This article has been accepted for publication in IEEE Transactions on Robotics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TRO.2024.3415235

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. For the Two Stage Approach, we assume that 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 simultaneously estimate the trustworthiness of robots, their reporting strategy, and the correct hypothesis. We exploit particular structures in the problem to show that this approach remains computationally tractable even with unknown problem parameters. We deploy both algorithms in a hardware experiment where a group of robots conducts crowdsensing of traffic conditions subject to a Sybil attack on a mock-up road network. We extract the trust observations for each robot from 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 algorithms respectively.**

***Index Terms*—Multi-Robot Systems, Sensor Networks, Networked Robots, Distributed Estimation.**

I. Introduction

Distributed sensing, involving a team of robots that sense their environment and combine their data to discern events of interest, is critical to many robotics applications such as coordinated coverage where robots seek to maximize their ability to sense events of interest [2]–[4], share target information for coordinated tracking [5]–[7], or merge map information to provide a global understanding of the environment [8]–[10]. However, these distributed sensing algorithms are vulnerable to manipulation by malicious attacks, where adversaries aim to skew the system's estimation regarding the event of interest. A notable example is in crowdsensing tasks such as traffic prediction, where a server uses GPS data to infer if a particular roadway is congested or not [11] (see Fig. 1). Unfortunately, this process is susceptible to manipulation by malicious robots [12], [13]. Prior works have shown that a Sybil attack—where an attacker creates and uses fake entities to feed false information to the system— can cause crowdsensing applications such as Google Maps to incorrectly perceive traffic conditions, resulting in erroneous reporting of traffic flows [14], [15].

In this work, we are interested in the problem where robots estimate the presence of an event of interest. Each robot relays its measurement to a *Fusion Center* (FC) that makes an informed binary decision on the occurrence of the event. An unknown subset of the network consists of malicious robots whose goal is to increase the likelihood that the FC makes a wrong decision [13], [16], [17]. 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.

The problem of binary adversarial hypothesis testing has been previously studied within the context of sensor networks [18]–[20]. Many of these approaches use data, such as a history of measurements and hypothesis outcomes, to assess the trustworthiness of the robots [21]–[24]. 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 robots in the network are legitimate. A growing body of work investigates additionally sensed quantities arising from the *physicality of cyberphysical systems* such as multi-robot networks, to cross-validate and assess the trustworthiness of robots [2], [25]–[27]. 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 [27]–[29]. Importantly, this class of trust observations can often be obtained from a one-shot observation, independent of the transmitted measurement. The works in [30]–[32] use trust observations to recover resilient consensus and distributed optimization even in the case where more than half of the network elements are malicious. In this paper *we wish to derive a framework for adversarial hypothesis testing that exploits trust observations to arrive at a similar level of resilience, such that a FC can conceivably reduce its error probability, even in the one-shot scenario and where legitimate robots do not hold a majority in the network.*

To this end, we derive algorithms for achieving resilient

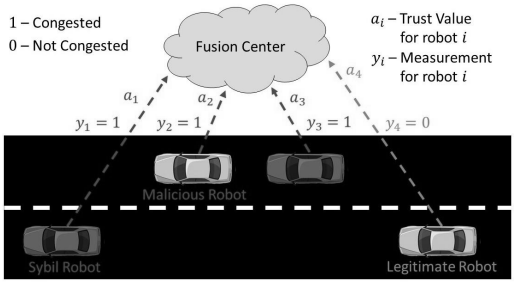

Fig. 1: Malicious robots can perform a Sybil Attack to try to force a FC to incorrectly perceive traffic conditions on a road. The FC can aggregate measurements and trust values from robots to accurately estimate the true traffic condition of the road despite the attack.

(*Co-primary authors). M. Cavorsi, O. E. Akgün, and S. Gil are with the School of Engineering and Applied Sciences, Harvard University, USA. M. Yemini is with the Faculty of Engineering, Bar-Ilan University, Israel. A. J. Goldsmith is with the Department of Electrical and Computer Engineering, Princeton University, USA. The authors gratefully acknowledge partial support through AFOSR grant FA9550-22-1-0223, AFOSR award #002484665 and NSF awards #CNS-2147631 and #CNS-2147694. This paper was accepted in part for presentation at the 2023 IEEE International Conference on Robotics and Automation (ICRA) [1].

hypothesis testing by exploiting stochastic trust observations between the FC and a group of robots participating in event detection. We exploit one-shot trust observations, hereafter called *trust values*, to arrive at *tractable, closed-form solutions* when the majority of the network may be malicious and the strategy of the malicious robots is unknown – a challenging and otherwise intractable problem to solve in general [33].

For the case where an upper limit on the proportion of malicious robots is known, we develop the *Two Stage Approach* (2SA) for hypothesis testing. The first stage of this algorithm uses trust values to determine the most likely set of malicious robots. In the second stage it applies a Likelihood Ratio Test (LRT) only over the trusted robots identified in the first 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) [34]. We show that the addition of trust values allows us to decouple the trustworthiness estimation from the strategy of adversaries, allowing us to calculate the exact MLE of unknown parameters in polynomial time, instead of approximating them as in previous works [33], [35]. Our simulations show that the A-GLRT yields a lower probability of error than the 2SA under a worst-case attack, 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 to validate our results and assess the effectiveness of our methods in a practical setting.

This paper extends the results of its conference version [1] in several ways. For the Two Stage Approach algorithm, additional analysis is provided regarding the probability of error for a fixed proportion of malicious robots as the number of robots in the network increases. In particular, we show that if the probability of trusting a legitimate robot is much higher than the probability of trusting a malicious robot, then the probability of error from using the 2SA will decay at least exponentially as the number of robots in the network increases. Additionally, we investigate the limiting behavior of the 2SA as the proportion of malicious robots increases. We show that if there are too many malicious robots in the network, the 2SA will resort to rejecting all information received, and choose a decision purely based on the prior probability of the event in question occurring. We also characterize the critical proportion of malicious robots that will cause this behavior as a function of the probability of trusting legitimate and malicious robots. For the A-GLRT algorithm, we introduce two different modifications to the algorithm given additional information. One modification utilizes knowledge of prior legitimacy probabilities, i.e., the probability any robot chosen at random will be legitimate or malicious. The other modification is helpful if there is a known upper bound on the number of malicious robots in the network. We investigate the behavior of the A-GLRT as the *quality* of the trust values improves, where a high quality trust value corresponds to a lower probability of misclassifying a legitimate robot as malicious, and vice versa. We show asymptotic results in the observation quality; i.e., as the trust values approach the true legitimacy with probability 1, the A-GLRT converges to the LRT using only the legitimate measurements. Finally, we include all proofs that were excluded from the conference paper due to space.

II. Related Works

The system where a group of sensors detect an event locally and convey their binary measurements to the FC is well-studied in the literature [34], [36]. The LRT minimizes the probability of error in the FC given that the probability of false alarm and missed detection for all sensors as well as the prior probability of the event is known by the FC [34]. However, these distributed sensor networks are known to be susceptible to adversaries, as demonstrated by many previous works [12], [16], [17]. In the presence of adversaries, the assumption of knowing the probability of false alarm and missed detection of all sensors in the network no longer holds. Since the identities of the sensors in the network and the strategy of adversaries are unknown to the FC, the LRT cannot be employed.

The problem of decision making with unknown parameters is known as composite hypothesis testing [37]. A common approach in composite hypothesis testing is to apply the GLRT, which replaces the unknown parameters with their MLEs [34]. The previous works in [33], [35] approach the problem from this perspective by jointly determining the true hypothesis and estimating the unknown parameters in the system. The authors in [33] use an expectation-maximization algorithm to approximate the MLEs of the unknown parameters iteratively. At each iteration, the algorithm estimates sensor identities using the previous false alarm and missed detection probability estimations, then refines its estimation of these probabilities using the new identity estimation. After convergence, the LRT is applied using the estimated parameters. Similarly, the authors in [35] propose a likelihood-based estimation algorithm for determining the identities of the sensors and their corresponding false-alarm and missed detection probabilities. The algorithm fixes all unknowns in the system but one, and then optimize over that free parameter. The proposed algorithm improves the computational complexity over [33], yet it still generates approximations to the MLEs. The A-GLRT algorithm we present is also based on the GLRT. It incorporates the trust observations into the GLRT framework. Moreover, our algorithm finds the exact MLEs, instead of approximating them as was done in previous works.

Another common way to anticipate the effect of adversaries in the network is to try to identify explicitly which robots are malicious. Previous works such as [22], [23] identify malicious sensors using a reputation-based approach. In these approaches the FC compares the information received from each of the sensors with the final decision it arrives at over the course of several hypothesis tests. During this comparison, if the FC notices that certain sensors are consistently sending information that disagrees with the final decision, then those sensors can be flagged as potentially malicious. Other common ways to identify malicious sensors involve leveraging specific communication network structures. For example, the authors in [38] pair the sensors in groups of two. They implement an architecture where each sensor sends its information to the FC and also to the other sensor in its group. Then, each sensor also relays the information from its group member to the FC so the FC can examine the information for inconsistencies. In our work we also look to identify which sensors

are potentially malicious in order to use that information to make a more informed decision. However, most of the related literature identifies malicious sensors by exploiting specific network structures or by referring to previous observations. This either restricts the network architectures that can be used or allows the FC to be susceptible for some time before it can develop a strategy to defend against the attack. We propose a different approach without these shortcomings

In our approach we utilize physical properties of the robot network that may elucidate some additional information as to the trustworthiness of a particular robot, as has been done in previous works. For example, in [39], [40] the robots physically interact with each other and each interaction has an expected outcome. The authors show that robots can determine the trustworthiness of neighboring robots by rating the outcome of their interaction as either successful or unsuccessful. The works in [27]–[29] use physical properties of wireless transmissions to thwart Sybil attacks. They show that by analyzing the wireless profiles from incoming transmissions, certain transmissions can be determined to be malicious if their signal profiles are dishonest or too similar to another, hinting that the robot may be performing a spoofing attack. In all of these methods, it can be shown that the ability to confidently discern trust of a neighboring robot increases as more observations are made; however, useful information from even a single observation can be made. Moreover, the authors in [30] showed that since their method uses physical information that is independent of the information the robots transmit, the system can even handle scenarios where more than half of the robots in the network are malicious. We seek to leverage the benefits of these physical trust observations in order to improve the performance of a FC performing a binary hypothesis test in the presence of adversarial robots.

III. 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 is 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 by $\mathcal{H}_1$ the hypothesis that $\Xi = 1$ and $\mathcal{H}_0$ the hypothesis that $\Xi = 0$. Each robot forwards its local decision to a centralized fusion center (FC).

We consider the scenario where some robots are *malicious* and may manipulate the data that they send to the FC, 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 $\mathcal{L} \cap \mathcal{M} = \emptyset$. Since we consider a one-shot detection, both these sets and the total number of robots N are allowed to vary over time (per hypothesis test), provided our assumptions hold consistently. This flexibility in our model contrasts with common assumptions in methods dependent on historical data, which often necessitate static numbers of robots and unchanging sets of legitimate and malicious robots. 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 we are interested in estimating it.

We assume the following robot behavioral models:

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 legitimate robot $i \in \mathcal{L}$, the measured bit Y_i is subject to noise with the following false alarm and missed detection probabilities*

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

where $P_{\text{FA,L}} \in (0, 0.5)$ and $P_{\text{MD,L}} \in (0, 0.5)$ without loss of generality. We assume that all legitimate robots have the same probability of false alarm and missed detection. Moreover, we assume that the measurement of a legitimate robot is independent of all other robots, and identically distributed given the true hypothesis. Finally, we also assume that $P_{FA,L}$ and $P_{MD,L}$ are known by the FC.

We note that the nonzero probabilities of missed detection and false alarm, P_{MD} and P_{FA} respectively, capture the the realistic assumption for robot systems that legitimate robots have imperfect information due to noisy sensors.

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 a malicious robot $i \in \mathcal{M}$ can flip its measurement with probability $p_f \in [0,1]$ after making an observation, i.e., measures $y_i = 1$ but sends $y_i = 0$, or vice versa, and that all malicious robots flip their bit with the same probability. Let $p_{FA,M}, p_{MD,M} \in [0, 0.5)$ be the probability of false alarm and missed detection of a malicious robot before altering the bit. We assume that all malicious robots have the same probability of false alarm and missed detection. The effective probabilities of false alarm and missed detection of a malicious robot after altering the bit are:*

$$\begin{aligned} P_{\text{FA,M}} &= \Pr(Y_i = 1 | \Xi = 0, t_i = 0) \\ &= (1 - p_{\text{f}}) \cdot p_{\text{FA,M}} + p_{\text{f}} \cdot (1 - p_{\text{FA,M}}), \\ P_{\text{MD,M}} &= \Pr(Y_i = 0 | \Xi = 1, t_i = 0) \\ &= (1 - p_{\text{f}}) \cdot p_{\text{MD,M}} + p_{\text{f}} (1 - p_{\text{MD,M}}). \end{aligned} \tag{2}$$

We assume that a measurement coming from a malicious robot is independent of other measurements given the true hypothesis. Furthermore, we assume that $p_{\text{FA,M}}$, $p_{\text{MD,M}}$, and the strategy of the malicious robots, which is the flipping probability p_{f}, are not known by the FC. This implies that the FC does not know $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$ either.

In our attack model, malicious robots cooperate to set the probability of flipping their bit, p_{f}. Then, attacks are carried out independently where each malicious robot flips its bit independently of the others. This setup allows for initial coordination in determining p_{f}, but avoids the complexities of further collaboration during the attack, which would necessitate additional communication infrastructure and computational resources. This attack model is consistent with common assumptions in the field [12], [17], [19], [20], [33], [41]–[43].

We assume that each measurement 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. We assume that these estimations are obtained as "side information" through

the network's physicality, and are available independently of the robots' measurements.

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 (aka sample space) by $\mathcal{A}$ and denote a realization for robot i by a_i.*

To illustrate how to derive trust values, consider a system under a Sybil attack where robots communicate with the FC using wireless signals. In this scenario, each robot's signal, acting as a unique "spatial fingerprint", is compared with others to generate a similarity or uniqueness score, forming the basis for their trust values. A higher similarity score indicates a lower trust value, as it suggests potential spoofing [28]. We also use this method to obtain trust values in our hardware experiment (see Section VI).

Assumption 1. *We assume that the set $\mathcal{A}$ is finite and that the trust value distributions are homogeneous across all legitimate robots $i\in\mathcal{L}$. To this end, we denote the probability mass function of the trust values of robots by $p_\alpha(a|t)$. We assume the probability mass functions are known or can be estimated by the FC.*[1] *We assume that the trust values are i.i.d. given the true legitimacy of the robot. Moreover, the trust values are assumed to be independent of the measurements, Y_i, and the true hypothesis. Finally, to omit trivial or noninformative cases, we assume that $p_\alpha(a|t=0)\in(0,1)$, $p_\alpha(a|t=1)\in(0,1)$, and $p_\alpha(a|t=0)\neq p_\alpha(a|t=1)$ for all $a\in\mathcal{A}$.*

We do not impose any restrictions over the conditional probability distributions $p_\alpha(a|t=1)$ and $p_\alpha(a|t=0)$. However, for the trust values to be meaningful, they should have different probability mass functions, i.e., $p_\alpha(a|t=1)\neq p_\alpha(a|t=0)$. How distinguishable the two probability mass functions are is termed the *quality* of the trust value, where a better quality corresponds to a larger distinction between the distributions $p_\alpha(a|t=1)$ and $p_\alpha(a|t=0)$. Based on these definitions, we provide the objective of the FC.

A. The objective of the FC

Denote the vector of all measurements with $\boldsymbol{Y}=(Y_1,\ldots,Y_N)$ and its realization $\boldsymbol{y}=(y_1,\ldots,y_N)$, and the vector of stochastic trust values by $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_N)$ and its realization by $\boldsymbol{a}=(a_1,\ldots,a_N)$. 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 $(\boldsymbol{a},\boldsymbol{y})$. Similarly $(\boldsymbol{a},\boldsymbol{y})\in\mathcal{D}_1$ if the FC chooses hypothesis $\mathcal{H}_1$. To simplify our notations we denote $\mathcal{D}:=\{\mathcal{D}_0,\mathcal{D}_1\}$.

Denote by P_{FA} and P_{MD} the false alarm and missed detection probabilities of the rule used by the FC, that is

$$P_{\text{FA}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}}) = \sum_{(\boldsymbol{a},\boldsymbol{y})\in\mathcal{D}_1}\Pr(\boldsymbol{\alpha}=\boldsymbol{a},\boldsymbol{Y}=\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}}), \tag{3}$$

$$P_{\text{MD}}(\mathcal{D},\boldsymbol{t},P_{\text{MD,M}}) = \sum_{(\boldsymbol{a},\boldsymbol{y})\in\mathcal{D}_0}\Pr(\boldsymbol{\alpha}=\boldsymbol{a},\boldsymbol{Y}=\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}}). \tag{4}$$

[1] Example of a trust value α_i: One example of such trust values comes from the works in [28]–[30]. In these works, the trust values $\alpha_i\in[0,1]$ are stochastic and are determined from physical properties of wireless transmissions. We measure and use these trust values in our hardware experiment in Section VI 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 [25], [44], [45].

Note that these probabilities are affected by the strategy of the malicious robots, i.e., $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$.

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

$$\begin{aligned}&P_{\text{e}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}},P_{\text{MD,M}})=\\&\Pr(\Xi=0)P_{\text{FA}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}})+\Pr(\Xi=1)P_{\text{MD}}(\mathcal{D},\boldsymbol{t},P_{\text{MD,M}}).\end{aligned} \tag{5}$$

In this case, the vector of trust values $\boldsymbol{\alpha}$ would not affect the optimal decision rule, and it would only depend on the vector of measurements $\boldsymbol{Y}$. We note that the goal of minimizing the expected error probability in (5) can be easily generalized to minimizing the expected loss, where constants can be added as coefficients to $P_{\text{FA}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}})$ and $P_{\text{MD}}(\mathcal{D},\boldsymbol{t},P_{\text{MD,M}})$ that represent costs corresponding to Type I errors (false alarm) and Type II errors (missed detection). We focus on minimizing the expected error probability to simplify the exposition.

There are two main obstacles to the optimization of the probability of error (5), namely:

1) The FC does not know the identity of the malicious robots, and thus it does not know the correct vector $\boldsymbol{t}$. Therefore, the FC needs to estimate the true trust vector, where the estimated trust vector is denoted by $\hat{\boldsymbol{t}}$.
2) The FC does not know how the malicious robots alter their measurements before sending them. In our setup, this means that the FC does not know the values $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$. Therefore the FC needs to estimate them, where the estimates are denoted by $\hat{P}_{\text{FA,M}}$ and $\hat{P}_{\text{MD,M}}$.

The FC needs to make a decision with these unknown parameters, which is known as the composite hypothesis testing problem. Since the minimization of (5) 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 assuming that the upper bound on the number of malicious robots in the network is known in order to make (5) tractable. Then, we can ignore the measurements from robots deemed to be malicious and choose the decision regions $\mathcal{D}_0$ and $\mathcal{D}_1$ using the measurements from the remaining 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 $\boldsymbol{\alpha}$. Then, the FC makes a decision about the hypothesis using only the vector of measurements $\mathbf{Y}$, from robots it identifies as legitimate. Given an upper bound $\overline{m}\in(0,1)$ 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:*

$$\min_{\mathcal{D}}\max_{P_{\text{FA,M}},P_{\text{MD,M}},\boldsymbol{t}:\sum_{i\in\mathcal{N}}t_i\leq\overline{m}N} P_{\text{e}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}},P_{\text{MD,M}}). \tag{6}$$

The definition in Problem 1 requires an approach that estimates the trustworthiness of a robot i using only the trust value a_i associated with that robot, while assuming a known upper bound on the proportion of malicious robots. 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, which are $\boldsymbol{t}$, $P_{\text{FA,M}}$, and $P_{\text{MD,M}}$ as well as the hypothesis $\mathcal{H}_0$ or $\mathcal{H}_1$ jointly, without requiring any 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 [34], that is

$$\frac{p(\boldsymbol{z};\hat{\theta}_1,\mathcal{H}_1)}{p(\boldsymbol{z};\hat{\theta}_0,\mathcal{H}_0)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)} \triangleq \gamma_{\text{GLRT}}, \tag{7}$$

where $\hat{\theta}_1$ is the maximum likelihood estimator (MLE) of the unknown parameter θ_1 assuming $\Xi=1$ and $\hat{\theta}_0$ is the MLE of θ_0 assuming $\Xi=0$. The operator $\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}$ is interpreted as choosing $\mathcal{H}_1$ when the left-hand side (LHS) of the expression is greater than the right-hand side (RHS), and choosing $\mathcal{H}_0$ when the LHS is less than or equal to the RHS. 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 of the test

$$\frac{\max_{\boldsymbol{t}\in\{0,1\}^N,P_{\text{MD,M}}\in[0,1]}\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})}{\max_{\boldsymbol{t}\in\{0,1\}^N,P_{\text{FA,M}}\in[0,1]}\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{\text{GLRT}}. \tag{8}$$

Note that in this setup 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 (8) 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. This leads us to the formulation in Problem 2.

Problem 2. *Find a computationally tractable algorithm that calculates the GLRT in* (8) *with unknown* $\boldsymbol{t}$*,* $P_{MD,M}$ *and* $P_{FA,M}$*.*

We now put our problem in the context of a running example that we will use throughout for clarity of exposition, and that forms the basis of our hardware experiment in Section VI.

Example 1 (Crowdsourced Traffic Detection)**.** *Consider a traffic detection scenario where robotic agents are driving on a roadway, and a service (the FC), e.g., Waze, Google Maps, etc., aims to estimate the current state of traffic to inform users. The service relies on crowdsourced information from the robots on the road to detect the state of traffic. This is akin to simplifying traffic detection into a binary decision where* N *robots assist the server in determining if there is significant traffic present on a road or not. Each robot senses the traffic locally and transmits* $y_i=1$ *to the server if it believes there is traffic, or* $y_i=0$ *otherwise. The server's decision corresponds to the hypothesis* $\mathcal{H}_0$ *(no traffic) or* $\mathcal{H}_1$ *(traffic), which ideally matches the true state of the event, represented by* $\Xi=0$ *(no traffic) or* $\Xi=1$ *(traffic). We assume the traffic information that robots gather from their sensors can be noisy. This can lead to a legitimate robot detecting that there is traffic when there, in fact, is not, characterized by the probability of false alarm,* $P_{FA,L}$*, or that there is not much traffic when there actually is, characterized by the probability of missed detection,* $P_{MD,L}$*. Malicious robots also estimate the current state of traffic, but may intentionally send the wrong information to the server (FC). The probability that a malicious robot sends the wrong information to the server is characterized by the probabilities* $P_{FA,M}$ *and* $P_{MD,M}$*.*

We emphasize that the derived framework and results of this paper hold for general hypothesis testing problems, and that Example 1 is not meant to limit the scope of the work.

In the next sections we propose two different approaches: one approach to solve Problem 1 and another to solve Problem 2. Then, we investigate the performance of both methods in Section VI, and conclude the paper in Section VII. A table of common notation can be found in Table I.

Y,y	Measurements	α,a	Trust Values
$\mathcal{L}$,$\mathcal{M}$	Legitimate, Malicious	t	True Legitimacy
$\mathcal{H}_0$,$\mathcal{H}_1$	Hypothesis	N	Number of Robots
FA	False Alarm	MD	Missed Detection
γ_t,γ_{GLRT},γ_{TS}	Decision Thresholds	P_{e}	Probability of Error

TABLE I: COMMON NOTATION

IV. TWO STAGE APPROACH

The first approach, called the Two Stage Approach (2SA), finds the optimum decision rule that solves Problem 1.

A. Two Stage Approach Algorithm

In this section we present an approach where we separate 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 choose $\mathcal{H}_0$ 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 [34, Chapter 3].

The FC decides which robots to trust using the LRT

$$\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)} \underset{\hat{t}_i=0}{\overset{\hat{t}_i=1}{\gtrless}} \gamma_t, \tag{9}$$

where γ_t is a threshold value that we optimize. Note that when $\gamma_t=1$, (9) is equivalent to a maximum likelihood detection.

The FC decides who to trust and stores it in the vector $\hat{\mathbf{t}}$, where $\hat{t}_i=1$ if the FC chooses to trust the robot, 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 the FC chooses $\hat{t}_i=0$ with probability $1-p_t$, where p_t is another parameter to be optimized. This leads to the following trust probabilities, where $P_{\text{trust,L}}(\gamma_t,p_t)$ is the probability of trusting a legitimate robot, and $P_{\text{trust,M}}(\gamma_t,p_t)$ is the probability of trusting a malicious robot:

$$\begin{aligned}
P_{\text{trust,L}}(\gamma_t,p_t) &= \Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}>\gamma_t \middle| t_i=1\right) \\
&\quad + p_t\Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}=\gamma_t \middle| t_i=1\right), \\
P_{\text{trust,M}}(\gamma_t,p_t) &= \Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}>\gamma_t \middle| t_i=0\right) \\
&\quad + p_t\Pr\left(\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}=\gamma_t \middle| t_i=0\right).
\end{aligned} \tag{10}$$

The error probability $P_{\text{e}}(\mathcal{D},\boldsymbol{t},P_{\text{FA,M}},P_{\text{MD,M}})$ at the FC in (5) is affected by the trustworthiness classification. That is, if a legitimate robot i is classified as malicious, the FC discards its measurement Y_i, which increases the error probability since fewer measurements are used in the FC decision making. On the other hand, if a malicious robot is classified as legitimate, it can increase the error probability by sending falsified measurements to the FC. For that reason, we look to optimize the trustworthiness classification to

balance these two conflicting scenarios. Determining the best γ_t and p_t to minimize the overall error probability of the hypothesis detection by the FC is the main focus of this section.

b) Detecting the Event Ξ: To determine a hypothesis $\mathcal{H}$ on the event Ξ, the FC only considers the measurements it receives from robots that it classifies as legitimate in the first stage, i.e., $i:\hat{t}_i=1$. Equivalently, the FC discards all the received measurements of robots it classifies as malicious. Then, the FC uses the following decision rule:

$$\frac{\prod_{\{i:\hat{t}_i=1\}} P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}}{\prod_{\{i:\hat{t}_i=1\}}(1-P_{\text{FA,L}})^{1-y_i}P_{\text{FA,L}}^{y_i}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)} = \exp(\gamma_{\text{TS}}), \quad (11)$$

where $\exp(\gamma_{\text{TS}}) \triangleq \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$ is the exponential function with respect to γ_{TS}, and it is a constant decision threshold. We set $\frac{\Pr(\Xi=0)}{\Pr(\Xi=1)} = \exp(\gamma_{\text{TS}})$ so that when we take the logarithm in later expressions we can express the resultant decision threshold as γ_{TS} for ease of exposition. This decision rule is commonly used in standard binary hypothesis testing problems where no malicious robots are present, and will be referred to as the *standard binary hypothesis decision rule*. The standard binary hypothesis decision rule is optimal in a system with no malicious robots, i.e., $\mathcal{M}=\emptyset$, and thus we attempt to approximate the standard binary hypothesis decision rule by first removing information from all robots deemed to be malicious. However, since there may be detection errors in the first stage which classifies legitimate and malicious robots, the threshold γ_t and tie-break probability p_t should balance the need to exclude malicious robots from participating in the test (11) with the need to allow legitimate robots to participate in the test (11) and contribute their truthful measurements to decrease the probability of error resulting from (11). In what follows we show how to optimize the threshold γ_t and tie-break probability p_t by first computing the probability of error of the FC using the 2SA.

Recalling the Neyman-Pearson Lemma [34], we have that (9) minimizes the missed detection probability for a desired false alarm probability. This false alarm probability dictates the value of the threshold γ_t. After the FC discards measurements that it does not trust, the decision rule (11) leads to the following false alarm and missed detection probabilities,

$$\begin{aligned} &P_{\text{FA}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}}) \\ &=\Pr\Big(\sum_{i=1}^{N}\hat{t}_i[w_{1,\text{L}}y_i-w_{0,\text{L}}(1-y_i)]\geq\gamma_{\text{TS}} \\ &\qquad\qquad \Big|\mathcal{H}_0,\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}}\Big), \\ &P_{\text{MD}}(\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}}) \\ &=\Pr\Big(\sum_{i=1}^{N}\hat{t}_i[w_{1,\text{L}}y_i-w_{0,\text{L}}(1-y_i)]<\gamma_{\text{TS}} \\ &\qquad\qquad \Big|\mathcal{H}_1,\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}}\Big), \end{aligned} \quad (12)$$

where $w_{1,\text{L}} = \log\left(\frac{1-P_{\text{MD,L}}}{P_{\text{FA,L}}}\right)$, and $w_{0,\text{L}} = \log\left(\frac{1-P_{\text{FA,L}}}{P_{\text{MD,L}}}\right)$. Consequently, the overall error probability at the FC is:

$$\begin{aligned} P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}) &= \Pr(\Xi=0)P_{\text{FA}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}}) \\ &\quad +\Pr(\Xi=1)P_{\text{MD}}(\gamma_t,p_t,\mathbf{t},P_{\text{MD,M}}). \end{aligned} \quad (13)$$

We seek to minimize the probability of error (13) for the decision rule (11) by minimizing the false alarm and missed detection probabilities. Any sequence of 0's and 1's can occur for the detected trust vector $\hat{\mathbf{t}}$, each yielding a different error probability, so the error probability must be calculated for each possible vector $\hat{\mathbf{t}}$, along with each possible vector of observations $\mathbf{y}$. Unfortunately, this computation scales exponentially with the number of robots, N. Furthermore, the true trust vector $\mathbf{t}$ and the probabilities of false alarm and missed detection of the malicious robots are unknown, i.e., $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$, therefore, they cannot be used in minimizing (13).

To this end, we derive analytical guarantees regarding the error probability of the overall detection performance of the two-stage approach 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 (13). 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

$$(\gamma_t^*,p_t^*)=\underset{\gamma_t,p_t}{\operatorname{argmin}}\ \max_{\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}} P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}). \quad (14)$$

To this end, 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, denoted by m, is known, or we choose an upper bound for it ($\overline{m}$).

Referring back to Example 1, the server (Waze, Google Maps, etc.) gathers each robot's binary decision of the traffic condition along with a trust value corresponding to each robot's likely trustworthiness. With the 2SA method, the server uses the trust information to decide which robots' information should be trusted using (9), and discards the rest. Then, it makes a decision based on the remaining information, assuming it all to be legitimate information, using (11). Assuming the 2SA is used, in Lemma 1 we analytically determine what strategy the malicious robots should employ to increase the error probability. Subsequently, in Lemma 2 we derive the worst-case proportion of malicious robots that will lead the server to arrive at the wrong binary decision about the occurrence of traffic with the highest probability.

Lemma 1. *If $P_{\text{FA,L}}<0.5$ and $P_{\text{MD,L}}<0.5$, then the probability of false alarm and missed detection of the FC* (12) *is maximized for the Two Stage Approach when malicious robots choose $P_{FA,M}=P_{MD,M}=1$, for any vector $\boldsymbol{t}\in\{0,1\}^N$.*

Proof. Recall the false alarm and missed detection probabilities for the FC using (9) and (11) that lead to the overall false alarm and missed detection probabilities stated in (12).

Next, we show that the false alarm probability (12) is maximized when $P_{\text{FA,M}}=1$. The proof for $P_{\text{MD,M}}$ is analogous. In order to maximize P_{FA} in (12) the summation must be maximized. We rewrite the summation by separating it into the terms affected by legitimate robots that were trusted and those affected by malicious robots that were trusted

$$\begin{aligned} &\sum_{i:\{\hat{t}_i=1,t_i=0\}}[w_{1,\text{L}}y_j-w_{0,\text{L}}(1-y_j)]+ \\ &\sum_{i:\{\hat{t}_i=1,t_i=1\}}[w_{1,\text{L}}y_i-w_{0,\text{L}}(1-y_i)]. \end{aligned} \quad (15)$$

Any robot $j:\{\hat{t}_i=1,\,t_i=0\}$ can maximize (15) by

maximizing $[w_{1,\mathrm{L}}y_j - w_{0,\mathrm{L}}(1-y_j)]$. Note that when $P_{\mathrm{FA,L}} < 0.5$ and $P_{\mathrm{MD,L}} < 0.5$ then $w_{1,\mathrm{L}} > 0$ and $w_{0,\mathrm{L}} > 0$. Thus, $[w_{1,\mathrm{L}}y_j - w_{0,\mathrm{L}}(1-y_j)]$ is maximized when $y_j = 1$ since $Y_j \in \{0,1\}$. Given the true hypothesis is $\mathcal{H}_0$, the measurement $Y_j = 1$ occurs when robot j reports a false alarm. Therefore, the probability that robot j reports $Y_j = 1$ is maximized when the probability of false alarm is maximized, i.e., $\Pr(Y_j = 1|\mathcal{H}_0) = P_{\mathrm{FA,M}} = 1$. □

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* (13)*. If $P_{\mathrm{FA,L}} < 0.5$, $P_{\mathrm{MD,L}} < 0.5$, and $P_{\mathrm{FA,M}} = P_{\mathrm{MD,M}} = 1$, then the probability of error $P_e(\gamma_t, p_t, \bar{\mathbf{t}}, 1, 1)$ is maximized when $\bar{\mathbf{t}}$ contains the maximum number of malicious robots, i.e., $\sum_{i\in\mathcal{N}} \bar{t}_i = N - \overline{m}N$.*

Proof. By Lemma 1 the probability of false alarm and missed detection (12) are maximized when a robot is trusted and its measurement reports the wrong hypothesis, i.e., $y_i = 1$ when the true event is $\Xi = 0$ or $y_i = 0$ when the true event is $\Xi = 1$. Since the optimal policy for malicious robots is to report the wrong hypothesis with probability 1 (Lemma 1), any robot increases the false alarm and missed detection probability of the FC when it is malicious instead of legitimate. Thus, the probability of error $P_{\mathrm{e}}(\gamma_t, p_t, \mathbf{t}, 1, 1)$ is maximized when the proportion of malicious robots, m, is maximized, i.e., when $\bar{\mathbf{t}}$ has $\overline{m}N$ malicious robots, where $\overline{m}$ is the upper bound on the proportion of malicious robots in the network. □

Intuitively, the results of Lemmas 1 and 2 show that since the 2SA trusts all information that passes the first stage, the probability of error will be maximized when the proportion of malicious robots is maximized and they always send the wrong information. This maximizes the probability of misinformation reaching the second stage and affecting the final decision.

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 $P_{\mathrm{FA,M}} = P_{\mathrm{MD,M}} = 1$. In order to compute the probability of error exactly, we must compute the probability of false alarm and missed detection using (12). Let $k_{\mathrm{L}} \in K_{\mathrm{L}}$ be the number of legitimate robots trusted by the FC, where $K_{\mathrm{L}} = \{0,\ldots,(1-\overline{m})N\}$. Similarly, let $k_{\mathrm{M}} \in K_{\mathrm{M}}$ be the number of malicious robots trusted by the FC, where $K_{\mathrm{M}} = \{0,\ldots,\overline{m}N\}$. Let S_N represent the left side of the inequalities in (12) given by $S_{\mathrm{N}} = \sum_{i=1}^{N} \hat{t}_i[w_{1,\mathrm{L}}y_i - w_{0,\mathrm{L}}(1-y_i)]$. Using the law of total probability, the false alarm probability is

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

Similarly, the probability of missed detection of the FC is

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

The probability of false alarm for a particular instantiation of k_{L} and k_{M} can be written as a function of the binomial Cumulative Distribution Function:

$$\begin{aligned} &P_{\mathrm{FA}}(S_{\mathrm{N}} \geq \gamma_{\mathrm{TS}} | \mathcal{H}_0, k_{\mathrm{L}}, k_{\mathrm{M}}) \\ &= \Pr\left(\sum_{i:\{\hat{t}_i=1, t_i=1\}} y_i \geq \frac{\gamma_{\mathrm{TS}} - k_{\mathrm{M}}w_{1,\mathrm{L}} + k_{\mathrm{L}}w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} \middle| \mathcal{H}_0, k_{\mathrm{L}}, k_{\mathrm{M}}, \right), \\ &= 1 - F_{\mathrm{b}}\left(\lceil \frac{\gamma_{\mathrm{TS}} - k_{\mathrm{M}}w_{1,\mathrm{L}} + k_{\mathrm{L}}w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} \rceil - 1; P_{\mathrm{FA,L}}, k_{\mathrm{L}} \right), \end{aligned} \tag{16}$$

where $F_{\mathrm{b}}(g; p, n) = \sum_{i=0}^{g} \binom{n}{i} p^i (1-p)^{n-i} = \Pr(\sum_{i=1}^{n} y_i \leq g)$ is the binomial Cumulative Distribution Function evaluated at g for n variables and success probability p. Similarly, for the probability of missed detection, we have

$$\begin{aligned} &P_{\mathrm{MD}}(S_{\mathrm{N}} < \gamma_{\mathrm{TS}} | \mathcal{H}_1, k_{\mathrm{L}}, k_{\mathrm{M}}) = \\ &F_{\mathrm{b}}\left(\lceil \frac{\gamma_{\mathrm{TS}} + k_{\mathrm{M}}w_{0,\mathrm{L}} + k_{\mathrm{L}}w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} \rceil - 1; 1 - P_{\mathrm{MD,L}}, k_{\mathrm{L}} \right). \end{aligned} \tag{17}$$

Recall (10). We note that these probabilities depend on the distribution of the robot's trust values $\boldsymbol{a}$. Then, we have that

$$\begin{aligned} \Pr(K_{\mathrm{L}} = k_{\mathrm{L}}) &= \Pr\left(\sum_{i\in\mathcal{L}} \hat{t}_i = k_{\mathrm{L}} \right) \\ &= f_{\mathrm{b}}(k_{\mathrm{L}}; P_{\mathrm{trust,L}}(\gamma_t, p_t), (1-\overline{m})N), \\ \Pr(K_{\mathrm{M}} = k_{\mathrm{M}}) &= \Pr\left(\sum_{i\in\overline{\mathcal{M}}} \hat{t}_i = k_{\mathrm{M}} \right) \\ &= f_{\mathrm{b}}(k_{\mathrm{M}}; P_{\mathrm{trust,M}}(\gamma_t, p_t), \overline{m}N), \end{aligned}$$

where $f_{\mathrm{b}}(g; p, n) = \binom{n}{g} p^g (1-p)^{n-g} = \Pr(\sum_{i=1}^{n} y_i = g)$ is the binomial probability distribution function evaluated at g for n variables and success probability p. Thus, the probability of false alarm and missed detection are

$$\begin{aligned} &P_{\mathrm{FA}}(\gamma_t, p_t, \overline{m}, 1) \\ &= \sum_{k_{\mathrm{L}}\in K_{\mathrm{L}}, k_{\mathrm{M}}\in K_{\mathrm{M}}} f_{\mathrm{b}}(k_{\mathrm{L}}; P_{\mathrm{trust,L}}(\gamma_t, p_t), (1-\overline{m})N) \cdot \\ &\qquad f_{\mathrm{b}}(k_{\mathrm{M}}; P_{\mathrm{trust,M}}(\gamma_t, p_t), \overline{m}N) \cdot \\ &\qquad P_{\mathrm{FA}}(S_{\mathrm{N}} \geq \gamma_{\mathrm{TS}} | \mathcal{H}_0, k_{\mathrm{L}}, k_{\mathrm{M}}), \\ &P_{\mathrm{MD}}(\gamma_t, p_t, \overline{m}, 1) \\ &= \sum_{k_{\mathrm{L}}\in K_{\mathrm{L}}, k_{\mathrm{M}}\in K_{\mathrm{M}}} f_{\mathrm{b}}(k_{\mathrm{L}}; P_{\mathrm{trust,L}}(\gamma_t, p_t), (1-\overline{m})N) \cdot \\ &\qquad f_{\mathrm{b}}(k_{\mathrm{M}}; P_{\mathrm{trust,M}}(\gamma_t, p_t), \overline{m}N) \cdot \\ &\qquad P_{\mathrm{MD}}(S_{\mathrm{N}} < \gamma_{\mathrm{TS}} | \mathcal{H}_1, k_{\mathrm{L}}, k_{\mathrm{M}}). \end{aligned} \tag{18}$$

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

$$\begin{aligned} \overline{P}_{\mathrm{e}}(\gamma_t, p_t, \overline{m}, 1, 1) \triangleq &\Pr(\Xi = 0) P_{\mathrm{FA}}(\gamma_t, p_t, \overline{m}, P_{\mathrm{FA,M}} = 1) + \\ &\Pr(\Xi = 1) P_{\mathrm{MD}}(\gamma_t, p_t, \overline{m}, P_{\mathrm{MD,M}} = 1), \end{aligned} \tag{19}$$

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

Lemma 3. *Denote $\Gamma_t := \left\{ \frac{p_\alpha(a|t_i=1)}{p_\alpha(a|t_i=0)} \right\}_{a\in\mathcal{A}}$, where $\{\cdot\}_{a\in\mathcal{A}}$ represents a set consisting of all possible values of $a \in \mathcal{A}$ and $\mathcal{A}$ follows Assumption 1. Then, the minimal value of* (14) *with respect to γ_t can be achieved by $\gamma_t \in \Gamma_t$.*

Proof. The proof follows directly from the finiteness of the set $\mathcal{A}$ and since p_t can take values in the interval $[0,1]$. □

Let $\Gamma_p := \{0, \delta_p, 2\delta_p, \ldots, 1\}$ with a discretization constant δ_p. Algorithm 1 explains the Two Stage Approach step-by-step. Algorithm 1 takes a set Γ_t as input. Then, for each $\hat{\gamma}_t \in \Gamma_t$ and each $\hat{p}_t \in \Gamma_p$ we compute $P_{\mathrm{trust,L}}(\hat{\gamma}_t, \hat{p}_t)$, $P_{\mathrm{trust,M}}(\hat{\gamma}_t, \hat{p}_t)$, as well as $P_{\mathrm{FA}}(\hat{\gamma}_t, \hat{p}_t, \overline{m}, 1)$ and $P_{\mathrm{MD}}(\hat{\gamma}_t, \hat{p}_t, \overline{m}, 1)$. Then we compute the probability of error at the FC for the given $\hat{\gamma}_t$ and $\hat{p}_t$. The $\hat{\gamma}_t$ and $\hat{p}_t$ that yields the minimum probability of error is then used in the decision rule in (9) to determine which robots to trust or not trust

Algorithm 1 Two Stage Approach

Input: $P_{\text{FA,L}}$, $P_{\text{MD,L}}$, $\hat{P}_{\text{FA,M}}=\hat{P}_{\text{MD,M}}=1$, $\Pr(\Xi=0)$, $\Pr(\Xi=1)$, $\mathbf{y}$, $\boldsymbol{a}$, $\overline{m}$, Γ_t, δ_p

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

Set $\Gamma_p=\{0,\delta_p,2\delta_p,...,1\}$.
Set $\gamma_{t,\text{temp}}=0$, $p_{t,\text{temp}}=0$, $P_{\text{e,temp}}=2$.
for all $\hat{\gamma}_t\in\Gamma_t$, $\hat{p}_t\in\Gamma_p$ **do**
% Compute probability of error for each γ_t,p_t (lines 4-6)
    Compute $P_{\text{trust,L}}(\hat{\gamma}_t,\hat{p}_t)$, $P_{\text{trust,M}}(\hat{\gamma}_t,\hat{p}_t)$ by (10).
    Compute $P_{\text{FA}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1)$, $P_{\text{MD}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1)$ by (18).
    Compute $\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)$ by (19).
    if $\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)<P_{\text{e,temp}}$ **then**
% Store the current min values in $\gamma_{t,temp}$,$p_{t,temp}$,$P_{e,temp}$
        Set $(\gamma_{t,\text{temp}},p_{t,\text{temp}})=(\hat{\gamma}_t,\hat{p}_t)$.
        Set $P_{\text{e,temp}}=\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)$.
Set $(\gamma_t,p_t)=(\gamma_{t,\text{temp}},p_{t,\text{temp}})$. *% Set γ_t,p_t to the values that yielded the lowest probability of error*
Determine the vector $\hat{\mathbf{t}}$ using (9). *% Estimate trust*
Determine decision using (11). *% Perform the standard hypothesis test using measurements from the trusted robots*
Return decision $\mathcal{H}_0$ or $\mathcal{H}_1$.

(vector $\hat{\mathbf{t}}$). Finally, we use the chosen vector $\hat{\mathbf{t}}$ to make a decision using the standard binary hypothesis decision rule (11).

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. However, this only needs to be computed once, and then the returned γ_t and p_t can be used to run each subsequent hypothesis test. With a given γ_t and p_t, the hypothesis test requires $\mathcal{O}(N)$ comparisons.

Theorem 1. *Assume that the FC uses the decision rule in* (9) *to detect malicious robots, and then uses the decision rule* (11)*. 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 up to a discretization distance*

$$d(\delta_p):=\min_{p_t\in\Gamma_p}\overline{P}_e(\gamma_t^*,p_t,\overline{m},1,1)-\overline{P}_e(\gamma_t^*,p_t^*,\overline{m},1,1).$$

Furthermore, $d(\delta_p)\to 0$ as $\delta_p\to 0$.

Proof. The goal is to minimize the worst-case probability of error of the FC, i.e.,

$$\min_{\gamma_t,p_t\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}}\max P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}).$$

Using the results from Lemmas 1, 2 and (18) we upper bound the error probability using the worst-case error probability:

$$\begin{aligned}\min_{\gamma_t,p_t\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}}}\max\ & P_{\text{e}}(\gamma_t,p_t,\mathbf{t},P_{\text{FA,M}},P_{\text{MD,M}})\\ &=\min_{\gamma_t,p_t}\max_{\mathbf{t}}P_{\text{e}}(\gamma_t,p_t,\mathbf{t},1,1),\\ &=\min_{\gamma_t,p_t}\overline{P}_e(\gamma_t,p_t,\overline{m},1,1).\end{aligned}$$

The equality in the first line directly follows from Lemma 1. The second line follows from the first by inserting the worst-case vector $\mathbf{t}$, with $\overline{m}$ malicious robots, as the one that maximizes the probability of error P_{e} (Lemma 2).

Additionally, by Lemma 3, it is sufficient to optimize γ_t over the set Γ_t. Now, since we optimize p_t using a line search, we may not necessarily find an optimal pair (γ_t^*,p_t^*). However, we can express the distance from the optimal solution by:

$$\begin{aligned}&\min_{\gamma_t\in\Gamma_t,p_t\in\Gamma_p}\overline{P}_{\text{e}}(\gamma_t,p_t,\overline{m},1,1)-\overline{P}_{\text{e}}(\gamma_t^*,p_t^*,\overline{m},1,1)\\ &\le\min_{p_t\in\Gamma_p}\overline{P}_{\text{e}}(\gamma_t^*,p_t,\overline{m},1,1)-\overline{P}_{\text{e}}(\gamma_t^*,p_t^*,\overline{m},1,1)=d(\delta_p).\end{aligned}$$

For every fixed γ_t, the function $\overline{P}_{\text{e}}(\gamma_t,p_t,\overline{m},1,1)$ is a polynomial function of p_t, therefore, it is continuous in p_t (over the interval $p_t\in[0,1]$). Consequently, $d(\delta_p)\to 0$ as $\delta_p\to 0$. □

B. Error Bounds for the Two Stage Approach

In this section we characterize the behavior of the 2SA as the number of robots in the system increases. Namely, we show that when the probability of the FC trusting a legitimate robot in the first stage of the 2SA (9) is much greater than the probability of the FC trusting a malicious robot, that the overall worst-case probability of error at the FC decreases towards 0 as the number of robots in the network increases.

Let $\beta_{\text{M}}\in(0,1)$ and $\beta_{\text{L}}\in(0,1)$ denote the proportion of malicious (resp. legitimate) robots that are trusted by the FC after the first stage. The terms β_{M} and β_{L} are purely for analytical purposes. They will be utilized to split the probability of error analysis into four separate events, corresponding to differing numbers of trusted legitimate and malicious robots.

Recall that $\overline{m}$ is the upper bound on the true proportion of malicious robots in the network. Assume $\overline{m}\in(0,1)$. Let us consider a given threshold value γ_t and tie-break probability p_t at the first stage, and let $P_{\text{trust,M}}(\gamma_t,p_t)$ and $P_{\text{trust,L}}(\gamma_t,p_t)$ be the resulting probability of trusting a malicious (resp. legitimate) robot. Furthermore, let us consider a given β_{M} and β_{L} such that $\beta_{\text{M}}>P_{\text{trust,M}}(\gamma_t,p_t)$ and $\beta_{\text{L}}<P_{\text{trust,L}}(\gamma_t,p_t)$. Intuitively, the values $P_{\text{trust,M}}(\gamma_t,p_t)$ and $P_{\text{trust,L}}(\gamma_t,p_t)$ correspond to the expected proportion of malicious and legitimate robots that will be trusted by the FC under the 2SA algorithm. Consequently, when we consider $\beta_{\text{M}}>P_{\text{trust,M}}(\gamma_t,p_t)$ and $\beta_{\text{L}}<P_{\text{trust,L}}(\gamma_t,p_t)$ we are representing undesirable regions where more than the expected proportion of malicious robots are trusted and less than the expected proportion of legitimate robots are trusted. Finally, assume $\beta_{\text{L}}|\mathcal{L}|>>\max\{\beta_{\text{M}}|\mathcal{M}|,1\}$. This means we consider scenarios where many more legitimate robots than malicious robots are trusted by the FC. This is likely to occur when $P_{\text{trust,L}}(\gamma_t,p_t)>>P_{\text{trust,M}}(\gamma_t,p_t)$.

We summarize the assumptions used here for convenience:

Assumption 2. *(a) We denote by $\beta_M\in(0,1)$ and $\beta_L\in(0,1)$ sample proportions of malicious (resp. legitimate) robots that are trusted by the FC after the first stage of the 2SA for analytical purposes. We analyze undesirable scenarios where many malicious and few legitimate robots are trusted, i.e., scenarios that satisfy $\beta_M>P_{trust,M}(\gamma_t,p_t)$ and $\beta_L<P_{trust,L}(\gamma_t,p_t)$.*
(b) We assume that $\beta_L|\mathcal{L}|>>\max\{\beta_M|\mathcal{M}|,1\}$, which is likely to occur when $P_{trust,L}(\gamma_t,p_t)>>P_{trust,M}(\gamma_t,p_t)$.
(c) We assume $\overline{m}\in(0,1)$.

Recall that k_{L} and k_{M} denote the actual number of legitimate and malicious robots trusted by the FC. In what follows, we upper bound the worst-case probability of error by examining four cases, each considering a different regime with respect to the number of trusted legitimate and malicious robots:

Case 1) $k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|$, $k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|$,
Case 2) $k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|$, $k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|$,
Case 3) $k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|$, $k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|$,
Case 4) $k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|$, $k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|$.

In words, these cases correspond to scenarios where 1) few legitimate and malicious robots are trusted after the first stage of the 2SA, 2) few legitimate robots but many malicious robots are trusted, 3) many legitimate and malicious robots are trusted, and 4) many legitimate robots but few malicious robots are trusted. Intuitively, Cases 1, 2, and 3 will contribute the most to the detection error probability since they contain either many malicious robots or few legitimate robots, whereas the fourth event is the most desirable since it contains many legitimate robots and few malicious robots. In what follows, we investigate scenarios where the probabilities corresponding to Cases 1, 2, or 3 occurring decay at least exponentially as the number of robots increases, then show that the probability of error given Case 4 also decays at least exponentially as the number of robots increases. In other words, we show that Cases 1, 2, and 3 are unlikely to occur under the assumed conditions, leaving Case 4 as the most likely case. Finally, we show that the probability of error is low when Case 4 occurs.

Recall that $\gamma_{\mathrm{TS}} = \log\left(\frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}\right)$ is the decision threshold used in the second stage of the 2SA (11), $\hat{t}_i$ denotes the outcome of the first stage which tests the trustworthiness of robot i, and $S_{\mathrm{N}} = \sum_{i=1}^{N} \hat{t}_i[w_{1,\mathrm{L}} y_i - w_{0,\mathrm{L}}(1-y_i)]$, is the left side of the inequalities in (12). The probability of a particular case occurring corresponds to the probability of trusting k_{L} and k_{M} robots that fall into the region described by the particular case. These probabilities are conditioned upon the chosen threshold values γ_t and p_t, but we omit these threshold values from the case probabilities for ease of exposition. With these four cases, we can upper bound the worst-case probability of error by using the union bound:

$$\begin{aligned}
&\overline{P}_{\mathrm{e}}(\gamma_t,p_t,\overline{m},P_{\mathrm{FA,M}}=1,P_{\mathrm{MD,M}}=1)\\
&\leq \Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|) \max_{\substack{k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|,\\ k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|}} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}})\\
&+\Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|)) \max_{\substack{k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|,\\ k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|}} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}})\\
&+\Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|) \max_{\substack{k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|,\\ k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|}} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}})\\
&+\Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|) \max_{\substack{k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|,\\ k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|}} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}), \qquad (20)
\end{aligned}$$

where

$$\begin{aligned}
\overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}) \triangleq{}& \Pr(\Xi=0)P_{\mathrm{FA}}(S_{\mathrm{N}} \geq \gamma_{\mathrm{TS}}|\mathcal{H}_0,k_{\mathrm{L}},k_{\mathrm{M}})\\
&+\Pr(\Xi=1)P_{\mathrm{MD}}(S_{\mathrm{N}} < \gamma_{\mathrm{TS}}|\mathcal{H}_1,k_{\mathrm{L}},k_{\mathrm{M}}),
\end{aligned}$$

represents the probability of error for a given k_{L} and k_{M} corresponding to a particular case. Intuitively, (20) shows how we can upper bound the worst-case probability of error by the sum of the worst-case probabilities of error for each of the 4 cases. Note that $\overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}) \leq 1$ for any k_{L} and k_{M}. Consequently, we can simplify (20) to

$$\begin{aligned}
\overline{P}_{\mathrm{e}}(\gamma_t,p_t,\overline{m},1,1) \leq{}& \Big[\Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|)\\
&+\Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|))\Big]\cdot 1\\
&+\Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|)\cdot 1\\
&+\Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|) \max_{\substack{k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|,\\ k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|}} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}). \qquad (21)
\end{aligned}$$

We utilize the upper bound $\overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}) \leq 1$ for the cases where few legitimate robots are trusted, $k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|$, or many malicious robots are trusted, $k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|$, to simplify the analysis, i.e., we upper bound the error probabilities that are likely to be high by 1. We intend to show that the probability of these cases occurring decays at least exponentially as the number of robots increases. Let

$$\begin{aligned}
\Pr(\text{Case 1}) &\triangleq \Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|),\\
\Pr(\text{Case 2}) &\triangleq \Pr(k_{\mathrm{L}} \leq \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|),\\
\Pr(\text{Case 3}) &\triangleq \Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} \geq \beta_{\mathrm{M}}|\mathcal{M}|),\\
\Pr(\text{Case 4}) &\triangleq \Pr(k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|), \qquad (22)
\end{aligned}$$

be the probability of Case 1, Case 2, Case 3, and Case 4 occurring, respectively. We are interested in characterizing how the probability of error in (21) is affected when the number of robots increases while keeping the proportion of malicious robots the same. To see this more clearly, we rewrite (21) using (22), and then analyze each term separately:

$$\begin{aligned}
\overline{P}_{\mathrm{e}}(\gamma_t,p_t,\overline{m},1,1) \leq{}& [\Pr(\text{Case 1})+\Pr(\text{Case 2})]\\
&+\Pr(\text{Case 3})+\Pr(\text{Case 4})\cdot \max_{k_{\mathrm{L}} > \beta_{\mathrm{L}}|\mathcal{L}|, k_{\mathrm{M}} < \beta_{\mathrm{M}}|\mathcal{M}|} \overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}}). \qquad (23)
\end{aligned}$$

We derive our upper bound (23) on the error probability by examining its terms. To this end, we utilize the following upper bound [46] which is derived from the Chernoff bound

$$\Pr(X \leq g) = F_{\mathrm{b}}(g;n,p) \leq \exp\left(-nD\left(\frac{g}{n}||p\right)\right), \qquad (24)$$

where we assume X to be a binomial distribution, n is the number of trials, p is the success probability, i.e., the probability a trial results in a 1, and

$$D(p||q) = p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right)$$

is the Kullback–Leibler (KL) divergence between a Bernoulli random variable with success probability p and a Bernoulli random variable with success probability q. The Chernoff bound (24) provides an upper bound for the lower tail of the cumulative distribution function for $\Pr(X \leq g)$, and is valid when $\frac{g}{n} \in (0,p)$. The Chernoff bound can also provide an upper bound for the upper tail of the cumulative distribution function for $\Pr(X \geq g)$ for $\frac{g}{n} \in (p,1)$.

Next, we analyze the terms within (23). Specifically, for Cases 1-3 we show that the probability of them occurring decays at least exponentially as the number of robots increases. Then, we show that the probability of Case 4 occurring approaches 1, but the corresponding probability of error for Case 4 decays at least exponentially.

Referring again to Example 1, we want to show that if the proportion of robots that are malicious is held constant and the trust values received by the server are likely to yield the correct trustworthiness of each of the robots, then the most likely case is that the server will trust mostly legitimate robots in making its traffic condition estimation. Thus, the probability of the server making the correct estimation will be high.

*a) **Cases 1 and 2**:* Cases 1 and 2 correspond to cases where few legitimate robots are trusted. We show that the probability of them occurring decays at least exponentially as the number

of robots increases. First, we simplify the probability of them occurring using the law of total probability:

$$\begin{aligned}&\Pr(\text{Case 1})+\Pr(\text{Case 2})\\&\quad=\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}}<\beta_{\mathrm{M}}|\mathcal{M}|)\\&\qquad+\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|))\\&\quad=\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|).\end{aligned}$$

Next, observe that the number of trusted legitimate robots, i.e., $k_{\mathrm{L}}=\sum_{i\in\mathcal{L}}\hat{t}_i$, is distributed according to a binomial distribution with the probability for $\hat{t}_i=1$ equal to the probability of trusting a legitimate robot $i\in\mathcal{L}$. Therefore, the upper bound on $\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|)$ can be written by

$$\Pr(k_{\mathrm{L}}\leq\exp(-(1-\overline{m})ND(\beta_{\mathrm{L}}||P_{\text{trust,L}}(\gamma_t,p_t))). \tag{25}$$

The Chernoff bound is valid here since we consider the region where $\beta_{\mathrm{L}}<P_{\text{trust,L}}(\gamma_t,p_t)$ (Assumption 2.a). It can be seen from (25) that the upper bound on the probability of Case 1 or Case 2 occurring decays exponentially with a rate of $(1-\overline{m})ND(\beta_{\mathrm{L}}||P_{\text{trust,L}}(\gamma_t,p_t))$ assuming $\overline{m}\neq 1$ (Assumption 2.c). This is guaranteed to be an exponential decay because the KL divergence is always non-negative, $\beta_{\mathrm{L}}\neq P_{\text{trust,L}}(\gamma_t,p_t)$, and $N>0$.

*b) **Case 3**:* Case 3 corresponds to the case where many legitimate robots are trusted by the FC, but also many malicious robots are trusted. Similar to Cases 1 and 2, we show that the probability of Case 3 occurring decays at least exponentially as the number of robots increases. Recall that

$$\Pr(\text{Case 3})=\Pr(k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|).$$

For Cases 1 and 2 we showed in (25) that $\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|)$ decays toward 0 at least exponentially as N increases. Since $\Pr(k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|)=1-\Pr(k_{\mathrm{L}}\leq\beta_{\mathrm{L}}|\mathcal{L}|)\leq 1$, we conclude by the sandwich theorem [47] that $\Pr(k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|)$ approaches 1 as N tends to infinity. However, observe that the number of trusted malicious robots, i.e., $k_{\mathrm{M}}=\sum_{i\in\mathcal{M}}\hat{t}_i$, is distributed according to a binomial distribution with the probability for $\hat{t}_i=1$ equal to the probability of trusting a robot i given that $i\in\mathcal{M}$. Then, using the Chernoff bound (24), the upper bound on $\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|)$ can be written by

$$\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|)\leq\exp(-\overline{m}ND(\beta_{\mathrm{M}}||P_{\text{trust,M}}(\gamma_t,p_t))). \tag{26}$$

The Chernoff bound is valid here since we consider the region where $\beta_{\mathrm{M}}>P_{\text{trust,M}}(\gamma_t,p_t)$ (Assumption 2.a). It can be seen from (26) that the upper bound on $\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|)$, and thus the probability of Case 3 occurring, decays exponentially with a rate of $\overline{m}ND(\beta_{\mathrm{M}}||P_{\text{trust,M}}(\gamma_t,p_t))$ assuming $\overline{m}\neq 0$ (Assumption 2.c). Again, this is guaranteed to be an exponential decay because the KL divergence is always non-negative, $\beta_{\mathrm{M}}\neq P_{\text{trust,M}}(\gamma_t,p_t)$, and $N>0$.

*c) **Case 4**:* Case 4 is the ideal case, where many legitimate robots are trusted by the FC and few malicious robots are trusted. We show that the probability of Case 4 occurring approaches 1 as the number of robots increases, and the corresponding probability of error decays at least exponentially. Recall that

$$\Pr(\text{Case 4})=\Pr(k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|)\Pr(k_{\mathrm{M}}<\beta_{\mathrm{M}}|\mathcal{M}|).$$

We already showed that $\Pr(k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|)\to 1$ as $N\to\infty$. Similarly, $\Pr(k_{\mathrm{M}}<\beta_{\mathrm{M}}|\mathcal{M}|)\to 1$ as $N\to\infty$ since $\Pr(k_{\mathrm{M}}\geq\beta_{\mathrm{M}}|\mathcal{M}|)\to 0$.

For Case 4 we must also upper bound the probability of error, which requires upper bounding the probability of false alarm and missed detection for a given number of trusted legitimate, k_{L}, and malicious robots, k_{M}. For both of these, we use the Chernoff bound. First, we analyze the false alarm. Recall (16) which allows us to write the upper bound as

$$\begin{aligned}&P_{\text{FA}}(S_{\mathrm{N}}\geq\gamma_{\text{TS}}|\mathcal{H}_0,k_{\mathrm{L}},k_{\mathrm{M}})\\&\quad=\Pr\left(\textstyle\sum_{i:\{\hat{t}_i=1,t_i=1\}}y_i\geq\frac{\gamma_{\text{TS}}-k_{\mathrm{M}}w_{1,\mathrm{L}}+k_{\mathrm{L}}w_{0,\mathrm{L}}}{w_{0,\mathrm{L}}+w_{1,\mathrm{L}}}\middle|\mathcal{H}_0,k_{\mathrm{L}},k_{\mathrm{M}}\right)\\&\quad\leq\exp(-k_{\mathrm{L}}D(\tilde{\gamma}_{\text{FA}}(k_{\mathrm{L}},k_{\mathrm{M}})||P_{\text{FA,L}})),\end{aligned} \tag{27}$$

where

$$\tilde{\gamma}_{\text{FA}}(k_{\mathrm{L}},k_{\mathrm{M}}):=\frac{1}{k_{\mathrm{L}}}\frac{\gamma_{\text{TS}}-k_{\mathrm{M}}w_{1,\mathrm{L}}+k_{\mathrm{L}}w_{0,\mathrm{L}}}{(w_{0,\mathrm{L}}+w_{1,\mathrm{L}})}, \tag{28}$$

is the threshold on the RHS of the inequality in (16) and (27) normalized with respect to k_{L}.

Similarly, the probability of missed detection given k_{L} and k_{M} is upper bounded by

$$\begin{aligned}&P_{\text{MD}}(S_{\mathrm{N}}<\gamma_{\text{TS}}|\mathcal{H}_1,k_{\mathrm{L}},k_{\mathrm{M}})\\&\qquad\leq\exp(-k_{\mathrm{L}}D(\tilde{\gamma}_{MD}(k_{\mathrm{L}},k_{\mathrm{M}})||1-P_{\text{MD,L}})),\end{aligned} \tag{29}$$

where $\tilde{\gamma}_{\text{MD}}(k_{\mathrm{L}},k_{\mathrm{M}}):=\frac{1}{k_{\mathrm{L}}}\frac{\gamma_{\text{TS}}+k_{\mathrm{M}}w_{0,\mathrm{L}}+k_{\mathrm{L}}w_{0,\mathrm{L}}}{(w_{0,\mathrm{L}}+w_{1,\mathrm{L}})}$.

The Chernoff bounds in (27) and (29) are valid whenever $\tilde{\gamma}_{\text{FA}}(k_{\mathrm{L}},k_{\mathrm{M}})\in(P_{\text{FA,L}},1)$ and $\tilde{\gamma}_{\text{MD}}(k_{\mathrm{L}},k_{\mathrm{M}})\in(0,1-P_{\text{MD,L}})$.

From here we upper bound the probability of error corresponding to Case 4 by noticing that our upper bound on the probability of error is maximized when the least legitimate robots are trusted and the most malicious robots are trusted. Define $\underline{k_{\mathrm{L}}}\triangleq\beta_{\mathrm{L}}|\mathcal{L}|+1$ and $\overline{k_{\mathrm{M}}}\triangleq\beta_{\mathrm{M}}|\mathcal{M}|-1$ to be the minimum number of legitimate robots within the region $k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|$, and the maximum number of malicious robots within the region $k_{\mathrm{M}}<\beta_{\mathrm{M}}|\mathcal{M}|$ that can be trusted, respectively. The following lemma formulates this observation.

Lemma 4. *Consider Case 4 where many legitimate robots are trusted by the FC and few malicious robots are trusted. Without loss of generality, assume* $\Pr(\Xi=1)>\Pr(\Xi=0)$. *If* $\tilde{\gamma}_{FA}(k_L,k_M)\in(P_{FA,L},1)$ *and* $\tilde{\gamma}_{MD}(k_L,k_M)\in(0,1-P_{MD,L})$, *then the probability of error for a given* k_L *and* k_M *within Case 4 can be upper bounded by*

$$\begin{aligned}&\max_{\substack{k_L>\beta_L|\mathcal{L}|,\\k_M<\beta_M|\mathcal{M}|}}\overline{p}_e(k_L,k_M)\\&\leq\Pr(\Xi=0)\max_{\substack{k_L>\beta_L|\mathcal{L}|,\\k_M<\beta_M|\mathcal{M}|}}P_{FA}(S_N\geq\gamma_{TS}|\mathcal{H}_0,k_L,k_M)\\&+\Pr(\Xi=1)\max_{\substack{k_L>\beta_L|\mathcal{L}|,\\k_M<\beta_M|\mathcal{M}|}}P_{MD}(S_N<\gamma_{TS}|\mathcal{H}_1,k_L,k_M).\end{aligned} \tag{30}$$

Additionally, assume $\beta_L|\mathcal{L}| >> \max\{\beta_M|\mathcal{M}|, 1\}$. *Then, there exists values* $k_L>\beta_L|\mathcal{L}|$, *and* $k_M<\beta_M|\mathcal{M}|$ *such that* $\tilde{\gamma}_{FA}(k_L,k_M)\in(P_{FA,L},1)$ *and* $\tilde{\gamma}_{MD}(k_L,k_M)\in(0,1-P_{MD,L})$.

Proof. The proof of Lemma 4 can be found in Appendix A. □

From Lemma 4 we have that

$$\begin{aligned}&\max_{k_{\mathrm{L}}>\beta_{\mathrm{L}}|\mathcal{L}|,k_{\mathrm{M}}<\beta_{\mathrm{M}}|\mathcal{M}|}\overline{p}_{\mathrm{e}}(k_{\mathrm{L}},k_{\mathrm{M}})\\&\qquad\leq\Pr(\Xi=0)q_{\text{FA}}+\Pr(\Xi=1)q_{\text{MD}},\end{aligned}$$

where

$$q_{\text{FA}}=e^{(-(\beta_{\mathrm{L}}(1-\overline{m})N+1)D(\tilde{\gamma}_{FA}(\beta_{\mathrm{L}}(1-\overline{m})N+1,\beta_{\mathrm{M}}\overline{m}N-1)||P_{\text{FA,L}}))}$$

and

$$q_{\text{MD}}=e^{(-(\beta_{\mathrm{L}}(1-\overline{m})N+1)D(\tilde{\gamma}_{MD}(\beta_{\mathrm{L}}(1-\overline{m})N+1,\beta_{\mathrm{M}}\overline{m}N-1)||1-P_{\text{MD,L}}))}.$$

Intuitively, Lemma 4 allows us to upper bound the worst-case probability of error under Case 4 by two terms consisting of $\Pr(\Xi=0)$ and $\Pr(\Xi=1)$, which are constants, and q_{FA} and q_{MD}. Therefore, if we show that q_{FA} and q_{MD} both decay exponentially with N, then so does the upper bound for the worst-case probability of error for Case 4. Indeed, we see that the upper bound for the probability of error when Case 4 occurs decays exponentially with a rate of

$$\begin{aligned}&(\beta_{\text{L}}(1-\overline{m})N+1)\cdot\\ &\min\{D(\tilde{\gamma}_{\text{FA}}(\beta_{\text{L}}(1-\overline{m})N+1,\beta_{\text{M}}\overline{m}N-1)||P_{\text{FA,L}}),\\ &\quad D(\tilde{\gamma}_{\text{MD}}(\beta_{\text{L}}(1-\overline{m})N+1,\beta_{\text{M}}\overline{m}N-1)||1-P_{\text{MD,L}})\}.\end{aligned}$$

We summarize the main result of this section with the following proposition.

Proposition 1. *Assume Assumptions 2.a - 2.c hold. Then the worst-case probability of error, $\overline{P}_{\text{e}}(\gamma_t,p_t,\overline{m},1,1)$ in* (19)*, decays towards zero at least exponentially as the number of robots in the network increases.*

Proof. This result follows from the results of Sections IV-B.a - IV-B.c, which show that $[\Pr(\text{Case 1})+\Pr(\text{Case 2})]\to 0$, $\Pr(\text{Case 3})\to 0$, $\Pr(\text{Case 4})\to 1$ as $N\to\infty$, and the upper bound on the probability of error corresponding to Case 4, $\overline{p}_{\text{e}}(\underline{k_{\text{L}}},\overline{k_{\text{M}}})\to 0$, as $N\to\infty$. Since all upper bounds exhibit exponential decay rates, we conclude that the probability of error decays towards 0 at least exponentially as the number of robots in the network increases. □

C. Analyzing the Limits of the Two Stage Approach

If the proportion of malicious robots in the network, i.e., $\overline{m}$, is high enough, the probability of error for the 2SA will plateau. Intuitively, this is due to the fact that if there are too many malicious robots it becomes more beneficial for the FC to guess between $\mathcal{H}_0$ or $\mathcal{H}_1$ using the prior probabilities $\Pr(\Xi=0)$ and $\Pr(\Xi=1)$ rather than utilize any measurements from robots. In the context of Example 1, this would correspond to a scenario where most, if not all of the robots on the road are sending malicious information. In this case, the server likely has a better chance of correctly estimating the traffic conditions by making an informed guess based on previously known traffic patterns. In this section, we wish to find the critical proportion of robots on the road that need to be malicious for this to happen. We note that since trust values are used, the proportion of malicious robots that causes the 2SA to plateau is not necessarily the typical $\overline{m}=0.5$. We formally state and prove this observation with the following lemma. Recall that Algorithm 1 chooses the classification threshold γ_t and tie-break probability p_t by computing the probability of error in the presence of a worst-case attack over all values $\hat{\gamma}_t\in\Gamma_t$ and $\hat{p}_t\in\Gamma_p$, where $\Gamma_t=\left\{\frac{p_\alpha(a|t_i=1)}{p_\alpha(a|t_i=0)}\right\}_{a\in\mathcal{A}}$, $\Gamma_p=\{0,\delta_p,2\delta_p,...,1\}$, and δ_p is a given discretization.

Lemma 5. *If the worst-case probability of error for every choice of $\hat{\gamma}_t$ and $\hat{p}_t$ is no better than performing event detection with no information, i.e.,*

$$\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)\geq\min\{\Pr(\Xi=0),\Pr(\Xi=1)\},$$

for all $\hat{\gamma}_t\in\Gamma_t$ and all $\hat{p}_t\in\Gamma_p$, then the optimal worst-case probability of error becomes the probability of the less likely event between $\Xi=0$ and $\Xi=1$ occurring, i.e.,

$$\overline{P}_e(\gamma_t^*,p_t^*,\overline{m},1,1)=\min\{\Pr(\Xi=0),\Pr(\Xi=1)\}.$$

Furthermore, the Two Stage Approach chooses thresholds γ_t and p_t that lead to not trusting any robots, i.e., $\gamma_t^*=\max_{a_i\in\alpha}\left\{\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}\right\}$ and $p_t^*=0$.

Proof. Let $\hat{\gamma}_t=\max_{a_i\in\alpha}\left\{\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}\right\}$ and $\hat{p}_t=0$. This corresponds to the scenario where the measurements from all robots will be discarded by the FC in the first stage. Discarding all measurements simplifies the decision rule (11) to

$$1\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}}\frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}.$$

If $\Pr(\Xi=0)\geq\Pr(\Xi=1)$ then the FC chooses $\mathcal{H}_0$ which leads to an error probability of $\Pr(\Xi=1)$. If $\Pr(\Xi=0)<\Pr(\Xi=1)$ then the FC chooses $\mathcal{H}_1$ which leads to an error probability of $\Pr(\Xi=0)$. Therefore, the probability of error

$$\overline{P}_{\text{e}}(\hat{\gamma}_t,\hat{p}_t,\overline{m},1,1)=\min\{\Pr(\Xi=0),\Pr(\Xi=1)\}.$$

By Algorithm 1 if the probability of error is greater for all other $\hat{\gamma}_t\in\Gamma_t$ and $\hat{p}_t\in\Gamma_p$, then $\gamma_t^*=\max_{a_i\in\alpha}\left\{\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}\right\}$ and $p_t^*=0$. □

This lemma formally shows that if at some point the probability of error when trusting any robots is always greater than the probability of error from using the prior probabilities $\Pr(\Xi=0)$ and $\Pr(\Xi=1)$ then Algorithm 1 chooses γ_t^* and p_t^* such that no robots are ever trusted. This reduces to the case where the hypothesis test is done using the event probabilities.

Let m^* denote the critical proportion of malicious robots that causes the 2SA to reject all information in the first stage, i.e., for all $\overline{m}\geq m^*$ we have $\overline{P}_{\text{e}}(\gamma_t^*,p_t^*,\overline{m},1,1)=\min\{\Pr(\Xi=0),\Pr(\Xi=1)\}$. Next we develop an understanding of how m^* is affected by the quality of the trust values, i.e., as a function of the probability of trusting legitimate and malicious robots, $P_{\text{trust,L}}(\gamma_t,p_t)$ and $P_{\text{trust,M}}(\gamma_t,p_t)$. In order to do so, we assume as a simplification that there is no noise in the sensor measurements of legitimate robots, i.e., $P_{\text{FA,L}}=P_{\text{MD,L}}=0$. This allows us to simplify the probability of false alarm in (16) by considering y_i to be a deterministic variable with respect to the legitimacy of robot i:

$$P_{\text{FA}}(S_{\text{N}}\geq\gamma_{\text{TS}}|\mathcal{H}_0)=\Pr(-K_{\text{L}}w_{0,\text{L}}+K_{\text{M}}w_{1,\text{L}}\geq\gamma_{\text{TS}}|\mathcal{H}_0),\quad(31)$$

where $K_{\text{L}}\in\{0,1,\dots,(1-\overline{m})N\}$ and $K_{\text{M}}\in\{0,1,...,\overline{m}N\}$ are random variables that represent the possible number of trusted legitimate and malicious robots, respectively. When $P_{\text{FA,L}}=P_{\text{MD,L}}$ we have that $w_{0,\text{L}}=w_{1,\text{L}}$. Then (31) becomes

$$P_{\text{FA}}(S_{\text{N}}\geq\gamma_{\text{TS}}|\mathcal{H}_0)=\Pr(K_{\text{M}}-K_{\text{L}}\geq 0|\mathcal{H}_0),\quad(32)$$

where we use the fact that $w_{0,\text{L}}\to\infty$ and $w_{1,\text{L}}\to\infty$ as $P_{\text{FA,L}}\to 0$ and $P_{\text{MD,L}}\to 0$. Similarly, the probability of missed detection becomes

$$P_{\text{MD}}(S_{\text{N}}<\gamma_{\text{TS}}|\mathcal{H}_1)=\Pr(K_{\text{M}}-K_{\text{L}}>0|\mathcal{H}_1).\quad(33)$$

The variables K_{L} and K_{M} are distributed according to binomial distributions:

$$K_{\text{L}}\sim\text{BIN}((1-\overline{m})N,P_{\text{trust,L}}),K_{\text{M}}\sim\text{BIN}(\overline{m}N,P_{\text{trust,M}}),$$

where $\text{BIN}(n,p)$ corresponds to a binomial distribution with n trials and success probability p.

Define $Z\triangleq K_{\text{M}}-K_{\text{L}}$ to be a discrete random variable corresponding to the difference of the two binomial random

variables K_{M} and K_{L}. We are interested in $\Pr(Z \geq 0)$ for the probability of false alarm (see (32)), and $\Pr(Z > 0)$ for the probability of missed detection (see (33)). Then, m^* could be found by finding the minimum $\overline{m}$ such that

$$\begin{aligned}\Pr(\Xi=0)\Pr(Z\geq 0)+\Pr(\Xi=1)\Pr(Z>0)\\ \geq \min\{\Pr(\Xi=0),\Pr(\Xi=1)\}.\end{aligned}$$

where $\Pr(Z \geq 0)$ and $\Pr(Z > 0)$ are a function of $P_{\text{trust,L}}(\gamma_t, p_t)$, $P_{\text{trust,M}}(\gamma_t, p_t)$, $\overline{m}$, and N.

When N is large the distribution of Z is approximately normal with mean and variance

$$\mu = \overline{m}N(P_{\text{trust,L}}+P_{\text{trust,M}})-NP_{\text{trust,L}},$$

$$\sigma^2 = \overline{m}NP_{\text{trust,M}}(1-P_{\text{trust,M}})+(1-\overline{m})NP_{\text{trust,L}}(1-P_{\text{trust,L}}),$$

respectively. The mean is found using the linearity of expectation, and the variance is found by utilizing the fact that the binomial random variables K_{M} and K_{L} are statistically independent, given $P_{\text{trust,L}}$, $P_{\text{trust,M}}$, $\overline{m}$, and N. Then, we can approximate the probability $\Pr(Z > z)$ using the complement distribution function $Q(g) = \frac{1}{\sqrt{2\pi}}\int_g^\infty \exp -(u^2/2)du$ where $g = \frac{z-\mu}{\sigma}$. Utilizing this, we have

$$\begin{aligned}\Pr(Z>0) &\approx Q\left(\frac{-\mu}{\sigma}\right)\\ &= Q\left(\frac{NP_{\text{trust,L}}-\overline{m}N(P_{\text{trust,L}}+P_{\text{trust,M}})}{\sqrt{\overline{m}NP_{\text{trust,M}}(1-P_{\text{trust,M}})+(1-\overline{m})NP_{\text{trust,L}}(1-P_{\text{trust,L}})}}\right),\end{aligned} \tag{34}$$

for the probability of missed detection. The probability of false alarm can be upper bounded by $\Pr(Z > -1/2)$ using the continuity correction [48, Ch 4] and computed similarly.

1) Simulation study for m^*: We conclude this section by running a simple simulation study where we compute m^* by varying $\overline{m}$ from 0 to 1 and choosing the first value such that $\overline{P}_{\mathrm{e}}(\hat{\gamma}_t, \hat{p}_t, \overline{m}, 1, 1) \geq \min\{\Pr(\Xi=0), \Pr(\Xi=1)\}$. We also compare this to an estimation of m^*, denoted by $\widehat{m^*}$, done by approximating $P_{\text{MD}}(S_{\text{N}} < \gamma_{\text{TS}} | \mathcal{H}_1)$ by $\Pr(Z > 0)$ in (34) and $P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0)$ by $\Pr(Z > -1/2)$. We compare the results for a case where we set $N = 50$ and $\Pr(\Xi = 0) = \Pr(\Xi = 1) = 0.5$, and vary $P_{\text{trust,L}}(\gamma_t, p_t) \in [0.1, 0.9]$ with $P_{\text{trust,M}}(\gamma_t, p_t) = 1 - P_{\text{trust,L}}(\gamma_t, p_t)$. From Fig. 2 it can be seen that our method of using the normal approximation to estimate m^* closely matches the true value. It can also be seen that a fairly linear relationship exists between the probability of trusting legitimate and malicious robots and the critical proportion m^*. Moreover, for the special case where $\Pr(\Xi=0) = \Pr(\Xi=1)$ this relationship can be estimated by

$$m^* \approx \frac{P_{\text{trust,L}}(\gamma_t, p_t)}{P_{\text{trust,L}}(\gamma_t, p_t)+P_{\text{trust,M}}(\gamma_t, p_t)}.$$

V. Adversarial Generalized Likelihood Ratio Test

In this section, we introduce our second approach, called the Adversarial Generalized Likelihood Ratio Test (A-GLRT). The A-GLRT uses both the trust values and measurements simultaneously to arrive at a final decision while estimating the unknown parameters using the maximum likelihood estimation rule. The A-GLRT approach addresses Problem 2.

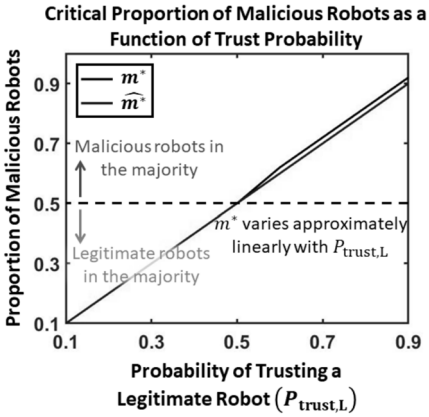

Fig. 2: Case study to compare m^* with $\widehat{m^*}$, an estimate of m^* using the normal approximation. The estimate closely matches m^*, and there is a linear relationship between the probability of trusting legitimate robots and m^*.

A. A-GLRT Algorithm

The main purpose of this section is to construct an efficient algorithm that implements the GLRT in (8). We can simplify (8) by recalling that given the true trustworthiness of a robot t_i and the true hypothesis $\mathcal{H}$, the trust value α_i and the measurement Y_i are statistically independent. Thus,

$$\begin{aligned}&\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})\\ &\quad=\Pr(\boldsymbol{a}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})\Pr(\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}}),\\ &\Pr(\boldsymbol{a},\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})\\ &\quad=\Pr(\boldsymbol{a}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})\Pr(\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}}).\end{aligned}$$

Furthermore, the trust value α_i is independent of the true hypothesis $\mathcal{H}$. Thus,

$$\Pr(\boldsymbol{a}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})=\Pr(\boldsymbol{a}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})=\Pr(\boldsymbol{a}|\boldsymbol{t}).$$

Hence, we obtain

$$\frac{\max\limits_{\boldsymbol{t}\in\{0,1\}^N, P_{\text{MD,M}}\in[0,1]} \Pr(\boldsymbol{a}|\boldsymbol{t})\Pr(\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}})}{\max\limits_{\boldsymbol{t}\in\{0,1\}^N, P_{\text{FA,M}}\in[0,1]} \Pr(\boldsymbol{a}|\boldsymbol{t})\Pr(\boldsymbol{y}|\mathcal{H}_0,\boldsymbol{t},P_{\text{FA,M}})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma_{\text{GLRT}}. \tag{35}$$

We choose $\gamma_{\text{GLRT}} = \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$ since we do not assume that we have the prior distribution of $\boldsymbol{t}$. In Example 1, using the GLRT method, the server initially assumes traffic presence on a roadway (the hypothesis $\mathcal{H}_1$) and determines the strategy of malicious robots and the trustworthiness of all robots that maximize the likelihood of the observed trust information and robot measurements. Subsequently, the server performs a similar calculation assuming no traffic on the roadway (the hypothesis $\mathcal{H}_0$). The traffic condition is then deduced by comparing the likelihoods calculated under both assumptions.

The challenging part of using the GLRT in this problem is calculating the MLEs for both the numerator and denominator. The unknown $\boldsymbol{t}$ is a discrete multidimensional variable while $P_{\text{MD,M}}$ and $P_{\text{FA,M}}$ are continuous variables restricted to the domain $[0,1]$. Therefore, calculating the MLE is not trivial. Due to symmetry in the calculation of the numerator and denominator in (35), we focus only on the numerator.

Using Assumption 1 regarding independence of trust values, we obtain $\Pr(\boldsymbol{a}|\boldsymbol{t}) = \prod_{i=1}^N p_\alpha(a_i|t_i)$. Additionally, we obtain the

following using the i.i.d. assumption about measurements:

$$\Pr(\boldsymbol{y}|\mathcal{H}_1,\boldsymbol{t},P_{\text{MD,M}}) = \prod_{i:t_i=1}(1-P_{\text{MD,L}})^{y_i}\cdot P_{\text{MD,L}}^{1-y_i} \cdot \prod_{i:t_i=0}(1-P_{\text{MD,M}})^{y_i}\cdot P_{\text{MD,M}}^{1-y_i}.$$

Using these equations, we write the numerator as:

$$\max_{\boldsymbol{t}\in\{0,1\}^N, P_{\text{MD,M}}\in[0,1]} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \right\}. \tag{36}$$

Since there is no clear way to optimize (36) over variables $\boldsymbol{t}$ and $P_{\text{MD,M}}$ at the same time, we reformulate the problem as two nested optimizations using the Principle of Iterated Suprema [49, p. 515], that is:

$$\sup\{f(z,w): z\in\mathcal{Z}, w\in\mathcal{W}\} = \sup_{z\in\mathcal{Z}}\{\sup_{w\in\mathcal{W}}\{f(z,w)\}\} = \sup_{w\in\mathcal{W}}\{\sup_{z\in\mathcal{Z}}\{f(z,w)\}\},$$

where $f\colon \mathcal{Z}\times\mathcal{W}\to\mathbb{R}$, and $\mathcal{Z},\mathcal{W}\subseteq\mathbb{R}^d$. By the Principle of Iterated Suprema we can calculate the maximization in (36) in two ways. We rewrite the maximization problem as:

$$\max_{\boldsymbol{t}\in\{0,1\}^N} \left\{ \max_{P_{\text{MD,M}}\in[0,1]} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \right\} \right\}.$$

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 calculate the inner maximization in the following lemma.

Lemma 6. *Let $\boldsymbol{t}$ and $\boldsymbol{y}$ be given vectors in $\{0,1\}^N$. Assume that $p_\alpha(a_i|t_i)$ is known both $t_i=0$ and $t_i=1$, and that $\sum_{i:t_i=0}1>0$. Then,*

$$\prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \tag{37}$$

is maximized by $\widehat{P}_{\text{MD,M}} = \frac{\sum_{i:t_i=0}(1-y_i)}{\sum_{i:t_i=0}1}$. Additionally, if $\sum_{i:t_i=0}1=0$, i.e., $|\{i: t_i=0\}|=0$, any choice $\widehat{P}_{\text{MD,M}}\in[0,1]$ maximizes (37).

Proof. First, observe that given the vector $\boldsymbol{t}$, (37) is maximized by the MLE of $\prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}$. Furthermore, since

$$\prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} = \left(\prod_{i:t_i=0} p_\alpha(a_i|t_i)\right)\left(\prod_{i:t_i=0} P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}\right),$$

it follows that (37) is maximized by the MLE of $\prod_{i:t_i=0} P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}$. This is a well-known estimation problem [50, Problem 7.8], that together with the invariance property of the MLE [50, Theorem 7.2] leads to the optimal estimator $\widehat{P}_{\text{MD,M}} = \frac{\sum_{i:t_i=0}(1-y_i)}{\sum_{i:t_i=0}1}$. Note, that this estimator is equal to the empirical missed detection probability of the measurements sent by the malicious robots. Finally, it is easy to validate that if $|\{i: t_i=0\}|=0$, any $\widehat{P}_{\text{MD,M}}\in[0,1]$ maximizes (37). □

Unfortunately, since the set $\{0,1\}^N$ grows exponentially with the number of robots in the network, this approach is computationally intractable for large robot networks. Therefore, we look for an alternative solution. Another equivalent formulation of the maximization problem in (36) that is obtained by the Principle of Iterated Supremum is

$$\max_{P_{\text{MD,M}}\in[0,1]} \left\{ \max_{\boldsymbol{t}\in\{0,1\}^N} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \right\} \right\}, \tag{38}$$

where the order of variables that the maximization is taken over is flipped. Since the variable $P_{\text{MD,M}}$ belongs to an uncountably infinite set, it is impossible to perform the maximization with this formulation. However, assuming that we have a given $P_{\text{MD,M}}$, the inner maximization can still be calculated. The following lemma calculates the inner maximization.

Lemma 7. *Let $P_{\text{MD,M}}$, $\boldsymbol{a}$, and $\mathbf{y}$ be given. Additionally, assume that $p_\alpha(a_i|t_i)$ is known for both $t_i=0$ and $t_i=1$. Let*

$$c_{\text{L},i} = p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i},$$
$$c_{\text{M},i} = p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}.$$

If the estimated robot identity vector $\hat{\mathbf{t}}$ is constructed by choosing $\hat{t}_i=1$ if $c_{\text{L},i}\geq c_{\text{M},i}$ and $\hat{t}_i=0$ otherwise, where $\hat{t}_i$ is the i^{th} component of $\hat{\mathbf{t}}$, then, $\hat{\mathbf{t}}$ is a vector that maximizes the expression (37). *Moreover, maximization with this approach requires $\mathcal{O}(N)$ comparisons.*

Proof. First, we reformulate (37) as:

$$\prod_{i=1}^{N} (p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i})^{t_i} \cdot (p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i})^{1-t_i}, \tag{39}$$

where the product is calculated by going through all robots rather than going through legitimate and malicious robots separately. We define $c_{\text{L},i} = p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}$, and $c_{M,i} = p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}$. Then, (39) becomes:

$$\prod_{i=1}^{N} c_{\text{L},i}^{t_i} \cdot c_{\text{M},i}^{1-t_i}. \tag{40}$$

Let $0\log 0 = 1$, thus $0^0=1$. Then, the expression (40) is maximized when choosing $t_i=1$ if $c_{L,i}\geq c_{M,i}$ and $t_i=0$ otherwise. Since this comparison needs to be performed for every $i\in\mathcal{N}$, $\mathcal{O}(N)$ comparisons need to be performed. □

Now, we consider these two perspectives together to introduce an efficient calculation of the numerator of the GLRT given in (36). By Lemma 6, we can see that the optimum value of $P_{\text{MD,M}}$ has a special structure. Exploiting this knowledge, we can restrict the set that

$P_{\text{MD,M}}$ belongs to in (38). Then, the inner maximization can be calculated using Lemma 7. The following theorem builds on this intuition to provide an efficient calculation of (36). First, we define the set $\mathcal{P}$:

$$\mathcal{P} \triangleq \left\{ \frac{T_n}{T_d} \right\}_{T_n \in \{0,\dots,T_d\}, T_d \in \{1,\dots,N\}}.$$

Theorem 2. *Assume that $(\boldsymbol{t}^*, P^*_{MD,M})$ attains the maximization in* (36)*. Then, for each vector of measurements* $\mathbf{y}$ *and trust values* $\mathbf{a}$*, $P^*_{MD,M}$ belongs to the set $\mathcal{P}$ where $|\mathcal{P}| \leq N^2+1$. Moreover, the maximization in* (36) *can be calculated by iterating over $\mathcal{O}(N^2)$ different values in $\mathcal{P}$ and performing $\mathcal{O}(N)$ comparisons.*

Proof. First, we will approach the problem by rewriting it as (38) using the Principle of Iterated Suprema:

$$\max_{P_{\text{MD,M}} \in [0,1]} \left\{ \max_{\boldsymbol{t} \in \{0,1\}^N} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i} (1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i} (1-P_{\text{MD,M}})^{y_i} \right\} \right\}.$$

By Lemma 7, we can calculate the inner maximization for a given $P_{\text{MD,M}}$. Notice that, since the calculation requires a comparison for each robot, $\mathcal{O}(N)$ comparisons need to be performed for this maximization. Now, consider the other formulation of the problem given by (V-A). From Lemma 6, we can see that the optimal $P_{\text{MD,M}}$ only depends on the number of ones and zeros of malicious robots for a given $\boldsymbol{t}$. Moreover, the permutation of ones and zeros of malicious robots for a given $\boldsymbol{t}$ does not change the optimum; only the total number of ones and zeros does. We will restrict the set that the outer maximization process iterates over in (38) based on this observation. It follows from the Lemma 6 that for each value $\boldsymbol{t}$ in the outer maximization of (V-A), except the case where $\boldsymbol{t}$ consist of all ones, the optimal value of $P_{\text{MD,M}}$ belongs to the set $\mathcal{P}$. Moreover, in the case where $\boldsymbol{t}$ consists of all ones, any choice of $P_{\text{MD,M}}$ maximizes the expression. Hence, without loss of generality, it suffices to look for an optimizer $P_{\text{MD,M}}$ of (V-A) in the set $\mathcal{P}$. Observe that $|\mathcal{P}| \leq N^2+1$. Therefore, there are only $\mathcal{O}(N^2)$ possible values that the optimal $P_{\text{MD,M}}$ can take. Thus, we can reformulate (38) as:

$$\max_{P_{\text{MD,M}} \in \mathcal{P}} \left\{ \max_{\boldsymbol{t} \in \{0,1\}^N} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i} (1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i} (1-P_{\text{MD,M}})^{y_i} \right\} \right\}.$$

Therefore, this maximization can be calculated by iterating over $\mathcal{O}(N^2)$ different values of $P_{\text{MD,M}}$ and for each value, performing $\mathcal{O}(N)$ comparisons. A similar approach can be adapted for calculating the denominator as well. □

Now, using Theorem 2, we introduce the A-GLRT algorithm, which makes a decision based on the GLRT in (35).

Corollary 2.1. *The GLRT given by* (35) *can be calculated by Algorithm 2 which is referred to as the A-GLRT algorithm. The A-GLRT algorithm requires $\mathcal{O}(N^3)$ comparisons.*

Proof. Calculation of the maximization in the numerator can be calculated in $\mathcal{O}(N^2)$ iterations and performing $\mathcal{O}(N)$ comparisons at each iteration as described by Theorem 2. Therefore, it requires $\mathcal{O}(N^3)$ comparisons in total. Similarly, maximization of the denominator requires the same amount of computation and can be calculated in a similar manner using $P_{\text{FA,M}}$ instead of $P_{\text{MD,M}}$. After that, a final comparison is made by comparing the ratio of the numerator and denominator with $\gamma_{\text{GLRT}} = \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$. Algorithm 2 follows these steps, therefore, it requires $\mathcal{O}(N^3)$ comparisons in total. □

Algorithm 2 A-GLRT

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

Set $\mathcal{P} = \left\{ \frac{T_n}{T_d} \right\}_{T_n \in \{0,\dots,T_d\}, T_d \in \{1,\dots,N\}}$.
Set $\gamma_{\text{GLRT}} = \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}$, $l_{\text{num,max}} = 0$, $l_{\text{denom,max}} = 0$.
% Calculate the maximum likelihood estimations.
for all $P_M \in \mathcal{P}$ **do**
    Set $P_{\text{MD,M}} = P_M$, $l_{\text{num}} = 1$.
    for i=1 to N **do**
% Set the likelihoods for robot i according to Lemma 7.
        Set $c_{\text{L},i} = p_\alpha(a_i|t_i=1) P_{\text{MD,L}}^{1-y_i} (1-P_{\text{MD,L}})^{y_i}$.
        Set $c_{\text{M},i} = p_\alpha(a_i|t_i=0) P_{\text{MD,M}}^{(1-y_i)} (1-P_{\text{MD,M}})^{y_i}$.
        Set $l_{\text{num}} = l_{\text{num}} \cdot \max\{c_{\text{L},i}, c_{\text{M},i}\}$
    if $l_{\text{num}} > l_{\text{num,max}}$ **then** Set $l_{\text{num,max}} = l_{\text{num}}$.
    Repeat steps 4-9 for the denominator.
% Perform the standard hypothesis test using the maximum likelihood estimations.
if $\frac{l_{\text{num,max}}}{l_{\text{denom,max}}} > \gamma_{\text{GLRT}}$ **then** Return decision $\mathcal{H}_1$
else Return decision $\mathcal{H}_0$

Finally, we investigate how the measurements $\boldsymbol{y}$ and stochastic trust values $\boldsymbol{\alpha}$ are being used by the A-GLRT algorithm. Considering (39), an equivalent decision rule to the one derived in Lemma 7 is given as:

$$\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)} \underset{\hat{t}_i=0}{\overset{\hat{t}_i=1}{\gtrless}} \frac{P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}}{P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}}. \tag{41}$$

With this new perspective, we can gain more insights about the A-GLRT. First, we can see that the A-GLRT is essentially performing a likelihood ratio test with trust values for each robot to decide if they are legitimate or not, using different threshold values based on the measurement coming from that robot. For now, let's assume that $P_{\text{MD,M}}$ is not 0 or 1. Then, we can see that as trust values become more accurate, meaning that the ratio $\frac{p_\alpha(a_i|t_i=1)}{p_\alpha(a_i|t_i=0)}$ approaches ∞ if $t_i = 1$ or approaches zero otherwise, for all values that α_i can take, the finite threshold value becomes insignificant and the decision is made using trust values only. This situation agrees with intuition since, in this regime, trust values become true indicators of robot identities. In the next section, we formalize this intuition.

B. Behavior of A-GLRT as Quality of Trust Values Improve

In this section, we characterize the behavior of the A-GLRT algorithm as the quality of the trust values increase. For the rest of this section only, we focus on the special case where α_i is a discrete random variable drawn from a Bernoulli distribution:

Assumption 3. *Let the vector $\boldsymbol{t}$ denote the true identities of the robots in the network. We assume that the distribution of α_i when*

robot i is legitimate is the Bernoulli distribution with probability $1-p_{e,1}$,

$$p_\alpha(a_i|t_i=1) \sim \textit{Bernoulli}(1-p_{e,1}).$$

Similarly, we assume that the distribution given that robot i is malicious is the Bernoulli distribution with probability $p_{e,0}$,

$$p_\alpha(a_i|t_i=0) \sim \textit{Bernoulli}(p_{e,0}).$$

Under this assumption, we are interested in the case where the conditional expectation that $i \in \mathcal{M}$ approaches 0 such that $\mathbb{E}[\alpha_i|t_i=0] \to 0$, and the conditional in the case that $i \in \mathcal{L}$ approaches 1 such that $\mathbb{E}[\alpha_i|t_i=1] \to 1$. Since the expected value of Bernoulli(p) is equal to p, a direct implication of this limit behavior and Assumption 3 is that both $p_{e,1}$ and $p_{e,0}$ approach 0. Let $(\boldsymbol{t_n}^*, P^*_{MD,M})$ and $(\boldsymbol{t_d}^*, P^*_{FA,M})$ be maximizers of the numerator and denominator in (35)

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

respectively: We want to show that both $\boldsymbol{t_n}^*$ and $\boldsymbol{t_d}^*$ are equal to the true trust vector $\boldsymbol{t}$ with high probability. Moreover, when these estimated vectors are equal to each other, i.e., $\boldsymbol{t_n}^* = \boldsymbol{t_d}^*$, A-GLRT is equivalent to the likelihood ratio test using the measurements of legitimate robots only.

Lemma 8. *Assume that Assumption 3 holds. Let $(\boldsymbol{t_n}^*, P^*_{MD,M})$ and $(\boldsymbol{t_d}^*, P^*_{FA,M})$ be maximizers of the numerator and denominator in the GLRT decision rule* (35)*, respectively. Then, both $\boldsymbol{t_n}^*$ and $\boldsymbol{t_d}^*$ are equal to $\boldsymbol{t}$ with high probability given that $\mathbb{E}[\alpha_i|t_i=0] \to 0$ for every $i \in \mathcal{M}$ and $\mathbb{E}[\alpha_i|t_i=1] \to 1$ for every $i \in \mathcal{L}$.*

Proof. We show the proof only for the numerator for conciseness. However, a symmetric argument applies to the denominator as well. Moreover, we drop the subscript in $\boldsymbol{t_n}^*$ for readability and instead we denote it with $\boldsymbol{t}^*$. We want to show that the probability $\Pr(\boldsymbol{t}^* \neq \boldsymbol{t}|\mathcal{H}_1, \boldsymbol{t}, P^*_{\text{MD,M}})$ goes to zero as $\mathbb{E}[\alpha_i|t_i=0] \to 0$ and $\mathbb{E}[\alpha_i|t_i=1] \to 1$. Our strategy is to split this probability into two cases using the law of total probability: the first case is the case where the vector of trust values does not match the true trust vector, i.e., $\boldsymbol{a} \neq \boldsymbol{t}$ and the second case where $\boldsymbol{a} = \boldsymbol{t}$. The intuition is that the probability of the first case goes to zero, and $\boldsymbol{t}^*$ will be equal to $\boldsymbol{t}$ with high probability in the second case since $p_{e,1}, p_{e,0} \to 0$. Now, we will show this formally. We have that

$$\begin{aligned}&\Pr(\boldsymbol{t}^* \neq \boldsymbol{t}|\mathcal{H}_1, \boldsymbol{t}, P^*_{\text{MD,M}})\\ &= \Pr(\boldsymbol{t}^* \neq \boldsymbol{t}|\boldsymbol{a} \neq \boldsymbol{t}, \boldsymbol{t}, \mathcal{H}_1, P^*_{\text{MD,M}})\Pr(\boldsymbol{a} \neq \boldsymbol{t}|\boldsymbol{t})\\ &\quad + \Pr(\boldsymbol{t}^* \neq \boldsymbol{t}|\boldsymbol{a} = \boldsymbol{t}, \boldsymbol{t}, \mathcal{H}_1, P^*_{\text{MD,M}})\Pr(\boldsymbol{a} = \boldsymbol{t}|\boldsymbol{t}).\end{aligned} \tag{42}$$

We can bound the probability $\Pr(\boldsymbol{a} \neq \boldsymbol{t}|\boldsymbol{t})$ as

$$\begin{aligned}\Pr(\boldsymbol{a} \neq \boldsymbol{t}|\boldsymbol{t}) &= \Pr\left(\bigcup_{i\in\mathcal{L}\cup\mathcal{M}} \{a_i \neq t_i\}|\boldsymbol{t}\right)\\ &\leq \sum_{i\in\mathcal{N}} \Pr(a_i \neq t_i|t_i) = |\mathcal{M}|p_{e,0} + |\mathcal{L}|p_{e,1}.\end{aligned}$$

Since $p_{e,1}, p_{e,0} \to 0$, $\Pr(\boldsymbol{a} \neq \boldsymbol{t}|\boldsymbol{t})$ goes to 0 and the first term in (42) vanishes. Now let's consider the second term. We want to show that the probability $\Pr(\boldsymbol{t}^* \neq \boldsymbol{t}|\boldsymbol{a} = \boldsymbol{t}, \boldsymbol{t}, \mathcal{H}_1, P^*_{\text{MD,M}})$ goes to 0. For contradiction, assume that $\boldsymbol{t}^* \neq \boldsymbol{t}$. Remember that $(\boldsymbol{t}^*, P^*_{\text{MD,M}})$ maximizes the numerator by definition. The numerator is calculated as:

$$\begin{aligned}&\prod_{i:t_i=1} p_\alpha(a_i|t_i^*) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}\\ &\qquad \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i^*) P_{\text{MD,M}}^{*1-y_i}(1-P^*_{\text{MD,M}})^{y_i}.\end{aligned}$$

Since $\boldsymbol{t}^* \neq \boldsymbol{t}$ and $\boldsymbol{a} = \boldsymbol{t}$, we have

$$\begin{aligned}&\prod_{i:t_i^*=1} p_\alpha(a_i|t_i^*) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}\\ &\quad \cdot \prod_{i:t_i^*=0} p_\alpha(a_i|t_i^*) P_{\text{MD,M}}^{*1-y_i}(1-P^*_{\text{MD,M}})^{y_i} \leq \max(p_{e,1}, p_{e,0}).\end{aligned}$$

We will show that there exist a pair of estimators different than $(\boldsymbol{t}^*, P^*_{\text{MD,M}})$ that results in a larger numerator. Let $(\boldsymbol{t}, \hat{P}_{\text{MD,M}})$ be another pair of estimators for the numerator where $\hat{P}_{\text{MD,M}} = 0.5$. Here, $\hat{P}_{\text{MD,M}} = 0.5$ is an arbitrary choice to simplify the calculations. Using this pair of estimators, the numerator is

$$\begin{aligned}&\prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}\\ &\qquad \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) \hat{P}_{\text{MD,M}}^{1-y_i}(1-\hat{P}_{\text{MD,M}})^{y_i}\\ &= \left(\frac{1-p_{e,0}}{2}\right)^{|\mathcal{M}|} \cdot \prod_{i:t_i=1} (1-p_{e,1}) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}.\end{aligned}$$

Since $p_{e,1}, p_{e,0} \to 0$, we have

$$\begin{aligned}&\prod_{i:t_i^*=1} p_\alpha(a_i|t_i^*) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}\\ &\quad \cdot \prod_{i:t_i^*=0} p_\alpha(a_i|t_i^*) P_{\text{MD,M}}^{*1-y_i}(1-P^*_{\text{MD,M}})^{y_i} \leq \max(p_{e,1}, p_{e,0})\\ &< \left(\frac{1-p_{e,0}}{2}\right)^{|\mathcal{M}|} \cdot \prod_{i:t_i=1} (1-p_{e,1}) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}.\end{aligned}$$

Therefore, $(\boldsymbol{t}, \hat{P}_{\text{MD,M}})$ results in a larger numerator than $(\boldsymbol{t}^*, P^*_{\text{MD,M}})$ where $\boldsymbol{t}^* \neq \boldsymbol{t}$ and $\boldsymbol{a} = \boldsymbol{t}$, which means that $\boldsymbol{t}^*$ cannot be the maximizer. Hence, the event $\boldsymbol{t}^* \neq \boldsymbol{t}$ in this case has probability 0, which concludes our proof. □

Now, we can state the main result of this section.

Proposition 2. *Assume that Assumption 3 holds. Let $(\boldsymbol{t_n}^*, P^*_{MD,M})$ and $(\boldsymbol{t_d}^*, P^*_{FA,M})$ be maximizers of the numerator and denominator in* (35)*, respectively. If $\mathbb{E}[a_i|t_i=0] \to 0$ for every $i \in \mathcal{M}$ and $\mathbb{E}[a_i|t_i=1] \to 1$ for every $i \in \mathcal{L}$, then, with high probability, the A-GLRT algorithm is equivalent to the likelihood ratio test using the measurements of legitimate robots only, that is*

$$\frac{\prod_{i\in\mathcal{L}} P_{\text{FA,L}}^{y_i} \cdot (1-P_{\text{FA,L}})^{1-y_i}}{\prod_{i\in\mathcal{L}} (1-P_{\text{MD,L}})^{y_i} \cdot P_{\text{MD,L}}^{1-y_i}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\Pr(\Xi=0)}{\Pr(\Xi=1)}. \tag{43}$$

Proof. By Lemma 8, we have $\boldsymbol{t_n}^* = \boldsymbol{t_d}^* = \boldsymbol{t}$ with high probability. We use $\boldsymbol{t}$ in place of both $\boldsymbol{t_n}^*$ and $\boldsymbol{t_d}^*$ for simplicity in the rest of the proof. First, in the trivial case where $\sum_{i:t_i=0} 1 = 0$, the GLRT has the form (43) because there are no malicious robots in the system. In other cases, $P^*_{\text{MD,M}} = \frac{\sum_{i:t_i=0}(1-y_i)}{\sum_{i:t_i=0} 1}$, and $P^*_{\text{FA,M}} = \frac{\sum_{i:t_i=0} y_i}{\sum_{i:t_i=0} 1}$ by Lemma 6. Notice that $P^*_{\text{MD,M}}$ equals $1 - P^*_{\text{FA,M}}$. Therefore, in the calculation of GLRT, the contribution coming from the malicious

robots in the numerator and denominator cancel each other out. As a result, the GLRT has the form (43). Therefore, in all cases, the GLRT has the form (43) with high probability. □

C. Utilizing the Prior Knowledge with A-GLRT

In this section, we introduce two modifications of the A-GLRT algorithm to optionally incorporate additional information about the malicious robots into the system if it is available.

1) Probability of Each Robot Being Malicious: In some cases, the probability of each robot being malicious is available or assumed to be known. Essentially, this information would quantify the vulnerability of the multi-robot system, where a higher probability would correspond to a more vulnerable system. For instance, some previous works including [19], [24] have this assumption. Referring back to Example 1, this scenario could correspond to one where the server, having access to historical ground truth data on trustworthiness of robots from a roadway, incorporates this information into its decision rule. In this part, we modify the A-GLRT algorithm to introduce a way to incorporate this additional information. First, we formalize this new assumption. For the analysis of this section only we assume the following:

Assumption 4. *Let t_i denote the true identity of a robot i in the network. We assume that if a prior distribution over t_i is known, given by $\Pr(t_i)$, then it is independent of other robots and known by the FC.*

Under this assumption, we modify the GLRT given by (35):

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

where we can calculate $\Pr(\boldsymbol{a},\boldsymbol{t})$ using $\Pr(\boldsymbol{a},\boldsymbol{t}) = \Pr(\boldsymbol{a}|\boldsymbol{t})\Pr(\boldsymbol{t})$. Now, we focus on how to calculate the numerator with this new formulation since the denominator follows a similar structure. We write the numerator as:

$$\max_{\boldsymbol{t}\in\{0,1\}^N, P_{\text{MD,M}}\in[0,1]} \left\{ \prod_{i:t_i=1} \Pr(a_i,t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} \Pr(a_i,t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \right\}.$$

Notice that this new formulation does not affect the results in Lemma 6. Moreover, let $c_{\text{L},i} \triangleq p_\alpha(a_i|t_i=1)\Pr(t_i=1)P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i}$, and $c_{\text{M},i} \triangleq p_\alpha(a_i|t_i=0)\Pr(t_i=0)P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i}$. With these new definitions, Lemma 7 and Theorem 2 still hold. Therefore, we can still use the A-GLRT algorithm given in Algorithm 2 just by replacing $c_{\text{L},i}$ and $c_{\text{M},i}$ with these new definitions that include $\Pr(t_i)$.

2) An Upper Bound on The Number of Malicious Robots: In this part, we assume that the the upper bound on the proportion of the malicious robots in the network, denoted by $\overline{m}$, is known similar to the case of the 2SA algorithm. We next show how to modify Algorithm 2 to incorporate this additional information. This upper bound can be expressed as $|\mathcal{M}| \triangleq \sum_{i\in\mathcal{N}} 1-t_i \leq \overline{m}N$. First, notice that this new constraint on $\boldsymbol{t}$ does not affect the results in Lemma 6. However, the inner maximization given in expression (38) turns into a constrained optimization problem,

$$\max_{\boldsymbol{t}\in\{0,1\}^N, |\mathcal{M}|\leq\overline{m}N} \left\{ \prod_{i:t_i=1} p_\alpha(a_i|t_i) P_{\text{MD,L}}^{1-y_i}(1-P_{\text{MD,L}})^{y_i} \cdot \prod_{i:t_i=0} p_\alpha(a_i|t_i) P_{\text{MD,M}}^{1-y_i}(1-P_{\text{MD,M}})^{y_i} \right\},$$

for a given $P_{\text{MD,M}}$. We provide Algorithm 3 to calculate this.

Algorithm 3 Input: $\mathbf{y}$, $\mathbf{a}$, $P_{\text{FA,L}}$, $P_{\text{MD,L}}$, $\{\Pr(\Xi)\}_{\Xi=0,1}$, $\{p_\alpha(a_i|t_i)\}_{t_i=0,1}$, N, $\overline{m}$
Output: Estimate $\hat{\boldsymbol{t}}$

Initialize $N\times 1$ vector $\boldsymbol{d}$ arbitrarily.
for i=1 to N **do**
  Set $c_{\text{L},i} = p_\alpha(a_i|t_i=1)P_{\text{MD,L}}^{(1-y_i)}(1-P_{\text{MD,L}})^{y_i}$.
  Set $c_{\text{M},i} = p_\alpha(a_i|t_i=0)P_{\text{MD,M}}^{(1-y_i)}(1-P_{\text{MD,M}})^{y_i}$.
  Set $d_i = c_{\text{M},i} - c_{\text{L},i}$.
Set $\tilde{\boldsymbol{d}}$ = Sorted($\boldsymbol{d}$) ▷ sorting is in descending order
Set $count = 0$
for all $\tilde{d}_j \in \tilde{\boldsymbol{d}}$ **do**
  Set i as the corresponding index of j in the unordered vector $\boldsymbol{d}$
  if $\tilde{d}_j > 0$ and $count < \overline{m}N$ **then**
    Set $\hat{t}_i = 0$, $count = count + 1$
  else Set $\hat{t}_i = 1$
Return $\hat{\boldsymbol{t}}$.

The main difference of Algorithm 3 compared to the unconstrained inner maximization described in Lemma 7 is that it requires sorting. One can use a sorting algorithm which takes $\mathcal{O}(N\log N)$ comparisons such as merge sort [51]. Notice that this additional computation increases the number of comparisons given in Lemma 7 from $\mathcal{O}(N)$ to $\mathcal{O}(N\log N)$.

VI. Hardware Experiment and Numerical Results

In this section we validate our theoretical results using a hardware experiment with robotic vehicles driving on a mock-up road network. In this setting the robots are tasked with reporting the traffic condition of their road segment to a FC, similar to the scenario in Example 1 used throughout the paper. The objective of the malicious robots is to cause the FC to incorrectly perceive the traffic conditions (see Fig. 3). A numerical study further demonstrates the performance with increasing proportions 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. The Oracle benchmark serves as a lower bound on the probability of error. We also benchmark the ***Oblivious FC***, where the FC treats every robot as legitimate, and a ***Baseline Approach*** [24] 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* approaches use the decision rule in (11). Malicious robots perform a Sybil attack where they spoof additional robots into the network. These spoofed entities are not physically present on the roadmap, and thus do not

affect the ground truth traffic conditions. However, they can act to help prevent the FC from making the correct decision by sending additional malicious measurements about the traffic conditions, thus helping malicious robots gain a majority in the network. The spoofed entities choose to send measurements to the FC following the same strategy as malicious robots, i.e., $P_{\text{FA,M}}$ and $P_{\text{MD,M}}$ defined in (2), and their measurements are assumed to be i.i.d. We use the opensource toolbox in [52] to obtain trust values from communicated WiFi signals by analyzing the similarity between different fingerprints to detect spoofed transmissions. The works in [28]–[30] 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 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 traversing one of four possible paths made up of six different road segments. The robots used were GoPiGo differential drive robots from Dexter Industries. 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 detection 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). Malicious robots know the true traffic conditions and report the wrong measurement with probability 0.99, i.e., $P_{\text{FA,M}}=P_{\text{MD,M}}=0.99$. Hypothesis tests were run on each road segment any time at least one robot was present on that segment. All tests were run using MATLAB 2020a on a 2.6GHz Intel Core i7-10750H CPU with 16GB RAM. The entire experiment was run for 15 minutes, with tests run as frequently as the computer could compute them in order to maximize the number of tests. This led to a frequency of 30 hypothesis tests on each road segment per second, for a total of 61233 hypothesis tests carried out (since hypothesis tests were only used on road segments that were currently occupied). 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. The empirical data from the experiment is stated in Table II, where *Baseline1* and *Baseline5* refer to the Baseline Approach from [24] with parameters T and η set to ($T=1$, $\eta = 0.5$) and ($T = 5$, $\eta = 2.5$). We determined the parameters in Table II by first running an experiment without performing hypothesis tests and observing the behavior of the compared to ground truth. The trust values gathered using the toolbox in [52] 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. 4) [2].

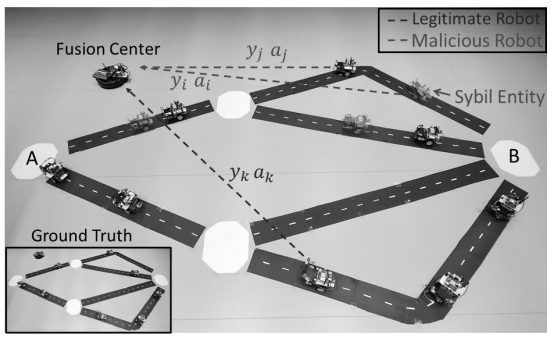

Fig. 3: Robots drive along a roadmap comprised of six road segments to get from point A to point B. While traversing the roadmap, robots estimate the congestion on their current road segment as either containing traffic (red) or not (green), and relay their estimates to the FC. All robots relay messages to the FC, but only a few are depicted on the figure for ease of readability.

In our hardware experiment the 2SA and A-GLRT outperform the Oblivious FC and the Baseline approach. The Baseline exhibits a high percent error due to the fact that it relies on the majority of the network being legitimate. Since 6 out of 11 robots are malicious, it is likely that many hypothesis tests are conducted where the majority are malicious. This points to a common vulnerability of reputation approaches that require legitimate robots to be in the majority.

*a) **Numerical Studies**:* 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. 5. From the plot it can be seen that the 2SA and the A-GLRT perform well even after malicious robots comprise the majority since they use additional trust information independent of the data, whereas the Baseline Approaches (abbreviated with 'B' in the figure) fail since they use only the data to assess the trustworthiness of the robots. Additionally, the existence of the critical proportion of malicious robots, m^*, beyond which the 2SA chooses to ignore all measurements and make the decision using the prior probabilities $\Pr(\Xi=0)$ and $\Pr(\Xi=1)$ can be seen. This value is approximately $m^*=0.8$ for this set of parameters. Finally, we investigate the effect of the malicious robots' strategy on the performance of our approaches. We use the same setup as the previous numerical study,

[2] A link to the code repository containing some of the functions used to run the hypothesis tests and experiment can be found here: https://github.com/mcavorsi/Adversarial_Hypothesis_Testing/tree/main

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. IV-A)	**30.5 %**	**A-GLRT (Sec. V-A)**	**29.0 %**
Oracle	19.5 %	Oblivious FC	52.0 %
Baseline1	50.8 %	Baseline5	49.1 %

TABLE II: EXPERIMENTAL RESULTS

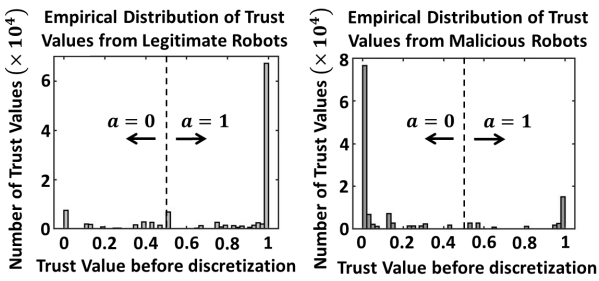

Fig. 4: Empirical distribution of the trust values gathered during the hardware experiment for legitimate and malicious robots. The trust value is thresholded to $a=1$ if it is ≥ 0.5, and $a=0$ otherwise.

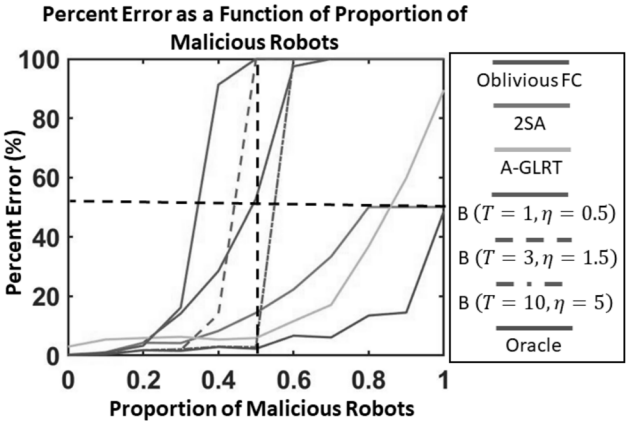

Fig. 5: The percent error for multiple approaches when the proportion of malicious robots is varied. The 2SA and A-GLRT outperform the Oblivious FC and Baseline (B) when the majority of the network is malicious. The performance of the Oracle declines as the proportion of malicious robots in the network increases since the FC is given access to less legitimate information.

except we fix the number of malicious robots to 6 and we vary the probability p_f of flipping their bit. The results are shown in Fig. 6.

VII. Conclusion

In this paper we present two methods to utilize trust values in solving the binary adversarial hypothesis testing problem. The first method, the Two Stage Approach, uses the trust values to determine which robots to trust in the first stage, and then makes a decision from the measurements of the trusted robots using a LRT in the second stage. We show that the probability of error when using the 2SA algorithm is provably minimized under a worst-case attack. Furthermore, we analyze some of the limiting behaviors of the algorithm. For the case where the trust value quality is high, making it more likely to trust a legitimate than malicious robot in the first stage of the 2SA, we show that the probability of error decays towards zero at least exponentially as the number of robots in the network increases. Additionally, we characterize m^*, the critical proportion of malicious robots in the network that would blind the FC using the 2SA. We show that m^* is a function of the quality of the trust values used, and that if the trust value quality is high, m^* can greatly exceed the typical limit that a majority of the network cannot be malicious seen by previous works that do not use trust values [12], [16], [22].

The second method, the Adversarial Generalized Likelihood Ratio Test (A-GLRT), jointly uses the trust values and measurements

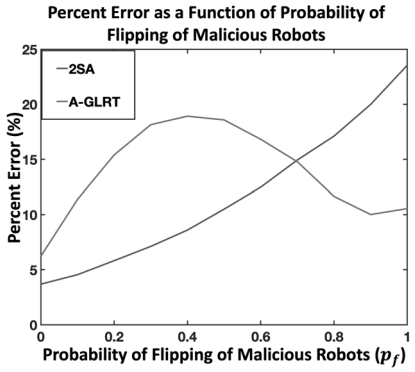

Fig. 6: Percent error comparing the performances of 2SA and A-GLRT as probability of flipping the bit p_f is varied. The worst performance of the 2SA is observed when $p_f = 1$ as proven in Lemma 1. Since the A-GLRT uses both measurements and trust values when making a decision, we see more performance degradation when the measurements have more randomness, i.e., when p_f is close to 0.5.

to estimate the trustworthiness of each robot, the strategy of malicious robots, and the true hypothesis. In general, GLRT-based approaches are computationally expensive since they require the MLE of the unknown parameters. However, with the addition of trust values we show that the A-GLRT algorithm can perform the hypothesis test in polynomial time. Additionally, we show that the A-GLRT algorithm reduces to performing a LRT using the measurements of legitimate robots only when the quality of trust values approaches perfect, i.e., $P_{\text{trust,L}} \to 1$ and $P_{\text{trust,M}} \to 0$, and thus the test in this case is optimal. In comparison to the 2SA, the A-GLRT performs slightly better in numerical studies under a worst-case attack, but at the expense of higher computation costs. The 2SA can perform hypothesis tests in linear time as opposed to polynomial time since the trust thresholds γ_t and p_t can be computed offline and applied repeatedly for each successive hypothesis test as long as the system parameters have not changed.

Finally, we test both methods in a hardware experiment where 61233 hypothesis tests were run by each method on a mock-up road network where robotic vehicles report traffic conditions. The hardware experiment validates our theoretical claims by showing that both methods perform well, even when malicious robots are in the majority. Specifically, the percent error seen by the 2SA and A-GLRT were 30.5% and 29.0%, respectively, compared to the Oblivious FC that does not use trust values, which yielded 52.0% error.

Appendix

A. Proof of Lemma 4

Proof. We start by proving the first part of the lemma, which upper bounds the error probability for Case 4. Recall that k_{L} and k_{M} denote the actual number of legitimate and malicious robots trusted by the FC. Furthermore, recall that $\underline{k_{\text{L}}} = \beta_{\text{L}}|\mathcal{L}|+1$ and $\overline{k_{\text{M}}} = \beta_{\text{M}}|\mathcal{M}|-1$ are the minimum number of legitimate robots within the region $k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|$ and the maximum number of malicious robots within the region $k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|$ that can be trusted, respectively. The probability of error for Case 4 and a given k_{L} and k_{M} is

$$\begin{aligned}\overline{p}_{\text{e}}(k_{\text{L}},k_{\text{M}}) =& \Pr(\Xi=0) P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0, k_{\text{L}}, k_{\text{M}}) \\ &+ \Pr(\Xi=1) P_{\text{MD}}(S_{\text{N}} < \gamma_{\text{TS}} | \mathcal{H}_1, k_{\text{L}}, k_{\text{M}}).\end{aligned}$$

The event probabilities $\Pr(\Xi = 0)$ and $\Pr(\Xi = 1)$ are constant, so in order to upper bound $\overline{p}_{\text{e}}(k_{\text{L}}, k_{\text{M}})$, we look to upper bound $P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0, k_{\text{L}}, k_{\text{M}})$ and $P_{\text{MD}}(S_{\text{N}} < \gamma_{\text{TS}} | \mathcal{H}_1, k_{\text{L}}, k_{\text{M}})$. We will only derive the result for $P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0, k_{\text{L}}, k_{\text{M}})$ since the proof is analogous for $P_{\text{MD}}(S_{\text{N}} < \gamma_{\text{TS}} | \mathcal{H}_1, k_{\text{L}}, k_{\text{M}})$.

From (27) we see that for every k_{L} and k_{M} such that $k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|$ and $k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|$ the following holds:

$$\begin{aligned}&P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0, k_{\text{L}}, k_{\text{M}}) \\ &\leq \max_{\substack{k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|, \\ k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|}} P_{\text{FA}}(S_{\text{N}} \geq \gamma_{\text{TS}} | \mathcal{H}_0, k_{\text{L}}, k_{\text{M}}) \\ &\leq \max_{\substack{k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|, \\ k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|}} \exp\left(-k_{\text{L}} D(\tilde{\gamma}_{\text{FA}}(k_{\text{L}}, k_{\text{M}}) || P_{\text{FA,L}})\right) \\ &\overset{(a)}{\leq} \exp\left(- \min_{\substack{k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|, \\ k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|}} k_{\text{L}} \cdot \min_{\substack{k_{\text{L}} > \beta_{\text{L}}|\mathcal{L}|, \\ k_{\text{M}} < \beta_{\text{M}}|\mathcal{M}|}} D(\tilde{\gamma}_{\text{FA}}(k_{\text{L}}, k_{\text{M}}) || P_{\text{FA,L}}) \right) \\ &\overset{(b)}{\leq} \exp\left(-\underline{k_{\text{L}}} \cdot D\left(\tilde{\gamma}_{\text{FA}}(\underline{k_{\text{L}}}, \overline{k_{\text{M}}}) || P_{\text{FA,L}}\right)\right),\end{aligned}$$

where (a) follows from the nonnegativity of k_{L} and the KL divergence. The inequality (b) follows by minimizing both terms

in the product in (a). The first term is trivially minimized when $k_{\mathrm{L}} = \underline{k_{\mathrm{L}}}$. The KL divergence term attains its minimum at 0 when $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, k_{\mathrm{M}}) = P_{\mathrm{FA,L}}$. When $\tilde{\gamma}_{\mathrm{FA}}(k_{\mathrm{L}}, k_{\mathrm{M}}) \in (P_{\mathrm{FA,L}}, 1)$, the KL divergence is minimized when $\tilde{\gamma}_{\mathrm{FA}}(k_{\mathrm{L}}, k_{\mathrm{M}})$ is minimized, which we show next to be when $k_{\mathrm{L}} = \underline{k_{\mathrm{L}}}$ and $k_{\mathrm{M}} = \overline{k_{\mathrm{M}}}$.

From (28) we see that $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, k_{\mathrm{M}})$ is minimized when k_{M} is maximized, i.e., $k_{\mathrm{M}} = \overline{k_{\mathrm{M}}}$. Now fix $k_{\mathrm{M}} = \overline{k_{\mathrm{M}}}$. Recall the assumption that $\beta_{\mathrm{L}}|\mathcal{L}| >> \max\{\beta_{\mathrm{M}}|\mathcal{M}|, 1\}$ (Assumption 2.b), thus $k_{\mathrm{L}} >> k_{\mathrm{M}}$. Therefore, we can rewrite (28) as

$$\tilde{\gamma}_{\mathrm{FA}}(k_{\mathrm{L}}, k_{\mathrm{M}}) \approx \frac{\gamma_{\mathrm{TS}}}{k_{\mathrm{L}}(w_{0,\mathrm{L}} + w_{1,\mathrm{L}})} + \frac{w_{0,\mathrm{L}}}{(w_{0,\mathrm{L}} + w_{1,\mathrm{L}})}. \tag{44}$$

Since $\Pr(\Xi = 1) > \Pr(\Xi = 0)$ we have that $\gamma_{\mathrm{TS}} < 0$. Therefore, the expression in (44) is minimized when k_{L} is minimized, i.e., $k_{\mathrm{L}} = \underline{k_{\mathrm{L}}}$, as long as $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, \overline{k_{\mathrm{M}}}) \in (P_{\mathrm{FA,L}}, 1)$.

The first part of the lemma shows that worst-case probability of error corresponding to Case 4 can be upper bounded by (30). However, this assumes that $\tilde{\gamma}_{\mathrm{FA}}(k_{\mathrm{L}}, k_{\mathrm{M}}) \in (P_{\mathrm{FA,L}}, 1)$ and $\tilde{\gamma}_{\mathrm{MD}}(k_{\mathrm{L}}, k_{\mathrm{M}}) \in (0, 1 - P_{\mathrm{MD,L}})$. We now proceed to prove the second part of the lemma by showing that the set of values for which $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, \overline{k_{\mathrm{M}}}) \in (P_{\mathrm{FA,L}}, 1)$ is nonempty. The proof is analogous for the missed detection case.

Consider $\underline{k_{\mathrm{L}}} \to \infty$. Notice that $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, \overline{k_{\mathrm{M}}}) \to \frac{w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}}$. In this case, $\tilde{\gamma}_{\mathrm{FA}}(\underline{k_{\mathrm{L}}}, \overline{k_{\mathrm{M}}}) \in (P_{\mathrm{FA,L}}, 1)$ if $\frac{w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} \in (P_{\mathrm{FA,L}}, 1)$. Since $P_{\mathrm{FA,L}} \in (0, 0.5)$ and $P_{\mathrm{MD,L}} \in (0, 0.5)$, and $w_{0,\mathrm{L}}, w_{1,\mathrm{L}} > 0$ we have that $\frac{w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} \in (0, 1)$, thus it remains to show that

$$\frac{w_{0,\mathrm{L}}}{w_{0,\mathrm{L}} + w_{1,\mathrm{L}}} > P_{\mathrm{FA,L}}. \tag{45}$$

We can manipulate (45) by multiplying both sides by $(w_{0,\mathrm{L}} + w_{1,\mathrm{L}})$, plugging in the expressions for $w_{0,\mathrm{L}}$ and $w_{1,\mathrm{L}}$, and using some algebra to yield

$$(1 - P_{\mathrm{FA,L}}) \log\left(\frac{1 - P_{\mathrm{FA,L}}}{P_{\mathrm{MD,L}}}\right) > P_{\mathrm{FA,L}} \log\left(\frac{1 - P_{\mathrm{MD,L}}}{P_{\mathrm{FA,L}}}\right). \tag{46}$$

Next, note that $1 - \frac{1}{x} \leq \log(x) \leq x - 1$. Then, we can lower bound the LHS of the expression in (46) and upper bound the RHS to give us

$$(1 - P_{\mathrm{FA,L}}) \left(1 - \frac{P_{\mathrm{MD,L}}}{1 - P_{\mathrm{FA,L}}}\right) > P_{\mathrm{FA,L}} \left(\frac{1 - P_{\mathrm{MD,L}}}{P_{\mathrm{FA,L}}} - 1\right).$$

This reduces to $1 > 1$. Therefore, the condition in (45) holds for all cases, except when $1 - \frac{1}{x} = \log(x) = x - 1$. This occurs at $x = 1$, which corresponds to $\frac{1 - P_{\mathrm{FA,L}}}{P_{\mathrm{MD,L}}} = 1$, and $\frac{1 - P_{\mathrm{MD,L}}}{P_{\mathrm{FA,L}}} = 1$. If we restrict the values of $P_{\mathrm{FA,L}}$ and $P_{\mathrm{MD,L}}$ to $(0, 0.5]$ then this corresponds to $P_{\mathrm{FA,L}} = P_{\mathrm{MD,L}} = 0.5$. Since $P_{\mathrm{FA,L}}$ and $P_{\mathrm{MD,L}}$ are bounded away from 0.5, the condition in (45) holds for all $P_{\mathrm{FA,L}} \in (0, 0.5)$ and $P_{\mathrm{MD,L}} \in (0, 0.5)$. □

REFERENCES

[1] M. Cavorsi, O. E. Akgun, M. Yemini, A. Goldsmith, and S. Gil, "Exploiting trust for resilient hypothesis testing with malicious robots," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.
[31] M. Yemini, A. Nedić, S. Gil, and A. J. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 4185–4192.
[32] M. Yemini, A. Nedić, A. Goldsmith, and S. Gil, "Resilient distributed optimization for multi-agent cyberphysical systems," *arXiv:2212.02459*, 2022.
[33] 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.
[34] S. M. Kay, *Fundamentals of statistical signal processing: Detection theory*. Prentice Hall PTR, 2008.
[35] 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.
[36] P. K. Varshney, *Distributed detection and data fusion*. Springer Science & Business Media, 2012.
[37] "Classical detection and estimation theory," in *Detection, Estimation, and Modulation Theory*. New York, USA: John Wiley & Sons, Inc, 2001, pp. 19–165.
[38] W. Hashlamoun, S. Brahma, and P. K. Varshney, "Audit bit based distributed bayesian detection in the presence of byzantines," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 4, pp. 643–655, 2018.
[39] C. Pippin and H. Christensen, "Trust modeling in multi-robot patrolling," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 59–66.
[40] W. Teacy, J. Patel, N. R. Jennings, and M. Luck, "Travos: Trust and reputation in the context of inaccurate information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 2, pp. 183–198, 2006.
[41] M. Abdelhakim, L. E. Lightfoot, J. Ren, and T. Li, "Distributed detection in mobile access wireless sensor networks under byzantine attacks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 4, pp. 950–959, 2014.
[42] A. Abrardo, M. Barni, K. Kallas, and B. Tondi, "A game-theoretic framework for optimum decision fusion in the presence of byzantines," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1333–1345, 2016.
[43] E. Soltanmohammadi and M. Naraghi-Pour, "Fast detection of malicious behavior in cooperative spectrum sensing," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 377–386, 2014.
[44] 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.
[45] 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.
[46] R. Arratia and L. Gordon, "Tutorial on large deviations for the binomial distribution," *Bulletin of mathematical biology*, vol. 51, no. 1, pp. 125–131, 1989.
[47] F. Clarke and Y. S. Ledyaev, "Mean value inequalities," *Proceedings of the American Mathematical Society*, pp. 1075–1083, 1994.
[48] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 2015.
[49] J. M. H. Olmsted, *Real variables: An introduction to the theory of functions*. Appleton-Century-Crofts, 1959.
[50] S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice-Hall PTR, 1993.
[51] D. E. Knuth, *The art of computer programming: Volume 3: Sorting and Searching*. Addison-Wesley Professional, 1998.
[52] 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.

Matthew Cavorsi received the Ph.D. degree in Electrical Engineering in 2023 from the School of Engineering and Applied Sciences at Harvard University, advised by Prof. Stephanie Gil. His thesis focus was on multi-robot coordination and resilience to adversaries in multi-robot systems. He was nominated for the Best Paper Award at the 2022 Robotics: Science and Systems (RSS) conference. He received his Bachelor's degree in Aerospace Engineering from the Pennsylvania State University in 2017.

Orhan Eren Akgün is a Computer Science Ph.D. student in the School of Engineering and Applied Sciences at Harvard University, advised by Prof. Stephanie Gil. His research focuses on the development of resilient algorithms to counter adversaries in networked multi-agent systems, specifically within the domain of multi-robot systems. He received his Bachelor's degree in Electrical and Electronics Engineering from the Bogazici University in 2021.

Michal Yemini is an assistant professor at Bar-Ilan University, Ramat-Gan, Israel. Prior to that, she was an associate research scholar at Princeton University, a postdoctoral researcher at Stanford University, Stanford, USA, and a visiting postdoctoral researcher at Princeton University. Her main research interests include distributed optimization, sequential decision-making, learning theory, information theory, and percolation theory. She received the Eric and Wendy Schmidt Postdoctoral Award for Women in Mathematical and Computing Sciences, the Council of Higher Education's Postdoctoral Fellowships Program for Outstanding Women in Science, and the Bar-Ilan University's Postdoctoral Fellowship for Women. She obtained her BSc in computer engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2011. In 2017 she received her PhD degree in the joint MSc-PhD program from the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel.

Andrea J. Goldsmith (Fellow, IEEE) is the Dean of Engineering and Applied Science and the Arthur LeGrand Doty Professor of Electrical Engineering at Princeton University. She was previously the Stephen Harris Professor of Engineering and Professor of Electrical Engineering at Stanford University. Her research interests are in information theory, communication theory, and signal processing, and their application to wireless communications, interconnected systems, and neuroscience. She founded and served as Chief Technical Officer of Plume WiFi and of Quantenna (QTNA), and she serves on the Board of Directors for Intel (INTC), Medtronic (MDT), and Crown Castle Inc (CCI). She also serves on the Presidential Council of Advisors on Science and Technology (PCAST) and as the founding Chair of the IEEE Board of Directors Committee on Diversity, Inclusion, and Equity. Dr. Goldsmith is a member of the National Academy of Engineering, the Royal Academy of Engineering, and the American Academy of Arts and Sciences. Her awards include the Marconi Prize, the IEEE Education Medal, the IEEE Sumner Technical Field Award, the ACM Athena Lecturer Award, the ComSoc Armstrong Technical Achievement Award, the Kirchmayer Graduate Teaching Award, and the WICE Mentoring Award. She is author of four books on wireless communications as well as an inventor on 29 patents. She is currently the Founding Chair of the IEEE Board of Directors Committee on Diversity, Inclusion, and Ethics. She served as the President for the IEEE Information Theory Society in 2009, as the Founding Chair for its student committee, and as the Founding Editor-in-Chief for the IEEE Journal on Selected Areas of Information Theory. She has also served on the Board of Governors for both the IEEE Information Theory and Communications Societies.

Stephanie Gil is an Assistant Professor in the Computer Science Department at the School of Engineering and Applied Sciences at Harvard University where she directs the Robotics, Embedded Autonomy and Communication Theory (REACT) Lab. Prior she was an Assistant Professor at Arizona State University. Her research focuses on multi-robot systems where she studies the impact of information exchange and communication on resilience and trusted coordination. She is the recipient of the 2019 Faculty Early Career Development Program Award from the National Science Foundation (NSF CAREER), the Office of Naval Research Young Investigator Program (ONR YIP) recipient, and has been selected as a 2020 Alfred P. Sloan Fellow. She obtained her PhD from the Massachusetts Institute of Technology in 2014.