize the sample space by letting $\mathcal{A} = \{0,1\}$ and find the probability mass functions to be $p_\alpha(a_i = 1|t_i = 1) = 0.8350$ and $p_\alpha(a_i = 1|t_i = 0) = 0.1691$. Other examples of observations can be found in [27], [37], [38].

where we omit the dependence of P_{FA} and P_{MD} on the parameters that are assumed to be given such as $P_{\text{FA,L}}$ and $P_{\text{MD,L}}$.

If the FC knows the true trust vector $\boldsymbol{t}$, and the probabilities $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$, it could optimize the decision regions $\mathcal{D}_0$ and $\mathcal{D}_1$ to minimize the expected error probability:

$$P_{\text{e}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}},P_{\text{MD,M}}) = \Pr(\Xi=0)P_{\text{FA}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}})+\Pr(\Xi=1)P_{\text{MD}}(\mathcal{D},\boldsymbol{t},P_{\text{MD,M}}). \tag{2}$$

However, there are two main obstacles to the optimization of (2), namely: 1) The FC does not know the true vector $\boldsymbol{t}$. 2) The FC does not know the strategy of malicious robots. In our setup, this means that the FC does not know the values $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$.

Because of this, the FC must estimate the vector $\boldsymbol{t}$, and the probabilities $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$, and make a decision with these estimates, denoted by $\hat{\boldsymbol{t}}$, $\hat{P}_{\text{FA,M}}$, and $\hat{P}_{\text{MD,M}}$. Since the minimization of (2) is not tractable, we explore different ways to circumvent this issue. One way is to start by estimating the legitimacy of the robots using trust values only and choosing the decision regions $\mathcal{D}_0$ and $\mathcal{D}_1$ using the measurements from the trusted robots. This approach leads us to the formulation in Problem 1.

Problem 1: Assume that the FC first estimates the identities of the robots in the network, i.e., it determines $\hat{\boldsymbol{t}}$, solely using the vector of trust values $\boldsymbol{a}$. Then, it makes a decision using only the vector of measurements $\mathbf{y}$, from robots it identifies as legitimate. Given an upper bound $\bar{m}$ on the proportion of malicious robots in the network, we wish to determine a strategy for the FC that minimizes the following worst-case scenario under these assumptions:

$$\min_{\mathcal{D}} \max_{P_{\text{FA,M}},P_{\text{MD,M}},\boldsymbol{t}:\sum_{i\in\mathcal{N}}t_i\leq\bar{m}N} P_{\text{e}}(\mathcal{D},\boldsymbol{t},P_{FA,M},P_{MD,M}).$$

Problem 1 estimates the trustworthiness of a robot i using only the trust value a_i associated with that robot. However, it is natural to seek additional information about the trustworthiness of the robots that can be obtained from the random measurement vector $\boldsymbol{y}$. Following this intuition, we seek a decision rule that estimates the unknown parameters in the system, i.e., $\boldsymbol{t}$, $P_{\text{FA,M}}$, and $P_{\text{MD,M}}$ as well as the hypothesis $\mathcal{H}_0$ or $\mathcal{H}_1$ jointly, without requiring a known upper bound on the proportion of malicious robots. A common approach to hypothesis testing with unknown parameters is to use the generalized likelihood ratio test (GLRT) [35], i.e.,

$$\frac{p(\boldsymbol{z};\hat{\theta}_1,\mathcal{H}_1)}{p(\boldsymbol{z};\hat{\theta}_0,\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)} \triangleq \gamma_{\text{GLRT}}, \tag{3}$$

where $\hat{\theta}_1$ is MLE of the unknown parameter θ_1 assuming $\Xi=1$ and $\hat{\theta}_0$ is the MLE of θ_0 assuming $\Xi=0$. For our problem, $\boldsymbol{z}=(\boldsymbol{a},\boldsymbol{y})$, $\theta_1=(\boldsymbol{t},P_{\text{MD,M}})$, and $\theta_0=(\boldsymbol{t},P_{\text{FA,M}})$, which results in the following formulation

$$\frac{\max_{\boldsymbol{t}\in\{0,1\}^N,P_{\text{MD,M}}\in[0,1]}\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})}{\max_{\boldsymbol{t}\in\{0,1\}^N,P_{\text{FA,M}}\in[0,1]}\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{\text{GLRT}}. \tag{4}$$

Note that the vector $\boldsymbol{t}$ is a parameter, thus, we do not make any prior assumption on its distribution. Calculating the MLE in the numerator and denominator in (4) is not trivial since the unknown $\boldsymbol{t}$ is a discrete multidimensional variable while $P_{\text{MD,M}}$ and $P_{\text{FA,M}}$ are continuous variables. Doing this in a tractable way leads us to the formulation in Problem 2.

Problem 2: Find a computationally tractable algorithm that calculates the GLRT given in (4).

In the next section we propose solutions to these problems. Then, we investigate the performance of both methods in Section IV, and conclude the paper in Section V.

III. Approach

In this section we present two different approaches: The Two Stage Approach (2SA) solves Problem 1, and the Adversarial Generalized Likelihood Ratio Test (A-GLRT) solves Problem 2. ***Proofs are delegated to our extended technical report in [39].***

A. Two Stage Approach Algorithm

We present an intuitive approach that separates the detection scheme into two stages where 1) a decision is made about the trustworthiness of each individual robot i based on the received value α_i, and then 2) only the measurements Y_i from robots that are trusted are used to detect $\mathcal{H}_0$ or $\mathcal{H}_1$.

a) Detection of Trustworthy Robots: We utilize the Likelihood Ratio Test (LRT) to detect *legitimate* robots. This test is guaranteed to have minimal missed detection probability (i.e., detecting a legitimate robot as malicious) for a given false alarm probability (i.e., detecting a malicious robot as legitimate) [35, Chapter 3].

The FC decides which robots to trust using the LRT:

$$\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)} \underset{\hat{t}_i=0}{\overset{\hat{t}_i=1}{\gtrless}} \gamma_t, \tag{5}$$

where γ_t is a threshold value that we wish to optimize.

The FC decides who to trust and stores it in the vector $\hat{\mathbf{t}}$, where $\hat{t}_i=1$ if the FC decides to trust robot i, and $\hat{t}_i=0$ otherwise. In the case of equality, a random decision is made where the FC chooses $\hat{t}_i=1$ with probability p_t and chooses $\hat{t}_i=0$ otherwise, where p_t is another parameter to be optimized. This leads to the following trust probability, where $P_{\text{trust}}(\gamma_t,p_t,\tilde{t}=1)$ is the probability of trusting a legitimate robot and $P_{\text{trust}}(\gamma_t,p_t,\tilde{t}=0)$ is the probability of trusting a malicious robot:

$$\begin{aligned} P_{\text{trust}}(\gamma_t,p_t,\tilde{t}) = &\Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}>\gamma_t \middle| t_i=\tilde{t}\right) \\ &+p_t\Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}=\gamma_t \middle| t_i=\tilde{t}\right). \end{aligned} \tag{6}$$

b) Detecting the Event Ξ: To determine a hypothesis $\mathcal{H}$ on the event Ξ, the FC only considers measurements from trusted robots, i.e., $i:\hat{t}_i=1$, and uses the following decision rule:

$$\frac{\prod_{\{i:\hat{t}_i=1\}}P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}}{\prod_{\{i:\hat{t}_i=1\}}(1-P_{\text{FA,L}})^{1-y_i}P_{\text{FA,L}}^{y_i}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \exp(\gamma_{\text{TS}}), \tag{7}$$

where $\exp(\gamma_{\text{TS}}) \triangleq \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$. This decision rule is commonly used in standard binary hypothesis testing problems, and it is known to be optimal in a system where no malicious robots are present, i.e., $\mathcal{M}=\emptyset$. Thus, we attempt to approximate this scenario by removing information from all robots deemed to be malicious. However, since there may be errors in classifying the trustworthiness of robots, the parameters γ_t and p_t should balance the need to exclude measurements from malicious robots with the need to allow measurements from legitimate robots in (7). We show how to optimize the threshold γ_t and tie-break probability p_t by computing the probability of error of the FC using the 2SA.

Let $w_{1,\text{L}}=\log(\frac{1-P_{\text{MD,L}}}{P_{\text{FA,L}}})$, $w_{0,\text{L}}=\log(\frac{1-P_{\text{FA,L}}}{P_{\text{MD,L}}})$, and denote $S_{\text{N}} \triangleq \sum_{i=1}^{N}\hat{t}_i[w_{1,\text{L}}y_i - w_{0,\text{L}}(1-y_i)]$ corresponding to the

logarithm of the left-hand side of (7). After the FC discards robot measurements that it does not trust, the decision rule (7) leads to the following missed detection error probability at the FC,

$$P_{\text{MD}}(\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}})=\Pr\Big(S_N<\gamma_{\text{TS}}\big|\mathcal{H}_1,\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}}\Big). \quad (8)$$

A similar expression can be found for the false alarm probability. Consequently, the overall error probability at the FC is:

$$P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}})= \quad (9)$$
$$\Pr(\Xi=0)P_{\text{FA}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}})+\Pr(\Xi=1)P_{\text{MD}}(\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}}).$$

We seek to minimize the probability of error (9) for the decision rule (7) by minimizing the false alarm and missed detection probabilities. Since the error probability must be calculated for every possible combination of vectors $\hat{\mathbf{t}}$ and $\mathbf{y}$ during the minimization of (7), the computation has exponential complexity with respect to the number of robots, N. Furthermore, the true trust vector $\mathbf{t}$ and the probabilities of false alarm $P_{\text{FA,M}}$ and missed detection $P_{\text{MD,M}}$ of the malicious robots are unknown, and therefore, they cannot be used in minimizing (9).

To this end, we derive analytical guarantees regarding the error probability of the overall detection performance of the 2SA as follows. We find the worst-case probability of error of the FC by considering all the possible trust vectors $\mathbf{t}\in\{0,1\}^N$ and false alarm and missed detection probabilities $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$, respectively, in the interval $[0,1]$, and choosing the $\mathbf{t}$, $P_{\text{FA,M}}$, and $P_{\text{MD,M}}$ that maximize (9). Then, we minimize this worst-case error probability by choosing the best threshold γ_t, i.e., choose $\gamma_t=\gamma_t^*$ and tie-break probability $p_t=p_t^*$ where

$$(\gamma_t^*,p_t^*)=\underset{\gamma_t,p_t}{\text{argmin}}\ \max_{\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}} P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}). \quad (10)$$

However, we must first determine the $P_{\text{FA,M}},P_{\text{MD,M}},\mathbf{t}$ that maximize P_{e}. In the remainder of this section, we assume that the proportion of malicious robots in the network is upper bound by $(\bar{m})$.

Lemma 1: If $P_{FA,L}<0.5$ and $P_{MD,L}<0.5$, then the probability of false alarm and missed detection of the FC (8) is maximized for the 2SA when malicious robots choose $P_{\text{FA,M}}=P_{\text{MD,M}}=1$, for any vector $\boldsymbol{t}\in\{0,1\}^N$.

Lemma 2: Let $\bar{\mathbf{t}}$ be the worst-case vector $\mathbf{t}$, i.e., the vector $\mathbf{t}$ that maximizes the probability of error (9). If $P_{FA,L}<0.5$, $P_{MD,L}<0.5$, and $P_{FA,M}=P_{MD,M}=1$, then the probability of error $P_{\text{e}}(\gamma_t,p_t,\bar{\mathbf{t}},1,1)$ is maximized when $\bar{\mathbf{t}}$ contains the maximum number of malicious robots, i.e., $\sum_{i\in\mathcal{N}}\bar{t}_i=N-\bar{m}N$.

The proofs of Lemmas 1-2 hinge on the fact that after information is discarded in the 2SA, the remaining information is trusted as legitimate, so any trusted malicious robots can maximize the probability of error by always sending information that contradicts the true hypothesis.

Utilizing Lemma 2, we calculate the exact probability of error for the FC for the worst-case attack where there are $\overline{m}N$ malicious robots and $P_{\text{FA,M}}=P_{\text{MD,M}}=1$. In order to compute the probability of error exactly, we must compute the probability of false alarm and missed detection using (8). Let $k_{\text{L}}\in K_{\text{L}}$ be the number of legitimate robots trusted by the FC, where $K_{\text{L}}=\{0,...,(1-\bar{m})N\}$, and $k_{\text{M}}\in K_{\text{M}}$ be the number of malicious robots trusted by the FC, where $K_{\text{M}}=\{0,...,\bar{m}N\}$.

Using the law of total probability, the probability of missed detection at the FC is given by

$$P_{\text{MD}}(\gamma_t,p_t,\overline{m},1)=\sum_{k_{\text{L}}\in K_{\text{L}},k_{\text{M}}\in K_{\text{M}}}\Pr(K_{\text{L}}=k_{\text{L}})\Pr(K_{\text{M}}=k_{\text{M}})$$
$$\cdot\Pr(S_{\text{N}}<\gamma_{\text{TS}}|\mathcal{H}_1,k_{\text{L}},k_{\text{M}}). \quad (11)$$

Hereafter, we denote $f_{\text{b}}(x;p,n)=\binom{n}{x}p^x(1-p)^{n-x}$ and $F_{\text{b}}(x;p,n)=\sum_{i=0}^{x}\binom{n}{i}p^i(1-p)^{n-i}$. These are the Binomial PDF and CDF, evaluated at x for n trials and success probability p.

Due to the assumption of homogeneity among legitimate robots and similarly among malicious robots, we can show for any particular instantiation of k_{L} and k_{M} that:

$$\Pr(S_{\text{N}}<\gamma_{\text{TS}}|\mathcal{H}_1,k_{\text{L}},k_{\text{M}})$$
$$=F_{\text{b}}\Big(\lceil\frac{\gamma_{\text{TS}}+k_{\text{M}}w_{1,\text{L}}+k_{\text{L}}w_{0,\text{L}}}{w_{0,\text{L}}+w_{1,\text{L}}}\rceil-1;1-P_{\text{MD,L}},k_{\text{L}}\Big). \quad (12)$$

Recall (6), then we have that

$$\Pr(K_{\text{L}}=k_{\text{L}})=f_{\text{b}}(k_{\text{L}};P_{\text{trust}}(\gamma_t,p_t,\tilde{t}=1),(1-\bar{m})N),$$
$$\Pr(K_{\text{M}}=k_{\text{M}})=f_{\text{b}}(k_{\text{M}};P_{\text{trust}}(\gamma_t,p_t,\tilde{t}=0),\bar{m}N). \quad (13)$$

We calculate $P_{\text{FA}}(\gamma_t,p_t,\overline{m},1)$ by plugging-in (12) and (13) in (11). We can follow similar arguments for the false alarm probability:

$$P_{\text{FA}}(\gamma_t,p_t,\overline{m},1)=\sum_{k_{\text{L}}\in K_{\text{L}},k_{\text{M}}\in K_{\text{M}}}\Pr(K_{\text{L}}=k_{\text{L}})\Pr(K_{\text{M}}=k_{\text{M}})$$
$$\cdot\Pr(S_{\text{N}}\geq\gamma_{\text{TS}}|\mathcal{H}_0,k_{\text{L}},k_{\text{M}}). \quad (14)$$

Therefore, we define the following total error probability in the worst-case

$$\overline{P}_{\text{e}}(\gamma_t,p_t,\overline{m},1,1)\triangleq\Pr(\Xi=0)P_{\text{FA}}(\gamma_t,p_t,\overline{m},P_{\text{FA,M}}=1)$$
$$+\Pr(\Xi=1)P_{\text{MD}}(\gamma_t,p_t,\overline{m},P_{\text{MD,M}}=1), \quad (15)$$

and we can choose the thresholds γ_t and p_t that minimize this expression. Once we choose γ_t and p_t, the rest of the 2SA becomes a standard binary hypothesis testing problem.

Lemma 3: Denote $\Gamma_t:=\{\frac{p_\alpha(a|t_i=1)}{p_\alpha(a|t_i=0)}\}_{a\in\mathcal{A}}$, where $\{\cdot\}_{a\in\mathcal{A}}$ represents a set corresponding to all possible values of $a\in\mathcal{A}$ and the set $\mathcal{A}$ follows Assumption 1. Then, the minimal value of (10) with respect to γ_t can be achieved by $\gamma_t\in\Gamma_t$.

Let $\Gamma_p:=\{0,\delta_p,2\delta_p,\dots,1\}$ with discretization constant δ_p. Algorithm 1 explains the 2SA step-by-step. Determining the threshold value γ_t and tie-break probability p_t requires computing the probability of error $|\Gamma_t|\cdot|\Gamma_p|$ times, where $|\cdot|$ represents the cardinality of the set. However, this can be computed offline, and

Algorithm 1 Two Stage Approach

Input: $\mathbf{y}$, $\mathbf{a}$, $P_{\text{FA,L}}$, $P_{\text{MD,L}}$, $\{\Pr(\Xi)\}_{\Xi=0,1}$, $\{p_\alpha(a_i|t_i)\}_{t_i=0,1}$, $\overline{m}$, Γ_t, δ_p

Output: Decision $\mathcal{H}_0$ or $\mathcal{H}_1$

1: Set $\Gamma_p=\{0,\delta_p,2\delta_p,...,1\}$.
2: **for all** $\hat{\gamma}_t\in\Gamma_t$, $\hat{p}_t\in\Gamma_p$ **do**
3: Compute $P_{\text{trust}}(\hat{\gamma}_t,\hat{p}_t,\tilde{t})$ for $\tilde{t}=0$, $\tilde{t}=1$ by (6).
4: Compute $P_{\text{MD}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1)$ by (11).
5: Compute $P_{\text{FA}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1)$ by (14).
6: Compute $\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)$ by (15).
7: Set $(\gamma_t,p_t)=\text{argmin}_{\hat{\gamma}_t\in\Gamma_t,\hat{p}_t\in\Gamma_p}\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)$.
8: Determine the vector $\hat{\mathbf{t}}$ using (5).
9: Determine decision using (7).

then the returned γ_t and p_t can be used for conducting the hypothesis test. With a given γ_t and p_t, the hypothesis test requires $\mathcal{O}(N)$ comparisons. We utilize Lemmas 1-3 to establish the following.

Theorem 1: Assume that the FC uses the decision rule in (5) to detect malicious robots, and then uses the decision rule (7). Then Algorithm 1 chooses the threshold value γ_t and tie-break probability p_t that minimize the worst-case probability of error of the FC (10) up to a discretization distance $d(\delta_p) \triangleq \min_{p_t \in \Gamma_p} \overline{P}_{\mathrm{e}}(\gamma_t^*, p_t, \overline{m}, 1, 1) - \overline{P}_{\mathrm{e}}(\gamma_t^*, p_t^*, \overline{m}, 1, 1)$. Furthermore, $d(\delta_p) \to 0$ as $\delta_p \to 0$.

B. A-GLRT Algorithm

Here, we construct an efficient algorithm that implements the GLRT in (4) and solves Problem 2 in Section II. Under Assumption 1 we reformulate (4) with the following:

$$\frac{\max\limits_{\boldsymbol{t}\in\{0,1\}^N, P_{\mathrm{MD,M}}\in[0,1]} \Pr(\boldsymbol{a}|\boldsymbol{t})\Pr(\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\mathrm{MD,M}})}{\max\limits_{\boldsymbol{t}\in\{0,1\}^N, P_{\mathrm{FA,M}}\in[0,1]} \Pr(\boldsymbol{a}|\boldsymbol{t})\Pr(\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\mathrm{FA,M}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{\mathrm{GLRT}}. \quad (16)$$

Due to the symmetry in the calculation of the numerator and denominator in (16), we focus our discussion on the numerator. Let

$$c_{\mathrm{L},i} \triangleq p_\alpha(a_i|t_i=1)P_{\mathrm{MD,L}}^{1-y_i}(1-P_{\mathrm{MD,L}})^{y_i}, \text{ and}$$
$$c_{\mathrm{M},i}(P_{\mathrm{MD,M}}) \triangleq p_\alpha(a_i|t_i=0)P_{\mathrm{MD,M}}^{1-y_i}(1-P_{\mathrm{MD,M}})^{y_i}, \quad (17)$$

and $C(\boldsymbol{t}, P_{\mathrm{MD,M}}) \triangleq \prod_{i=1}^N \left(c_{\mathrm{L},i}^{t_i} \cdot [c_{\mathrm{M},i}(P_{\mathrm{MD,M}})]^{1-t_i}\right)$. Applying Assumption 1 and the i.i.d. assumption about measurements, we reformulate the numerator in (16) as follows:

$$\max_{\boldsymbol{t}\in\{0,1\}^N, P_{\mathrm{MD,M}}\in[0,1]} C(\boldsymbol{t}, P_{\mathrm{MD,M}}). \quad (18)$$

Since there is no clear way to optimize (18) over variables $\boldsymbol{t}$ and $P_{\mathrm{MD,M}}$ at the same time, we reformulate the problem as two nested optimizations using the Principle of Iterated Suprema [40, p. 515]. We rewrite (18) as follows:

$$\max_{\boldsymbol{t}\in\{0,1\}^N} \left\{ \max_{P_{\mathrm{MD,M}}\in[0,1]} C(\boldsymbol{t}, P_{\mathrm{MD,M}}) \right\}.$$

With this formulation, one possible way to calculate the maximization is iterating over all vectors $\boldsymbol{t}$ in the set $\{0,1\}^N$; then for each $\boldsymbol{t}$, calculating the inner maximization. We utilize the well-known estimation problem [41, Problem 7.8] to calculate the inner maximization as follows.

Lemma 4: Let $\boldsymbol{t}$ and $\boldsymbol{y}$ be given vectors in $\{0,1\}^N$. Assume that $p_\alpha(a_i|t_i)$ is known for both $t_i = 0$ and $t_i = 1$, and that $\sum_{i:t_i=0} 1 > 0$. Then, $C(\boldsymbol{t}, P_{\mathrm{MD,M}})$ is maximized by $\widehat{P}_{MD,M} = \frac{\sum_{i:t_i=0}(1-y_i)}{\sum_{i:t_i=0} 1}$. Additionally, if $\sum_{i:t_i=0} 1 = 0$, any choice $\widehat{P}_{MD,M} \in [0,1]$ maximizes $C(\boldsymbol{t}, P_{MD,M})$.

Thus, the optimum value of $P_{\mathrm{MD,M}}$ is of a special structure, which we can exploit to avoid an exponential complexity incurred by the 2^N possible values of the trust vector $\mathbf{t} \in \{0,1\}^N$. We next find an efficient algorithm using another formulation of (18) that is obtained by the Principle of Iterated Supremum

$$\max_{P_{\mathrm{MD,M}}\in[0,1]} \left\{ \max_{\boldsymbol{t}\in\{0,1\}^N} C(\boldsymbol{t}, P_{\mathrm{MD,M}}) \right\}. \quad (19)$$

The following shows how to calculate the inner maximization.

Lemma 5: Let $P_{MD,M}$, $\boldsymbol{a}$, and $\mathbf{y}$ be given. Additionally, assume that $p_\alpha(a_i|t_i)$ is known for both $t_i = 0$ and $t_i = 1$. If the estimated robot identity vector $\hat{\boldsymbol{t}}$ is constructed by choosing $\hat{t}_i = 1$ if $c_{L,i} \geq c_{M,i}(P_{MD,M})$ and $\hat{t}_i = 0$ otherwise, where $\hat{t}_i$ is the i^{th} component of $\hat{\boldsymbol{t}}$, then, $\hat{\boldsymbol{t}}$ is a vector that maximizes $C(\boldsymbol{t}, P_{MD,M})$.

By Lemma 5, $C(\boldsymbol{t}, P_{\mathrm{MD,M}})$ can be maximized by comparing two likelihoods for each robot, resulting in $\mathcal{O}(N)$ comparisons in total. We combine Lemma 4 and Lemma 5 to introduce an efficient calculation of the numerator of the GLRT (18). By exploiting the special structure that $\widehat{P}_{MD,M}$ has (Lemma 4), we can restrict the search space for $P_{\mathrm{MD,M}}$ in (19). Then, the inner maximization can be calculated using Lemma 5. The following theorem builds on this intuition to provide an efficient calculation of (18).

Theorem 2: Assume that $(\boldsymbol{t}^*, P_{\mathrm{MD,M}}^*)$ attains the maximization in (18). Then, for each vector of measurements $\mathbf{y}$ and trust values $\mathbf{a}$, $P_{\mathrm{MD,M}}^*$ belongs to the set $\mathcal{P}$ where $\mathcal{P} \triangleq \left\{ \frac{T_n}{T_d} \right\}_{T_n \in \{0,\ldots,T_d\}, T_d \in \{1,\ldots,N\}}$. Moreover, the maximization in (18) can be calculated by iterating over $\mathcal{O}(N^2)$ different values in $\mathcal{P}$ and performing $\mathcal{O}(N)$ comparisons.

Using Thm. 2, we present the A-GLRT algorithm (Alg. 2), which calculates the GLRT (16) using only $\mathcal{O}(N^3)$ comparisons.

Algorithm 2 A-GLRT

Input: $\mathbf{y}$, $\mathbf{a}$, $P_{\mathrm{FA,L}}$, $P_{\mathrm{MD,L}}$, $\{\Pr(\Xi)\}_{\Xi=0,1}$, $\{p_\alpha(a_i|t_i)\}_{t_i=0,1}$, N
Output: Decision $\mathcal{H}_0$ or $\mathcal{H}_1$

Set $\mathcal{P} = \left\{ \frac{T_n}{T_d} \right\}_{T_n \in \{0,\ldots,T_d\}, T_d \in \{1,\ldots,N\}}$.
Set $\gamma_{\mathrm{GLRT}} = \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$, $l_{\mathrm{num,max}} = 0$, $l_{\mathrm{denom,max}} = 0$.
for all $P_M \in \mathcal{P}$ **do**
    Set $P_{\mathrm{MD,M}} = P_M$, $l_{\mathrm{num}} = 1$.
    for i=0 to N **do**
        Set $l_{\mathrm{num}} = l_{\mathrm{num}} \cdot \max\{c_{\mathrm{L},i}, c_{\mathrm{M},i}(P_{\mathrm{MD,M}})\}$.
    if $l_{\mathrm{num}} > l_{\mathrm{num,max}}$ **then** Set $l_{\mathrm{num,max}} = l_{\mathrm{num}}$.
    Repeat steps 4-7 for the denominator.
if $\frac{l_{\mathrm{num,max}}}{l_{\mathrm{denom,max}}} > \gamma_{\mathrm{GLRT}}$ **then** Return decision $\mathcal{H}_1$
else Return decision $\mathcal{H}_0$

IV. Hardware Experiment and Numerical Results

We perform a hardware experiment with robotic vehicles driving on a mock-up road network where robots are tasked with reporting the traffic condition of their road segment to a FC. The objective of the malicious robots is to cause the FC to incorrectly perceive the traffic conditions (see Fig. 2). A numerical study further demonstrates the performance of this scenario with an increasing proportion of malicious robots.

We compare the performance of the 2SA and A-GLRT against several benchmarks including the ***Oracle***, where the FC knows the true trust vector $\mathbf{t}$ and discards malicious measurements, (this serves as a lower bound on the probability of error), the ***Oblivious FC***, where the FC treats every robot as legitimate, and a ***Baseline Approach*** [26] where the FC uses a history of T measurements to develop a reputation about each robot. The Baseline method ignores information from robots whose measurements disagree with the final decision at least $\eta < T$ times. The *Oracle*, *Oblivious FC*, and *Baseline Approach* use the decision rule in (7). Malicious robots perform a Sybil attack where they spoof additional robots into the network. We use the opensource toolbox in [42] to obtain trust values from communicated WiFi signals by analyzing

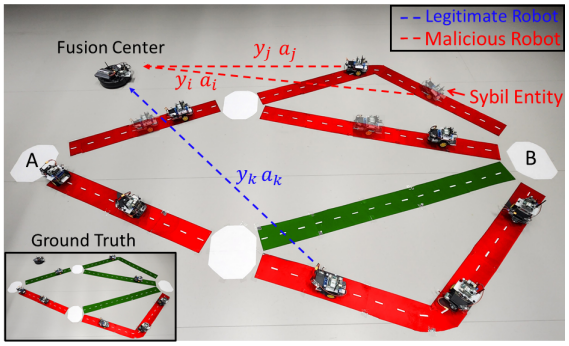

Fig. 2: Robots drive along six road segments to get from point A to point B. While traversing the roadmap, robots estimate the congestion on their current segment as either containing traffic (red) or not (green), and relay their estimates to the FC.

the similarity between different fingerprints to detect spoofed transmissions. The works [30]–[33] model these trust values $\alpha_i \in [0,1]$ as a continuous random variable. We discretize the sample space by letting $\mathcal{A} = \{0,1\}$ and setting $a_i = 1$ if the measured trust value is ≥ 0.5 and $a_i = 0$ otherwise.

*a) **Hardware Experiment**:* A group of $N = 11$ mobile robots drive in a loop from a starting point A to point B, approximately 4.5 meters apart, by randomly choosing to traverse one of four possible paths made up of six different road segments. As the robots drive between points A and B they are given noisy position information for themselves and neighboring robots from an OptiTrack motion capture system with added white Gaussian noise with a variance of $1m^2$. This serves as a proxy for GPS-reported measures used in crowdsourcing traffic estimation schemes like Waze, Google Maps, and others. A road segment is considered to have traffic ($y_i = 1$) if the number of robots on the segment is ≥ 2. Of the 11 robots in the group, 5 robots are legitimate, 3 are malicious, and 3 are spoofed by the malicious robots (making them also malicious). This leads to scenarios where hypothesis tests are performed with only legitimate robots, only malicious robots, or a combination of both, depending on where each robot is in the roadmap. Malicious robots know the true traffic conditions and report the wrong measurement with probability 0.99, i.e., $P_{\text{FA,M}} = P_{\text{MD,M}} = 0.99$. The empirical data from the experiment is stated

Parameters			
$P_{\text{FA,L}}$	0.0800	$P_{\text{MD,L}}$	0.2100
$\Pr(\Xi=0)$	0.6432	$\Pr(\Xi=1)$	0.3568
Percent Error			
2SA (Sec. III-A)	**30.5 %**	**A-GLRT (Sec. III-B)**	**29.0 %**
Oracle	19.5 %	Oblivious FC	52.0 %
Baseline1	50.8 %	Baseline10	48.7 %

TABLE I: EXPERIMENTAL RESULTS

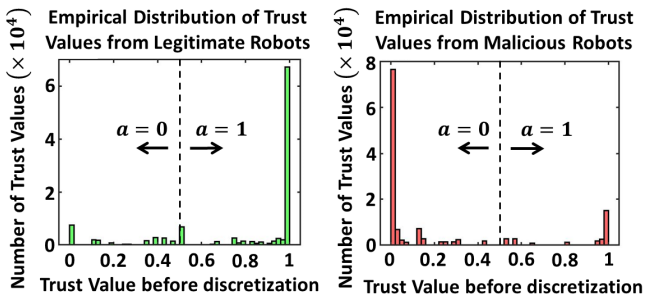

Fig. 3: Empirical distribution of the trust values gathered during the hardware experiment for legitimate and malicious robots. The trust value is thresholded to $a = 1$ if it is ≥ 0.5, and $a = 0$ otherwise.

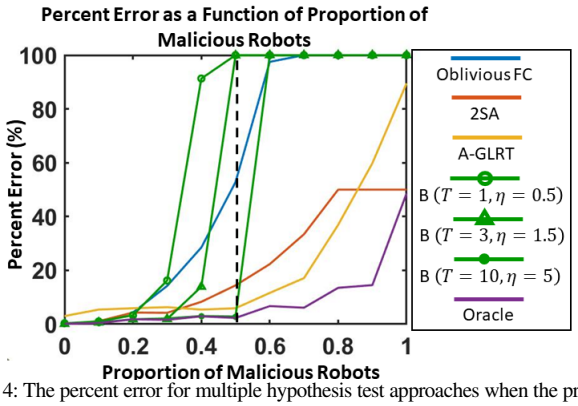

Fig. 4: The percent error for multiple hypothesis test approaches when the proportion of malicious robots is varied. The 2SA and A-GLRT outperform the Oblivious FC and Baseline Approaches (B) when the majority of the network is malicious (right side of the dashed line). The performance of the Oracle declines as the proportion of malicious robots increases since it has less legitimate information to leverage.

in Table I, where *Baseline1* and *Baseline10* refer to the Baseline Approach from [26] with parameters T and η set to $(T=1,\eta=0.5)$ and $(T=10,\ \eta=5)$. We determined the parameters in Table I by first running an experiment without performing hypothesis tests and observing the behavior of the system compared to ground truth. The trust values gathered using the toolbox in [42] led to the empirical probabilities $p_\alpha(a_i = 1|t_i = 1) = 0.8350$ and $p_\alpha(a_i = 1|t_i = 0) = 0.1691$ (see Fig. 3). The entire experiment was run for 15 minutes with a frequency of 30 hypothesis tests on each road segment per second. A test was only run if at least one robot was present in the segment. This led to a total of 61233 hypothesis tests carried out. Of the 61233 tests, 29.9% consisted of only legitimate robots, 28.1% of only malicious robots, and 42.0% contained both legitimate and malicious robots.

In our hardware experiment the 2SA and A-GLRT outperform the Oblivious FC and the Baseline Approach. The Baseline Approach exhibits a high percent error due to the fact that it relies on the majority of the network being legitimate. This points to a common vulnerability of reputation based approaches that assume only a small proportion of the network is malicious.

*b) **Numerical Study**:* Next, we perform a numerical study on the performance of each approach when the proportion of malicious robots is varied. In the numerical study we use $N = 10$ robots with $\Pr(\Xi=0) = \Pr(\Xi=1) = 0.5$, $P_{\text{FA,L}} = P_{\text{MD,L}} = 0.15$, and $P_{\text{FA,M}} = P_{\text{MD,M}} = 0.99$ and perform hypothesis tests over 1000 trials for each proportion of malicious robots. In the simulation study the trust value distributions are fixed at $p_\alpha(a_i = 1|t_i = 1) = 0.8$, $p_\alpha(a_i = 1|t_i = 0) = 0.2$, and the proportion of malicious robots varies from 0 to 1. The results of the simulation study are plotted in Fig. 4. From the plot it can be seen that the 2SA and the A-GLRT perform well even after the number of malicious robots exceeds majority since they use additional trust information independent of the data.

V. CONCLUSION

In this paper we present two methods to utilize trust values in solving the binary adversarial hypothesis testing problem. The 2SA uses the trust values to determine which robots to trust, and then makes a decision from the measurements of the trusted robots. The A-GLRT jointly uses the trust values and measurements to estimate the trustworthiness of each robot, the strategy of malicious robots, and the true hypothesis.

REFERENCES

[1] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Asymptotic analysis of distributed bayesian detection with byzantine data," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 608–612, 2014.

[2] X. Ren, J. Yan, and Y. Mo, "Binary hypothesis testing with byzantine sensors: Fundamental tradeoff between security and efficiency," *IEEE Transactions on Signal Processing*, vol. 66, no. 6, pp. 1454–1468, 2018.

[3] S. Althunibat, A. Antonopoulos, E. Kartsakli, F. Granelli, and C. Verikoukis, "Countering intelligent-dependent malicious nodes in target detection wireless sensor networks," *IEEE Sensors Journal*, vol. 16, no. 23, pp. 8627–8639, 2016.

[4] J. Wu, T. Song, Y. Yu, C. Wang, and J. Hu, "Generalized byzantine attack and defense in cooperative spectrum sensing for cognitive radio networks," *IEEE Access*, vol. 6, pp. 53 272–53 286, 2018.

[5] A. Pierson and M. Schwager, "Adaptive inter-robot trust for robust multi-robot sensor coverage," in *In International Symposium on Robotics Research*, 2013.

[6] Y. Xu, G. Deng, T. Zhang, H. Qiu, and Y. Bao, "Novel denial-of-service attacks against cloud-based multi-robot systems," *Information Sciences*, vol. 576, pp. 329–344, 2021.

[7] J. Song and S. Gupta, "Care: Cooperative autonomy for resilience and efficiency of robot teams for complete coverage of unknown environments under robot failures," *Autonomous Robots*, vol. 44, no. 3, pp. 647–671, 2020.

[8] S. Sariel-Talay, T. R. Balch, and N. Erdogan, "Multiple traveling robot problem: A solution based on dynamic task selection and robust execution," *IEEE/ASME TRANSACTIONS ON MECHATRONICS*, vol. 14, no. 2, 2009.

[9] B. Schlotfeldt, V. Tzoumas, D. Thakur, and G. J. Pappas, "Resilient active information gathering with mobile robots," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4309–4316.

[10] R. K. Ramachandran, N. Fronda, and G. S. Sukhatme, "Resilience in multi-robot target tracking through reconfiguration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4551–4557.

[11] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Resilient distributed state estimation with mobile agents: overcoming byzantine adversaries, communication losses, and intermittent measurements," *Autonomous Robots*, vol. 43, no. 3, pp. 743–768, 2019.

[12] A. Laszka, Y. Vorobeychik, and X. Koutsoukos, "Resilient observation selection in adversarial settings," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 7416–7421.

[13] J. Blumenkamp and A. Prorok, "The emergence of adversarial communication in multi-agent reinforcement learning," in *Conference on Robot Learning*. PMLR, 2021, pp. 1394–1414.

[14] R. Mitchell, J. Blumenkamp, and A. Prorok, "Gaussian process based message filtering for robust multi-agent cooperation in the presence of adversarial communication," *arXiv preprint arXiv:2012.00508*, 2020.

[15] G. Deng, Y. Zhou, Y. Xu, T. Zhang, and Y. Liu, "An investigation of byzantine threats in multi-robot systems," in *24th International Symposium on Research in Attacks, Intrusions and Defenses*, 2021, pp. 17–32.

[16] R. Wehbe and R. K. Williams, "Probabilistically resilient multi-robot informative path planning," *arXiv preprint arXiv:2206.11789*, 2022.

[17] N. Petrovska and A. Stevanovic, "Traffic congestion analysis visualisation tool," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. IEEE, 2015, pp. 1489–1494.

[18] T. Jeske, "Floating car data from smartphones: What google and waze know about you and how hackers can control traffic," *Proc. of the BlackHat Europe*, pp. 1–12, 2013.

[19] G. Wang, B. Wang, T. Wang, A. Nika, H. Zheng, and B. Y. Zhao, "Ghost riders: Sybil attacks on crowdsourced mobile mapping services," *IEEE/ACM transactions on networking*, vol. 26, no. 3, pp. 1123–1136, 2018.

[20] Y. S. Sandal, A. E. Pusane, G. K. Kurt, and F. Benedetto, "Reputation based attacker identification policy for multi-access edge computing in internet of things," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15 346–15 356, 2020.

[21] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, 2008.

[22] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed bayesian detection in the presence of byzantine data," *IEEE transactions on signal processing*, vol. 63, no. 19, pp. 5250–5263, 2015.

[23] R. Chen, J.-M. Park, and K. Bian, "Robust distributed spectrum sensing in cognitive radio networks," in *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*. IEEE, 2008, pp. 1876–1884.

[24] E. Nurellari, D. McLernon, and M. Ghogho, "A secure optimum distributed detection scheme in under-attack wireless sensor networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 325–337, 2017.

[25] E. Nurellari, D. McLernon, M. Ghogho, and S. Aldalahmeh, "Distributed binary event detection under data-falsification and energy-bandwidth limitation," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6298–6309, 2016.

[26] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2010.

[27] R. Liu, F. Jia, W. Luo, M. Chandarana, C. Nam, M. Lewis, and K. Sycara, "Trust-aware behavior reflection for robot swarm self-healing," *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, p. 122–130, 2019.

[28] V. Renganathan and T. Summers, "Spoof resilient coordination for distributed multi-robot systems," *2017 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*, pp. 135–141, Dec 2017.

[29] J. Xiong and K. Jamieson, "Securearray: Improving wifi security with fine-grained physical-layer information," *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, p. 441–452, 2013.

[30] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, "Guaranteeing spoof-resilient multi-robot networks," *AuRo*, p. 1383–1400, 2017.

[31] F. Mallmann-Trenn, M. Cavorsi, and S. Gil, "Crowd vetting: Rejecting adversaries via collaboration with application to multirobot flocking," *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 5–24, 2022.

[32] M. Yemini, A. Nedić, A. J. Goldsmith, and S. Gil, "Characterizing trust and resilience in distributed consensus for cyberphysical systems," *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 71–91, 2021.

[33] M. Yemini, A. Nedić, S. Gil, and A. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," in *Conference on Decision and Control (CDC)*, 2022.

[34] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, "Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 1, pp. 205–215, 2012.

[35] S. M. Kay, *Fundamentals of statistical signal processing: Detection theory*. Prentice Hall PTR, 2008.

[36] Z. Sun, C. Zhang, and P. Fan, "Optimal byzantine attack and byzantine identification in distributed sensor networks," in *2016 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2016, pp. 1–6.

[37] M. Cheng, C. Yin, J. Zhang, S. Nazarian, J. Deshmukh, and P. Bogdan, "A general trust framework for multi-agent systems," in *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, 2021, pp. 332–340.

[38] M. Peng, Z. Xu, S. Pan, R. Li, and T. Mao, "Agenttms: A mas trust model based on agent social relationship." *J. Comput.*, vol. 7, no. 6, pp. 1535–1542, 2012.

[39] M. Cavorsi, O. E. Akgun, M. Yemini, A. Goldsmith, and S. Gil, "Exploiting trust for resilient hypothesis testing with malicious robots," *ArXiv*, 2022.

[40] J. M. H. Olmsted, *Real variables: An introduction to the theory of functions*. Appleton-Century-Crofts, 1959.

[41] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice-Hall PTR, 1993.

[42] N. Jadhav, W. Wang, D. Zhang, S. Kumar, and S. Gil, "Toolbox release: A wifi-based relative bearing sensor for robotics," *ArXiv*, vol. abs/2109.12205, 2021.