Characterizing Trust and Resilience in Distributed Consensus for Cyberphysical Systems

Michal Yemini, Angelia Nedić, Andrea Goldsmith, Stephanie Gil

Abstract—This work considers the problem of resilient consensus where stochastic values of trust between agents are available. Specifically, we derive a unified mathematical framework to characterize convergence, deviation of the consensus from the true consensus value, and expected convergence rate, when there exists additional information of trust between agents. We show that under certain conditions on the stochastic trust values and consensus protocol: 1) almost sure convergence to a common limit value is possible even when malicious agents constitute more than half of the network connectivity, 2) the deviation of the converged limit, from the case where there is no attack, i.e., the true consensus value, can be bounded with probability that approaches 1 exponentially, and 3) correct classification of malicious and legitimate agents can be attained in finite time almost surely. Further, the expected convergence rate decays exponentially with the quality of the trust observations between agents.

I. INTRODUCTION

In this paper we address the problem of multi-agent distributed consensus in the presence of malicious agents. The consensus problem is of core importance to many algorithms and coordinated behaviors in multi-agent systems. It is well-known however, that these algorithms are vulnerable to malicious activity and/or non-cooperating agents and that several of the existing performance guarantees for the nominal case fail in the absence of cooperation.

Many works have investigated the possibility of attaining resilient consensus, or consensus in the face of non-cooperating or malicious agents. However, achieving resilience often relies on conservative assumptions of connectivity, and bounds on the maximum number of malicious actors. A classical result for consensus in networks with non-cooperating agents (faulty or malicious) holds from [1], [6]:

Well behaving agents can always agree upon a parameter if and only if the number of malicious agents is less than 1/2 of the network connectivity.

This result has long been held as a fundamental requirement for achieving consensus in unreliable networks [1], [6], [8].

M. Yemini and A. Goldsmith are with the Electrical and Computer Engineering Department, Princeton University, Princeton, NJ, 08544 USA, e-mails: myemini@princeton.edu; goldsmith@princeton.edu. A. Nedić is with Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, 85281 USA, email: Angelia.Nedich@asu.edu. S. Gil is with the Computer Science Department at the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02139 USA, email: sgil@seas.harvard.edu.

For purely cyber systems, where data is used to validate information, these limitations are difficult to circumvent. Cyberphysical systems (CPS) however, are becoming prevalent in many fields, from robotics to power systems to traffic systems and beyond [9]. These systems provide new opportunities to additionally use physical channels of information for data validation. This physical channel data can be exploited to understand “trust”, or the trustworthiness of data, among the agents in the system. Examples include using observers to compare against system models [10], [11], using camera or other physical channel data [12], or using transmitted communication channels by comparing against a known carrier signal to find whether a message has been manipulated [13], [14]. Wireless channels are of particular interest since they often exist in multi-agent systems as a medium for information exchange and can also be used to extract useful information such as, for example, tracking motion [15], [16], localizing robot agents [17], [18], as well as providing security to wireless systems [19].

Our recent work in [20] derives the characterization of a stochastic value of trust, \( \alpha_{ij} \in (0, 1) \), that approaches 1 for trusted transmission and 0 for an untrusted transmission between any two agents \( i \) and \( j \) in the case of a Sybil Attack. Importantly, these recent results show that information capturing the trustworthiness of communicating agents can be obtained from analyzing the communication signals themselves - alluding to the fact that the probability of trust between agents can be characterized using information that already exists in multi-agent networks.

However, there remains a need for a unified mathematical framework to guide how a probabilistic understanding of trust between agents can be used to arrive at important performance guarantees for resilient coordination and consensus.

We are motivated by these recent works to derive consensus results that take trust into account. This paper provides a unified framework that characterizes the value of trust when it comes to achieving resilient consensus in adversarial settings. We generalize the concept of trust between agents by considering the availability of stochastic observations \( \alpha_{ij} \) that can be thought of as a probability of trust between any two agents \( i \) and \( j \) in a system (see Fig. 1). We show that using this additional information can lead to strong results for resilience. Namely, under certain conditions on \( \alpha_{ij} \) that we characterize, we establish the following results:

1) Convergence: We show that convergence to a common limit value is possible almost surely, even when the number of malicious agents is larger than less than 1/2 of the network connectivity.
The studied resilient consensus protocol: 
\[ x_i(t + 1) = w_{ij}(t)x_j(t) + \sum_{j \in \mathcal{N}} W_{ij}(t)x_j(t) \]
\[ w_{ij}(t) = f(a_{ij}(0), \ldots, a_{ij}(t)) \]

Figure 1. A schematic of a multi-robot system with a consensus protocol modified to use agent transmitted values \(x_i\) and \(x_j\) as well as observations of inter-agent trust \(\alpha_{ij}\).

2) Deviation from nominal average consensus: Whereas without additional information of trust between agents the presence of malicious agents comprising more than 1/2 of the network connectivity could take the consensus value to any limit, we show that by using \(\alpha_{ij}\)'s, the influence of malicious agents can be bounded and that this bound can be characterized.

3) Convergence rate: We show that convergence can be attained in finite time with arbitrarily high probability, and that the expected convergence rate decays exponentially with the quality of the observations \(\alpha_{ij}\) and other problem parameters that we characterize.

A roadmap of our paper is as follows: We begin by providing some background and context for the resilient consensus problem in Section I. In Section II we present our consensus model, our model of trust between agents and the \(\alpha_{ij}\) observations, and we characterize the influence of malicious agents on multi-agent consensus. In Section III we provide our theoretical results and analysis of convergence, deviation, and the convergence rate for consensus with malicious agents. Finally, in Section IV we present simulation results, followed by Section V that provides a discussion and future work directions, and a conclusion in Section VI.

A. Related Work

The problem of coordination in multi-agent systems and consensus has a long history \([21]-[24]\). In particular, the consensus problem has enjoyed the development of many fundamental results providing insights into conditions for convergence \([25]-[29]\), convergence rate \([30]-[32]\), and general analysis and implementation \([33], [34]\). Similarly, in the current paper we hope to start building a more general mathematical foundation for resilient consensus using the concept of inter-agent trust.

Resilient consensus refers to consensus in the case where agents are non-cooperative either due to faulty or malicious behavior. This is an important problem due to the prevalence of the non-cooperative case. For multi-robot systems for example, security is of utmost importance as these robot systems enter the world as autonomous vehicles, delivery drones, or partners in security and defense \([35]-[37]\). For cyberphysical systems the issue of security is a highly investigated topic due to the increasing role that cyberphysical systems play in critical infrastructure such as power grids, traffic management systems, and beyond \([38]-[43]\). As a result, there has been increased attention on exposing the vulnerabilities of consensus in multi-agent systems \([1]-[4]\) and on developing new theory to support resilient consensus for these systems \([5], [6], [8], [10], [44]\).

A common theme in these works is to use transmitted data to identify and thwart attacks, and as a result they are often accompanied by important constraints on the minimum connectivity needed in the graph or requirements on the maximum number of malicious agents that can be tolerated. Take for example the classical result from \([1]\) which states that well behaving agents can always agree upon a parameter if and only if the number of malicious agents comprises less than 1/2 of the network connectivity. Additionally, the work \([3]\) presents conditions for an asymptotic convergence to an arbitrary point in the convex hull of the initial values of the legitimate agents.

Cyberphysical systems, including multi-robot systems as a prominent example, offer more physical channels of information that can be used for data validation and for quantifying “trust” between agents (robots, sensors, autonomous vehicles, and more). Camera data, communicated wireless signals, lidar, and radar, are a few examples of physical channels of information that often exist in cyberphysical systems and can be used to validate data transmitted between agents. Indeed, many works have exploited the physics of cyberphysical systems to validate transmitted data which presents great promise for the security of these systems \([4]-[6], [45], [46]\).

Advances in sensing using wireless signals is especially promising in this domain as many multi-agent systems transmit messages over a wireless medium \([47], [48]\). The wireless community has demonstrated that these wireless signals themselves contain important information that can be fruitfully exploited for security. For example it has been shown that wireless signals can be analyzed to provide tracking and localization of agents \([15], [16], [18], [49]-[51]\), authentication \([19], [52], [53]\), and thwarting of spoofing attacks \([54]\).

Deriving Trust Values using Wireless Signals: Here we provide some intuition on one particular example of quantifying trust between agents by exploiting physical channels of information; specifically, using wireless channels. Our previous work developed a method for measuring directional signal profiles using channel state information (CSI) from the wireless messages over each link \((i,j)\) in the network \([50], [55]\). These profiles measure signal strength arriving from every direction in the 3D plane. Directional signal profiles display two important properties: 1) transmissions originating from the same physical agent have very similar profiles and 2) energy can be measured coming from the direct-line path between physical agents. The paper \([20]\) quantifies these properties, providing an analysis that shows both analytically and experimentally, that a single scalar value \(\alpha_{ij} \in (-0.5, 0.5)\) (shifted by -0.5 from \([20]\)) can be computed for each signal profile that quantifies the likelihood that the transmission is coming from the same physical (spoofed) node, or a unique (legitimate) node; a property critical for thwarting Sybil Attacks. Intuitively, the \(\alpha_{ij}\) was shown experimentally and theoretically to be close to -0.5 if one of the agents \(j\) is a spoofed node and close to 0.5 if both agents are legitimate nodes in the network \([20]\). This is captured quantitatively by the bounds on
the expectation of the $\alpha_{ij}$. The current paper takes inspiration from previous work which shows the existence of such $\alpha_{ij}$ variables for stochastically characterizing trust between agents, but goes beyond to define a unified mathematical framework for using trust to obtain certain performance guarantees for consensus in multi-agent systems. As such, Definition II.2 specifies the mildest characteristics of such as $\alpha_{ij}$ term in order to achieve strong performance guarantees. Thus the results contained herein would be compatible with stochastic variables as derived in [20] or any other physical channels that satisfy the conditions for a probabilistic observation of $\alpha_{ij}$ values to arrive at resilient consensus. Key

Contributions in the context of previous work: This paper builds most closely off of our previous work in [20], [56], which shows that a stochastic variable $\alpha_{ij}(t)$ can be extracted from wireless signals between agents to verify the validity of an agent’s identity (in a probabilistic sense) during a Sybil Attack. Further, [56], [57] shows the potential of using these values to arrive at resilient consensus. In the current work we assume the availability of such probabilistic observations of $\alpha_{ij}$ variables for stochastically characterizing trust between agents, the $\alpha_{ij}$, and focus on the development of the mathematical foundations for why and how such values can be key to achieving certain performance guarantees that would be difficult or impossible to obtain otherwise.

We consider the problem of consensus for multi-agent systems. The agents in the system communicate over a network, which is represented by an undirected graph, $G = (V, E)$, where $V = \{1, \ldots, n\}$ denotes the set of node indices for the agents and $E \subset V \times V$ denotes the set of undirected edges. We use $\{i, j\}$ to denote the edge connecting agents $i$ and $j$. The set of neighbors of node $i$ is denoted by $N_i = \{j \in V | \{i, j\} \in E\}$. Note that by this definition if $i$ is a neighbor of $j$, then $j$ is also a neighbor of $i$. With the graph $G$, we will associate time-varying nonnegative weights $w_{ij}(t)$, $\{i, j\} \in E$, which capture the information exchange process among the agents at a given time. We use $w_{ij}(t)$ to denote the $ij$th entry of the weight matrix $W(t)$. For these matrices, we have $w_{ij}(t) \geq 0$ and $w_{ii}(t) \geq 0$ only if $\{i, j\} \in E$, and $w_{ij}(t) = w_{ji}(t) = 0$ otherwise. In this way, the positive entries of the matrices $W(t)$ are associated only with the edges in the set $E$. Additionally, the matrices $W(t)$ need not be symmetric. We consider the case where a subset of nodes with indices denoted by the set $\mathcal{M}$, $\mathcal{M} \subset V$, are either malicious or malfunctioning and, thus, do not reliably cooperate in the consensus protocol. The set $\mathcal{M}$ is assumed to be unknown, with the cardinality $|\mathcal{M}|$. An agent that is not malicious or malfunctioning is termed legitimate, and the set of legitimate agents is denoted $\mathcal{L}$, $\mathcal{L} \subset V$, with the cardinality $|\mathcal{L}|$.

In what follows, we use $\mathbf{0}$ to denote the zero matrix, where the dimension of $\mathbf{0}$ is to be understood from the context. We say that a matrix $A$ is nonnegative if its entries are nonnegative, i.e., $A_{ij} > 0$ for all $i, j$, and we write $A \succeq \mathbf{0}$ to denote that $A$ is nonnegative. We write $A > \mathbf{0}$ to denote a matrix having all positive entries, i.e., $A_{ij} > 0$ for all $i, j$. A nonnegative matrix is row-stochastic if the sum of its entries in every row is equal to 1, and it is row-substochastic if the sum of its entries in every row is equal to or less than 1. We use $\mathbf{1}$ to denote the vector with all entries equal to 1. Given a vector $x$, we use $x_i$ or $[x]_i$ (when $x$ has a subscript) to denote its $i$th entry. A similar notation is used for the entries of a matrix.

A. The Model

We are concerned with the problem of consensus where each legitimate agent updates its value according to the following update equation (see [3]) for all $t \geq T_0 - 1$,

$$x_i(t + 1) = w_{i'i}(t)x_i(t) + \sum_{j \in \mathcal{N}_i} w_{ij}(t)x_j(t),$$

(1)

where $x_i(t) \in \mathbb{R}$, while the weights $w_{ij}(t)$ are nonnegative and sum to 1, i.e., $w_{ii}(t) > 0$, $w_{ij}(t) \geq 0$ for $j \in \mathcal{N}_i$ and $w_{ii}(t) + \sum_{j \in \mathcal{N}_i} w_{ij}(t) = 1$. The process is initiated at some time $T_0 \geq 0$ with the agents’ initial values $x_i(T_0 - 1) = x_i(0)$ for all $i \in V$.

Definition II.1 (Malicious agent). Agent $i$ is said to be malicious if it does not follow the rule in (1) for updating its value $x_i(t)$ at some time $t \geq T_0 - 1$.

Let $x(t) \in \mathbb{R}^n$ denote the vector of values $x_i(t) \in \mathbb{R}$ for all the agents at time $t$. Without loss of generality, we assume that the agent indices are ordered in such a way that the vector $x(t)$ can be separated in two components; the first component representing the legitimate agent values $x_{\mathcal{L}}(t)$ and the second component representing the malicious agent values $x_{\mathcal{M}}(t)$. In this way, the consensus dynamics takes the following form:
\[
\begin{bmatrix}
x_L(t+1) \\
x_M(t+1)
\end{bmatrix} = \begin{bmatrix}
W_L(t) & W_M(t) \\
\Theta(t) & \Omega(t)
\end{bmatrix}
\begin{bmatrix}
x_L(t) \\
x_M(t)
\end{bmatrix},
\]

where \(W_L(t) \in \mathbb{R}^{|L| \times |L|}\) is the matrix multiplying the state component corresponding to the legitimate agents’ values, and \(W_M(t) \in \mathbb{R}^{|L| \times |M|}\) is the matrix multiplying the state component corresponding to the malicious agents’ values. The matrices \(\Theta(t)\) and \(\Omega(t)\) dictate the dynamics of the malicious agents’ values and are assumed to be unknown. For the sake of simplicity, we assume that a malicious agent sends the same data value to all of its neighbors, however, our analysis holds for the case where a malicious agent can send different data values to different neighbors at each iteration. We consider the case when a parameter \(\eta > 0\) is known to all legitimate agents and it is known that the values \(|x_j(t)|\) should not exceed \(\eta\) at any time. The assumption \(|x_j(t)| \leq \eta\) for all \(t\) and \(j \in \mathcal{M}\) is reasonable, since an agent \(j\) can be classified as malicious based on its value (i.e., if \(|x_j(t)| > \eta\) at some time \(t\)). Thus, in this paper, our focus is on the update rule in (2) where \(|x_j(t)| \leq \eta\) for all \(j \in \mathcal{L} \cup \mathcal{M}\) and time instants \(t\).

In this paper we are interested in the case where each transmission from agent \(j\) to agent \(i\) can be tagged with a random observation \(\alpha_{ij}(t) \in [0, 1]\) of a random variable \(\alpha_{ij} \in [0, 1]\). The random variable \(\alpha_{ij}\) represents a probability of trust that agent \(i\) can give to its neighbor \(j \in \mathcal{N}_i\).

**Definition II.2** (\(\alpha_{ij}\)). For every \(i \in \mathcal{L}\) and \(j \in \mathcal{N}_i\), the random variable \(\alpha_{ij}\) taking values in the interval \([0, 1]\) represents the probability that agent \(j \in \mathcal{N}_i\) is a trustworthy neighbor of agent \(i\). We assume the availability of such observations \(\alpha_{ij}(t)\) throughout the paper.

We refer to [20] for an example of such an \(\alpha_{ij}\) value. Intuitively, a random realization \(\alpha_{ij}(t)\) of \(\alpha_{ij}\) contains useful trust information if the legitimacy of the transmission can be thresholded.

In this paper we assume that a value of \(\alpha_{ij}(t) > 1/2\) indicates a legitimate transmission and \(\alpha_{ij}(t) < 1/2\) indicates a malicious transmission in a stochastic sense (misclassifications are possible). Note that \(\alpha_{ij}(t) = 1/2\) means that the observation is completely ambiguous and contains no useful trust information for the transmission at time \(t\).

Our ultimate goal is to understand the convergence, and convergence rate properties of consensus under the influence of a malicious attack if such observations of trust are available. Toward this goal, our first objective is to construct the appropriate weight matrix \(W(t) = [W_L(t) \ W_M(t)]\) for legitimate agents’ state dynamics in (2) by using some function of the observations \(\alpha_{ij}(t), \alpha_{ij}(t-1), \ldots, \alpha_{ij}(0)\), where \(i\) is a legitimate agent so that \(i \in \mathcal{L}\) and \(j\) is a neighbor of \(i\) so that \(j \in \mathcal{N}_i\). In this way, the legitimate agents’ ability to isolate the malicious agents is reflected in the convergence of the weight matrices \(W_M(t)\) to the zero matrix and the convergence of the weight matrices \(W_L(t)\) to a row-stochastic matrix, where the convergence is in a probabilistic sense (to be defined shortly). Necessary characteristics of \(\alpha_{ij}\) to allow for the convergence of the linear protocol in (2), possibly including its distribution, bounds on expectation, etc., will be derived in this work.

In line with this model, in addition to broadcasting its value \(x_i(t)\) at each time \(t\), we assume that a message from each agent \(i \in \mathcal{V}\) is tagged with an observation \(\alpha_{ij}(t)\) for every neighboring agent \(j\), i.e., \(\{i, j\} \in E\) (see Figure 1). The purpose of this work is to develop the necessary properties of \(\alpha_{ij}(t)\) such that consensus can be maintained in the presence of malicious or malfunctioning agents.

Suppose that, in the interval of time \([0, T_0 - 1]\), the legitimate agents only use their measurement of the trustworthiness values \(\alpha_{ij}, j \in \mathcal{N}_i, i \in \mathcal{L}\). We refer to this interval of time as an observation window where agents can amass a history of trust values \(\alpha_{ij}\) before starting the consensus. At time \(T_0 - 1\), the agents start the data passing phase with the weights \(w_{ij}(T_0 - 1) = f(\alpha_{ij}(0), \ldots, \alpha_{ij}(T_0 - 1)) \cdot \mathbb{1}_{\{j \in \mathcal{N}_i\}},\) where \(f\) is some function capturing the history of all observations \(\alpha_{ij}(0), \ldots, \alpha_{ij}(T_0 - 1)\) and \(\mathbb{1}_{\{\cdot\}}\) denotes the indicator function. Our goal is to study the resulting consensus dynamics over time. Since we start the consensus algorithm (2) from time \(t = T_0 - 1\), we have that \(x_L(t) = x_L(0)\) for all \(t \in [0, T_0 - 1]\).

Let \(\prod_{k=p}^{r} H_k\) denote the backward product of the matrices \(H_p, \ldots, H_r\), i.e.,

\[
\prod_{k=p}^{r} H_k = \begin{cases} 
H_r \cdots H_{p+1} H_p & \text{if } r \geq p, \\
I & \text{if } r < p,
\end{cases}
\]

where \(I\) denotes the identity matrix. Following (2) the value of the legitimate agents at any point in time \(t\) has two salient influence terms, the values contributed by other legitimate agents \(\tilde{x}_L(T_0, t)\) and the values contributed by other malicious agents \(\phi_M(T_0, t)\) so that for all \(t \geq T_0\)

\[
x_L(T_0, t) = \tilde{x}_L(T_0, t) + \phi_M(T_0, t), \tag{3}
\]

where

\[
\tilde{x}_L(T_0, t) = \left( \prod_{k=T_0-1}^{T_0-2} W_L(k) \right) x_L(0), \tag{4}
\]

and

\[
\phi_M(T_0, t) = \sum_{k=T_0-1}^{T_0-1} \left( \prod_{l=k+1}^{T_0-1} W_L(l) \right) W_M(k) x_M(k). \tag{5}
\]

Relation (3) shows explicitly how the states of legitimate agents depend on their initial states at time \(T_0\) and on the states of malicious agents. The term \(\phi_M(T_0, t)\) given in (5) captures the total influence of the malicious agents on the states of the legitimate agents from the initial time \(T_0\) to the current time \(t\) and will be the primary subject of study. Specifically, we will study attainable bounds on the malicious agents’ influence term \(\phi_M(T_0, t)\).

**Remark** (Arbitrary starting time \(T_0\) of the consensus,). We emphasize that the consensus updates from Eq. (2) begin at an arbitrary time \(T_0 \geq 0\) and that before this time agent values \(x_i(t)\) are not updated for \(i \in \mathcal{L}\). In other words, for \(t \in [0 \colon T_0\) only observations \(\alpha_{ij}(t)\) are collected by each agent \(i\) for its neighbors \(j \in \mathcal{N}_i\). The choice of \(T_0\) is a user-selected parameter that affects the overall influence of malicious agents in the network and the convergence rate.
There exists, almost surely, a random variable \( y(T_0) \) such that
\[
\lim_{t \to \infty} \hat{x}_C(T_0, t) = y(T_0) \mathbf{1},
\]
where \( \mathbf{1} \in \mathbb{R}^{|L|} \times 1 \), and \( y(T_0) \) is in the convex hull of the legitimate agent values \( x_i(0), i \in L \), and its distribution depends on the starting time \( T_0 \) of the consensus algorithm. Further, almost surely, a limit exists for the legitimate agent values which is a random variable \( z(T_0) \) such that
\[
\lim_{t \to \infty} x_L(T_0, t) = z(T_0) \mathbf{1}.
\]

Here, \( z(T_0) \) is in the convex hull of the legitimate and malicious agent values \( x_i(T_0), i = 1, \ldots, n \), and its deviation from the nominal consensus value (the case with no malicious agents) depends on the starting time \( T_0 \) of the consensus algorithm. See Def. [II.2] below.

2) Deviation: There is a probabilistic upper bound on the asymptotic distance between the legitimate agent values and a suitably defined weighted average of their initial values, i.e.,
\[
P \left( \max_{i \in L} \limsup_{t \to \infty} |x_L(T_0, t) - 1 \nu' x_C(0)|_i \leq \Delta(T_0, \delta) \right) 
\geq 1 - \delta,
\]
for some finite \( \Delta(T_0, \delta) \leq \Delta_{\text{max}}(T_0, \delta) \), where \( \delta > 0 \) is a user-defined error probability, while \( \nu \) is the Perron-Frobenius left-eigenvector of the matrix \( W_C \) defined in [II.2] such that \( \nu' \mathbf{1} = 1 \).

We next discuss the assumptions that we use in the sequel. For this, we let \( G_L = \langle L, E_L \rangle \) be the subgraph of the graph \( G \) that is induced by the legitimate agents, where \( E_L \subseteq E \) is given by \( E_L = \{ \{i, j\} \mid \{i, j\} \in E, i \in L, j \in L\} \). We use the following assumptions throughout the paper:

**Assumption 1.** Suppose that the following hold:
1. [Sufficiently connected graph] The subgraph \( G_L \) induced by the legitimate agents is connected.
2. [Homogeneity of trust variables] The expectation of the variables \( \alpha_{ij} \) are constant for the case of malicious transmissions and legitimate transmissions, respectively, i.e., for some scalars \( c, d \) with \( c \neq d \),
\[
c = E(\alpha_{ij}(t)) - 1/2 \quad \text{for all } i \in L, j \in N_i \cap M,
\]
\[
d = E(\alpha_{ij}(t)) - 1/2 \quad \text{for all } i \in L, j \in N_i \cap L.
\]
3. [Independence of trust observations] The observations \( \alpha_{ij}(t) \) are independent for all \( t \) and all pairs of agents \( i \) and \( j \), with \( i \in L, j \in N_i \). Moreover, for any \( i \in L \) and \( j \in N_i \), the observation sequence \( \{\alpha_{ij}(t)\} \) is identically distributed.

Assumption [II.1] captures the general connectivity requirement in consensus networks with a fixed topology. The assumption on the homogeneity of trust variables (cf. Assumption [II.2]) is made only for the ease of exposition. The assumption can be generalized to the heterogeneous case as discussed in Section [V].
∆ (the case of average consensus, we aim to find a bound on (Bounded deviation for average consensus) convergence.

α consensus to a common limit value is attainable even when (Convergence of the consensus protocol) we aim to address under the assumptions in Assumption 1:

Figure 2. A schematic showing the deviation from the nominal average in the case of a malicious agent attack and no resilience in the consensus protocol.

B. Problem Definition

In the following we summarize the three main problems that we aim to address under the assumptions in Assumption 1.

Problem 1 (Convergence of the consensus protocol). We aim to show that in the presence of an adversarial attack, consensus to a common limit value is attainable even when malicious agents comprise > 1/2 of the network connectivity if observations αij(t) are available. We seek conditions on αij(t), and a definition of the weights W(t) that allow convergence.

Problem 2 (Bounded deviation for average consensus). For the case of average consensus, we aim to find a bound on deviation from the average consensus value, ∆(T0, δ) that can be achieved with a probability at least 1 − δ, where the bound is given as a function of problem parameters (such as the number of legitimate and the number of malicious agents) and the characteristics of αij (such as bounds on its expected value E(αij)).

Problem 3 (Characterization of convergence rate τ). We aim to determine a finite convergence time for the correct classification of legitimate and malicious agents almost surely. Additionally, we aim to identify the rate of convergence as a function of the stochastic observations αij and other problem parameters that we characterize.

III. ANALYSIS

A. Convergence of Consensus of Legitimate Agents Inputs

In this section, we analyze the state dynamics among the legitimate agents, that is governed by the weight matrix WLC(k), as seen from (2). Recalling that W(t) = [WLC(t) WM(t)], from the definition of the matrix W(t) (cf. (8)), we see that the matrix WLC(t) is a sub-stochastic matrix, since the sum of some rows may be strictly less than one. This section analyzes the convergence of the (backward) product \( \prod_{k=0}^{\infty} W_{LC}(k) \), as \( t \to \infty \). We prove the strong ergodicity of this product, and show that the limit matrix \( \prod_{k=0}^{\infty} W_{LC}(k) \) has strictly positive entries, i.e., \( \prod_{k=0}^{\infty} W_{LC}(k) \neq 0 \).

Our goal here is to show convergence of the consensus system in Eq. (2) using the choice of weights from Eq. (8) that exploit the observations αij(t). Specifically, we aim to characterize the conditions on αij(t) necessary to achieve consensus in the presence of malicious agents even when the number of malicious agents is arbitrarily high.

We will achieve this in three parts: first we show that for a sufficiently connected network (cf. Assumption 1.1) the weight matrix in Eq. (8) reaches its limit described in Eq. (12) in finite time \( T_f \) almost surely and that this limit is a primitive matrix (cf. Lemma 1 and Proposition 1). Second we show that this in turn implies that the legitimate agents’ influence captured in Eq. (4) approaches a finite limit almost surely (cf. Proposition 2). By the same token, the malicious agents’ influence captured in Eq. (5) also approaches a finite limit almost surely (c.f. Proposition 3). Finally, putting these limits together we show that the values of the legitimate agents, \( x_L(T_0, t) \), approach a finite limit, i.e. we achieve convergence.

In what follows we use the notion of a primitive matrix, which is provided next.

Definition III.1 (Primitive matrix). A nonnegative square matrix A is said to be a primitive matrix if there exists an integer \( k \geq 1 \) such that \( A^k > 0 \).

Let us define a matrix \( W_{LC} \) with the entries given by: for every \( i, j \in \mathcal{L} \),

\[
[W_{LC}]_{ij} = \begin{cases} \frac{1}{\max\{\left|\mathcal{N}_i \cap \mathcal{L}\right| + 1, \kappa\}} & \text{if } j \in \mathcal{N}_i \cap \mathcal{L}, \\ 1 - \frac{1}{\max\{\left|\mathcal{N}_i \cap \mathcal{L}\right| + 1, \kappa\}} & \text{if } j = i, \\ 0 & \text{otherwise}. \end{cases}
\]

For the matrix \( W_{LC} \), we have the following result.

Lemma 1. The matrix \( W_{LC} \) is primitive.

Proof: The result follows directly by the definition of the matrix \( W_{LC} \) in (12) and the assumption that the graph \( G_L \) is connected (cf. Assumption 1.1).

Under Assumptions 1.2–1.3, we next provide a result for the random quantities \( \beta_{ij}(t) \) defining the weight matrix \( W(t) \).

Lemma 2. Consider the random variables \( \beta_{ij}(t) \) as defined in (6). Then, for every \( t \geq 0 \) and every \( i, j \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{L}, \)

\[
\Pr(\beta_{ij}(t) < \theta) \leq \max\{\exp(-2(t+1)d^2), 1_{(d<0)}\}.
\]

Additionally, for every \( t \geq 0 \) and every \( i, j \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M}, \)

\[
\Pr(\beta_{ij}(t) \geq \theta) \leq \max\{\exp(-2(t+1)c^2), 1_{(c>0)}\}.
\]

Proof: By the linearity of the expectation, we have \( E(\beta_{ij}(t)) = \sum_{k=0}^{\infty} (E(\alpha_{ij}(k)) - 1/2) \). For every \( k \geq 0 \), and for all \( i, j \) with \( i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{L} \), by Assumptions 1.2–1.3, we have \( d = E(\alpha_{ij}(k)) - 1/2 \). Additionally, for every \( k \geq 0 \) and \( i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M} \), by Assumptions 1.2–1.3, we have \( c = E(\alpha_{ij}(k)) - 1/2 \). The result then follows directly from the Chernoff-Hoeffding Inequality (see [58] Theorem 1.1).

From Lemma 2 one can observe the following.

Corollary 1 (Bounds on Expectation of \( \alpha_{ij}(t) \)). For the choice \( \beta_{ij}(t) = \sum_{k=0}^{\infty} \alpha_{ij}(k) - 1/2 \), we must have that \( c = E(\alpha_{ij}(k)) - 1/2 < 0 \) for all \( i \in \mathcal{L}, j \in \mathcal{M}, k \geq 0 \) and \( d = E(\alpha_{ij}(k)) - 1/2 > 0 \), for all \( i, j \in \mathcal{L}, k \geq 0 \) to have a decaying misclassification probabilities.
We provide some intuition about this corollary as it has significant implications for the characteristics of the \( \alpha_{ij}(t) \) necessary for obtaining several important results. The condition that \( c \) and \( d \) be bounded away from zero in Corollary [1] intuitively means that there is information captured by the \( \alpha_{ij}(t) \) observations. The case of \( c = 0, d = 0 \) is equivalent to saying that, in expectation, the observations \( \alpha_{ij}(t) \) will be the same regardless of the transmission being legitimate or malicious (the observations \( \alpha_{ij}(t) \) contain no useful information).

Hereafter, we assume that \( c < 0 \) and \( d > 0 \), which means that \( \alpha_{ij}(t) \) tends to be less than \( 1/2 \) for malicious transmissions (has expectation value less than \( 1/2 \)) and \( \alpha_{ij}(t) \) tends to be greater than \( 1/2 \) for legitimate transmissions (has expectation value greater than \( 1/2 \)).

**Proposition 1.** There exists a finite time instant \( T_f > 0 \) such that \( W_{L}(k) = \overline{W}_{L} \) for all \( k \geq T_f \) almost surely. Moreover, there exists a stochastic vector \( v > 0 \) such that

\[
\lim_{k \to \infty} W^k_{L} = 1 v'.
\]

In particular, \( \prod_{k=T_0-1}^{\infty} W_{L}(k) > 0 \) for every \( T_0 \geq 0 \) almost surely.

**Proof:** By Lemma [2] we have that for every \( k \geq 0 \) and \( i \in L, j \in N_i \cap L \),

\[
P(\beta_{ij}(k) < 0) \leq \exp(-2(k+1)d^2).
\]

Additionally, for every \( k \geq 0 \) and \( i \in L, j \in N_i \cap M \),

\[
P(\beta_{ij}(k) \geq 0) \leq \exp(-2(k+1)c^2).
\]

Note that we have \( \sum_{k=0}^{\infty} \exp(-2(k+1)c^2) < \infty \) and \( \sum_{k=0}^{\infty} \exp(-2(k+1)d^2) < \infty \). Thus, by the Borel-Cantelli Lemma, the events \( \{ \beta_{ij}(k) < 0 \ \forall i \in L, j \in N_i \cap M \} \) and \( \{ \beta_{ij}(k) \geq 0 \ \forall i \in L, j \in N_i \cap L \} \) occur only finitely often almost surely. Therefore, there exists a (random) finite time \( T_f > 0 \) such that \( W_{L}(k) = \overline{W}_{L} \) for all \( k \geq T_f \) almost surely.

Now, almost surely, we have

\[
\prod_{k=T_0-1}^{\infty} W_{L}(k) = \prod_{k=\max(T_f,T_0)}^{\infty} W_{L}(k) \prod_{k=T_0-1}^{\max(T_f,T_0)-1} W_{L}(k) = \lim_{k \to \infty} W^k_{L} W_{L}(k) \prod_{k=T_0-1}^{\max(T_f,T_0)-1} W_{L}(k).
\]

By Lemma [1] the matrix \( \overline{W}_{L} \) is primitive. Therefore, by the Perron-Frobenius Theorem (see [59]), there exists a stochastic column vector \( v > 0 \) such that \( \lim_{k \to \infty} W^k_{L} = 1 v' \).

Consequently, almost surely we have

\[
\prod_{k=T_0-1}^{\infty} W_{L}(k) = 1 v' \prod_{k=T_0-1}^{\max(T_f,T_0)-1} W_{L}(k).
\]

Since \( v > 0 \) and the diagonal entries of the matrices \( W_{L}(k) \) are positive (by construction), it follows that \( v' \prod_{k=\max(T_f,T_0)-1}^{\infty} W_{L}(k) > 0 \), implying that almost surely \( \prod_{k=T_0-1}^{\infty} W_{L}(k) > 0 \).

Our next results are aimed at proving that the legitimate agents’ values reach consensus almost surely. In particular, using the decomposition \( x_{L}(T_0,t) = \tilde{x}_{L}(T_0,t) + \phi_{M}(T_0,t) \) of the legitimate agent states (see (3)), we focus on the limiting behavior of \( \tilde{x}_{L}(T_0,t) \), as defined in (4). We show that \( \tilde{x}_{L}(T_0,t) \) converges almost surely to a consensus value, as \( t \to \infty \), as seen in the following proposition.

**Proposition 2 (Convergence of legitimate agents’ values).** Consider the dynamics in (2). Let \( x_{L}(0) \) be the vector of initial values of legitimate agents, and define

\[
\tilde{x}_{L}(T_0) = \left( \prod_{k=T_0-1}^{\infty} W_{L}(k) \right) x_{L}(0).
\]

Then, we have \( [\tilde{x}_{L}(T_0)]_i = [\tilde{x}_{L}(T_0)]_j \) for every \( i, j \in L \), almost surely.

**Proof:** By Proposition [1] there exists a finite time instant \( T_f \) such that \( W_{L}(k) = \overline{W}_{L} \) for all \( k \geq T_f \) almost surely, and \( \lim_{k \to \infty} W^k_{L} = 1 v' \) for a stochastic vector \( v > 0 \). Hence, it follows that almost surely

\[
\tilde{x}_{L}(T_0) = \left( \prod_{k=T_0-1}^{\infty} W_{L}(k) \right) x_{L}(0) = \left( \prod_{k=\max(T_f,T_0)}^{\infty} W_{L}(k) \prod_{k=T_0-1}^{\max(T_f,T_0)-1} W_{L}(k) \right) x_{L}(0) = \left( \prod_{k=T_0-1}^{\max(T_f,T_0)-1} W_{L}(k) \right) x_{L}(0).
\]
By letting \( \tilde{z}_{T_j}^L(T_0) = \left( \prod_{k=T_0-1}^{\max\{T_j,T_0\}-1} W_L(k) \right) x_L(0) \), we obtain that \( \tilde{z}_{T_j}^L(T_0) = 1v'\tilde{z}_{T_j}^L(T_0) \), almost surely.

We next consider the limiting effect that malicious agents’ states have on the legitimate agents’ states. Specifically, in the state decomposition relation \( x_L(T_0,t) = \tilde{x}_L(T_0,t) + \phi_M(T_0,t) \) for the legitimate agents’ states (see (3)), we focus on the limiting behavior of \( \phi_M(T_0,t) \) defined in (5), as \( t \to \infty \). We show that \( \phi_M(T_0,t) \) converges to a consensus almost surely given our choice of weight matrices, as seen in the following proposition.

**Proposition 3.** Consider the vector \( \phi_M^\infty(T_0) \) defined by

\[
\phi_M^\infty(T_0) = \sum_{k=T_0-1}^{\infty} \left( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \right) W_M(k)x_M(k).
\]

Then, the entries of this vector are the same almost surely, i.e., \( [\phi_M^\infty(T_0)]_{i,j} = [\phi_M^\infty(T_0)]_{j,i} \), almost surely for every \( i,j \in L \).

**Proof:** By Proposition 1, there exists a finite time instant \( T_J > 0 \) such that almost surely,

\[
W_L(k) = W_L_0 \quad \text{for all} \quad k \geq T_J.
\]

This implies that \( W_M(k) = 0 \) for all \( k \geq T_J \) almost surely. Therefore,

\[
\phi_M^\infty(T_0) = \sum_{k=T_0-1}^{\infty} \left( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \right) W_M(k)x_M(k)
\]

\[= \sum_{k=T_0-1}^{T_J-1} \left( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \right) W_M(k)x_M(k).\]

Now, for the product \( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \), we have that almost surely

\[
\prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) = \prod_{l=\max\{T_0,T_0\}-1}^{\infty} W_L(l)
\]

\[= \lim_{t \to \infty} \prod_{l=\max\{T_0,T_0\}-1}^{t} W_L(l),\]

where the second relation follows from (13).

By Proposition 1, we also have for a stochastic vector \( v \), such that \( \lim_{k \to \infty} W_L = 1v' \). Hence, almost surely,

\[
\phi_M^\infty(T_0) = \sum_{k=T_0-1}^{T_J-1} \left( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \right) W_M(k)x_M(k)
\]

\[= 1v' \left[ \sum_{k=T_0-1}^{T_J-1} \left( \prod_{l=k+1}^{\max\{T_0,T_0\}-1} W_L(l) \right) W_M(k)x_M(k) \right],\]

for some finite \( T_J \), implying that all the coordinates of the vector \( \phi_M^\infty(T_0) \) are identical almost surely.

As a direct consequence of Propositions 2 and 3, we have the following result.

**Corollary 2.** There exists, almost surely, a random variable \( y(T_0) \) such that

\[
\lim_{t \to \infty} \tilde{x}_L(T_0,t) = y(T_0)1,
\]

where \( y(T_0) \) is in the convex hull of the legitimate agents’ initial values \( x_i(T_0) \), \( i \in L \), and its distribution depends on the starting time \( T_0 \) of the consensus algorithm in (2). Furthermore, almost surely, there exists a random variable \( z(T_0) \) such that

\[
\lim_{t \to \infty} x_L(T_0,t) = z(T_0)1,
\]

where \( z(T_0) \) is in the convex hull of the initial values \( x_i(T_0) \), \( i = 1, \ldots, n \), of the legitimate and malicious agents, and its distribution depends on the starting time \( T_0 \) of the consensus algorithm.

**Proof:** The results follow directly from Proposition 2 and Proposition 3 and the consensus algorithm in (2).

We note that, in the view of the fact that the legitimate agents start consensus protocol at time \( T_0 \), we have that the random variable \( y(T_0) \) is in the convex hull of the initial values \( x_i(0) \), \( i \in L \), of legitimate agents. The random variable \( z(T_0) \) is in the convex hull of \( x_i(0), i \in L \), and the values \( x_m(T_0), m \in M \), since the malicious agents may have deviated from their initial values at time \( t = 0 \).

We are now ready to state our main convergence result.

**Theorem 1** (Convergence of Resilient Consensus). Consider the consensus system described by the dynamics in Eq. (2) with an arbitrary start time of \( T_0 \geq 0 \) and weights as described by Eq. (3). Given that 1) \( G \) is a sufficiently connected network and that 2) bounds on the expected values of the observations \( \alpha_{ij}(k) \) (see Defn. II.2) satisfy \( c = E(\alpha_{ij}(k)) - 1/2 < 0 \) for all \( i \in L, j \in N_i \cap M, k \geq 0 \) and \( d = E(\alpha_{ij}(k)) - 1/2 > 0 \) for all \( i \in L, j \in N_i \cap L, k \geq 0 \), the consensus algorithm converges almost surely, i.e.,

\[
\lim_{t \to \infty} x_L(T_0,t) = z(T_0)1,
\]

almost surely, independently of the number of malicious agents in the network. The value \( z(T_0) \) is in the convex hull of the initial values \( x_i(T_0), i = 1, \ldots, n \), of both the legitimate and malicious agents.

**Proof:** The proof follows from Propositions 2 and 3. Specifically, we have that the legitimate agent values satisfy \( x_L(T_0,t) = \tilde{x}_L(T_0,t) + \phi_M(T_0,t) \), where the first term \( \tilde{x}_L(T_0,t) \) converges by Proposition 2 and the second term \( \phi_M(T_0,t) \) converges by Proposition 3.

Note that this result asserts convergence in the presence of an arbitrary number of malicious nodes and characterizes necessary conditions on the observations \( \alpha_{ij}(t) \) to attain this convergence. However, the distance of the limit value \( z(T_0) \) to the nominal average consensus value in the ideal case (when the malicious agents are known at the start time) has not been discussed. The characterization of this distance is the subject of the following section.

**B. Deviation from Nominal Consensus Value**

In this section we characterize the deviation from the nominal consensus value under our consensus model in the presence of a malicious attack. The nominal consensus value is the consensus value over the graph \( G_L \) of legitimate agents in the case of no malicious agents in the network. More
specifically, by Proposition 1 we know that there exists a finite time instant \( T_f > 0 \) such that \( W_L(k) = \overline{W}_L \) for all \( k \geq T_f \) almost surely, where \( \overline{W}_L \) is given in (12). The nominal consensus value corresponds to consensus that would have been reached among the legitimate agents if the dynamics in (2) were governed by \( [\overline{W}_L \ 0] \) instead of \([W_L(k) \ W_M(k)]\). Hence, in view of Proposition 1 the nominal consensus value over the graph \( G_L \) of legitimate agents is \( 1^{v'}x_L(0) \).

In what follows we investigate the deviation of the consensus value \( \lim_{T \rightarrow \infty} x_L(T_0, t) \) reached by the legitimate agents (as predicted by Theorem 1) from the nominal value \( 1^{v'}x_L(0) \). The user-defined start time \( T_0 \), for running the consensus dynamics in Eq. (2), plays a major role in determining the amount of the resulting deviation due to the malicious agents’ inputs to the system. Intuitively, the longer \( \alpha_{ij}(t) \) values are observed for all pairs \( \{i, j\} \in \mathcal{E} \) before starting the consensus, the probability of making a misclassification error (classifying a legitimate agent as malicious and vice versa) decays exponentially.

Our approach in this section is to show that for larger values of \( T_0 \), the probability of our weight matrix not having reached its limit value decays exponentially (see Lemma 3). The implication of this manifests itself as a smaller overall deviation. Specifically, deviation of the convergence limit (cf. Theorem 1) from the nominal average consensus value can be bounded by the deviation introduced by legitimate agents being misclassified as malicious and by malicious agents being misclassified as legitimate. Propositions 4 and 5 proven in this section provide definitive bounds for these deviations with probability \( p(T_0) \) which is characterized as a function of \( T_0 \).

We begin by showing that the probability of the weights from Eq. (9) not taking their ideal form \( \overline{W}_L \) decays exponentially with \( T_0 \).

**Lemma 3.** For every \( T_0 \geq 1 \),

\[
\mathbb{P}(\exists k \geq T_0 - 1 : W_L(k) \neq \overline{W}_L) \leq |\mathcal{L}|^2 \frac{|L|}{1 - \exp(-2T_0d^2)} \cdot \frac{|M|}{1 - \exp(-2\delta c^2)}.
\]

**Proof:** For every \( T_0 \geq 1 \), we have

\[
\mathbb{P}(\exists k \geq T_0 - 1 : W_L(k) \neq \overline{W}_L) = \Pr \left( \bigcup_{k \geq T_0 - 1} \left\{ W_L(k) \neq \overline{W}_L \right\} \right) \leq \sum_{k = T_0 - 1}^{\infty} \mathbb{P}(W_L(k) \neq \overline{W}_L),
\]

where the inequality follows from applying the union bound to the event \( \bigcup_{k \geq T_0 - 1} \left\{ W_L(k) \neq \overline{W}_L \right\} \). Further, for every \( k \) we have that

\[
\{W_L(k) \neq \overline{W}_L\} = \bigcup_{i \in \mathcal{L}, j \in N_i \cap \mathcal{L}} \{\beta_{ij}(k) < 0\} \bigcup_{i \in \mathcal{L}, j \in N_i \cap \mathcal{M}} \{\beta_{ij}(k) \geq 0\}.
\]

Therefore, by the union bound

\[
\mathbb{P}(W_L(k) \neq \overline{W}_L) \leq \sum_{i \in \mathcal{L}, j \in N_i \cap \mathcal{L}} \mathbb{P}(\beta_{ij}(k) < 0) + \sum_{i \in \mathcal{L}, j \in N_i \cap \mathcal{M}} \mathbb{P}(\beta_{ij}(k) \geq 0) \leq |\mathcal{L}|^2 \cdot \exp(-2(k + 1)d^2) + |\mathcal{L}| |\mathcal{M}| \cdot \exp(-2(k + 1)c^2),
\]

where the last inequality follows from Lemma 3. The result then follows by combining relations (22) and (23). ■

Next, we consider the deviation contributed by misclassifying legitimate agents as malicious at some times and, thus, discarding their values. In the sequel, we use the bound of Lemma 3 to define a quantity \( g_{L}(T_0, \delta) \) for an error tolerance \( \delta \). Specifically, given \( \delta > 0 \), we define

\[
g_{L}(T_0, \delta) = \frac{2|\mathcal{L}|}{\delta} \cdot \frac{\exp(-2T_0d^2)}{1 - \exp(-2\delta c^2)} + \frac{2|\mathcal{L}| |\mathcal{M}|}{\delta} \cdot \frac{\exp(-2T_0c^2)}{1 - \exp(-2\delta c^2)},
\]

where \( \eta \geq \sup_{i \in \mathcal{L} \cup \mathcal{M}, t \in \mathbb{N}} |x_i(t)| \) is a finite upper bound on the absolute input value.

In what follows, we make use of the following lemma.

**Lemma 4.** Let \( \tilde{X}, X \in \mathbb{R}_+^{m \times m} \) be two row-substochastic matrices, and let \( \gamma > 0 \) be such that \( \tilde{X}_{ii} \geq \gamma > 0 \) and \( X_{ii} \geq \gamma \) for all \( i \). Then \[\left| \tilde{X} - X \right|_1 \leq 2(1 - \gamma), \]

for all \( i \), where \( |X| \) denote the matrix with entries \( |x_{ij}| \).

**Proof:** Let coordinate index \( i \) be arbitrary, \( 1 \leq i \leq m \). First, we prove that \( |\tilde{x}_{ii} - x_{ii}| + 2\gamma \leq |\tilde{x}_{ii} + x_{ii}|. \) By the triangle inequality, we obtain

\[
|\tilde{x}_{ii} - x_{ii}| + 2\gamma = |(\tilde{x}_{ii} - \gamma) - (x_{ii} - \gamma)| + 2\gamma = |\tilde{x}_{ii} - \gamma| + |x_{ii} - \gamma| + 2\gamma.
\]

Since \( \tilde{x}_{ii} - \gamma \geq 0 \) and \( x_{ii} - \gamma \geq 0 \) we have that

\[
|\tilde{x}_{ii} - \gamma| + |x_{ii} - \gamma| + 2\gamma = \tilde{x}_{ii} + x_{ii}.
\]

Therefore, it follows that

\[
\left| \tilde{X} - X \right|_1 = \tilde{x}_{ii} + x_{ii} + \sum_{j=1, j \neq i}^{m} |\tilde{x}_{ij} - x_{ij}| \leq \tilde{x}_{ii} + x_{ii} + \sum_{j=1, j \neq i}^{m} (\tilde{x}_{ij} + x_{ij}),
\]

where the last inequality is obtained by using the triangle inequality and the fact that the matrices \( X \) and \( \tilde{X} \) have nonnegative entries. Since both \( X \) and \( \tilde{X} \) are stochastic matrices, we further have that \( \sum_{j=1}^{m} \tilde{x}_{ij} + x_{ij} \leq 2 \). From the preceding two relations, it follows that \[\left| \tilde{X} - X \right|_1 \leq 2(1 - \gamma). \]

Denote by \( \varphi_i(T_0, t) \) the deviation suffered by legitimate agent \( i \) that is caused by its incorrect classification of the legitimate agents, i.e., for all \( i \in \mathcal{L} \),

\[
\varphi_i(T_0, t) = \left[ \tilde{x}_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} \overline{W}_L \right) x_L(0) \right]_i.
\]

**Proposition 4.** Given an error tolerance \( \delta > 0 \), for the
deviations $\varphi_i(T_0, t), i \in \mathcal{L}$, as defined in (25), we have
\[
P \left( \max_{i \in \mathcal{L}} \limsup_{t \to \infty} \varphi_i(T_0, t) > \tilde{y}_\mathcal{L}(T_0, \delta) \right) < \delta,
\]
where $\tilde{y}_\mathcal{L}(T_0, \delta)$ is given by (24).

Proof: Let $T_f(T_0, t)$ be a random variable equal to 0 if $W_\mathcal{L}(k) = \overline{W}_\mathcal{L}$ for all $k \in [T_0 - 1, t - 1]$, and equal to $\sup\{1 + 1 : W_\mathcal{L}(t + T_0 - 1) = \overline{W}_\mathcal{L}, t \in [0, t - T_0]\}$, otherwise. In view of the evolution of $\tilde{x}_\mathcal{L}(T_0, t)$ as given in (4), we have
\[
\varphi_i(T_0, t) = \left| \left( \left( \prod_{k=T_0+1}^{T_0+t-1} \overline{W}_L \right) - \overline{W}_L^{t-T_0} \right) x_\mathcal{L}(0) \right|_i.
\]
Define
\[
\Delta(W_\mathcal{L}, T_f) = \left( \prod_{k=T_0+1}^{T_0+t-1} \overline{W}_L \right) - \left( \prod_{k=T_0+1}^{T_0+t-1} \overline{W}_L \right).
\]
Then,
\[
\varphi_i(T_0, t) = \left| \left( \left( \prod_{k=T_0+1}^{T_0+t-1} \overline{W}_L \right) - \overline{W}_L^{t-T_0} \right) \Delta(W_\mathcal{L}, T_f)x_\mathcal{L}(0) \right|_i \leq (a) \max_{i \in \mathcal{L}} \left| \Delta(W_\mathcal{L}, T_f)x_\mathcal{L}(0) \right|_i,
\]
where (a) follows since $W_\mathcal{L}$ is a row-stochastic matrix. Further, since $|x_\mathcal{L}(0)| \leq \eta$, it follows that
\[
\max_{i \in \mathcal{L}} \left| \Delta(W_\mathcal{L}, T_f)x_\mathcal{L}(0) \right|_i \leq \eta \max_{i \in \mathcal{L}} \left| \Delta(W_\mathcal{L}, T_f)x_\mathcal{L}(0) \right|_i,
\]
where for a matrix $A = [a_{ij}]$, we use $|A|$ to denote the matrix with entries $|a_{ij}|$. Therefore,
\[
\varphi_i(T_0, t) \leq \eta \max_{i \in \mathcal{L}} \left| \Delta(W_\mathcal{L}, T_f)x_\mathcal{L}(0) \right|_i.
\]
Let $n_w = \max\{|\mathcal{L}| + |\mathcal{M}|, \kappa\}$ and note that $W_\mathcal{L}$ is a stochastic matrix with $[W_\mathcal{L}]_{ii} \geq \frac{1}{n_w}$ (see (12)). Additionally, for every $k$, the matrix $W_\mathcal{L}$ is substochastic with $[W_\mathcal{L}(k)]_{ii} \geq \frac{1}{n_w}$ for every $i \in \mathcal{L}$. Therefore, by Lemma 4 we have
\[
\left| \Delta(W_\mathcal{L}, T_f) \right|_i \leq 2 \left[ 1 - \left( \frac{1}{n_w} \right)^{T_f(T_0, t)} \right].
\]
Define
\[
\varphi(T_0, t) = 2 \eta.
\]
Then, we have $\varphi_i(T_0, t) \leq \varphi(T_0, t)$ for all $i \in \mathcal{L}$ and $t \geq T_0$, thus implying that
\[
\max_{i \in \mathcal{L}} \limsup_{t \to \infty} \varphi_i(T_0, t) \leq \lim_{t \to \infty} \varphi(T_0, t).
\]
By Proposition 1 for $t \geq T_f$, we have $T_f(T_0, t) = t$ almost surely, implying that $\varphi(T_0, t)$ almost surely converges as $t \to \infty$. By Markov’s inequality it follows that
\[
P \left( \lim_{t \to \infty} \varphi(T_0, t) > \tilde{y}_\mathcal{L}(T_0, \delta) \right) \leq \frac{\varphi(T_0, t)}{\tilde{y}_\mathcal{L}(T_0, \delta)}.
\]
By definition, for every $t \geq T_0$ we have that $0 \leq T_f(T_0, t) \leq T_f(T_0, t + 1)$, therefore, $0 \leq \varphi(T_0, t) \leq \varphi(T_0, t + 1) \leq 2$ for every $t \geq T_0$. Hence, by the Monotone Convergence Theorem (see [60]), it follows that
\[
E \left( \lim_{t \to \infty} \varphi(T_0, t) \right) = \lim_{t \to \infty} E \left( \varphi(T_0, t) \right) = 2\eta \left[ 1 - \lim_{t \to \infty} \frac{1}{T_f(T_0, t)} \right] \leq 2\eta \left[ 1 - \lim_{t \to \infty} \mathbb{P} \left( T_f(T_0, t) = 0 \right) \right] = 2\eta \lim_{t \to \infty} \mathbb{P} \left( T_f(T_0, t) = 0 \right).
\]
Since $1 - \mathbb{P}(T_f(T_0, t) = 0) = \mathbb{P}(T_f(T_0, t) > 0)$, it further follows that
\[
E \left( \lim_{t \to \infty} \varphi(T_0, t) \right) = 2\eta \lim_{t \to \infty} \mathbb{P}(T_f(T_0, t) > 0) \leq 2\eta \mathbb{P}(\exists k \geq T_0 : W_\mathcal{L}(k) \neq \overline{W}_\mathcal{L}) \leq \delta \cdot \tilde{g}_\mathcal{L}(T_0, \delta),
\]
where the last inequality follows from Lemma 3 and the definition of $\tilde{g}_\mathcal{L}(T_0, \delta)$ in (24).

Now we consider the deviation contributed by malicious agents that are misclassified as legitimate. We denote by $\phi_i(T_0, t), i \in \mathcal{L}$, the worst case effect on legitimate agent $i$ due to its incorrect malicious agents’ classification (labeling an untrustworthy agent as trustworthy) at some time $t \geq T_0$. Specifically, let $\eta = \max_{i \in \mathcal{L} \cup \mathcal{M}} |x_\mathcal{L}(0)_i|$ and define
\[
\phi_i(T_0, t) = \eta \sum_{k=T_0+1}^{T_0+t-1} \sum_{j \in \mathcal{L} \cup \mathcal{M}} \left[ \prod_{l=k+1}^{T_0+t-1} W_\mathcal{L}(l) \right] W_\mathcal{M}(k)_{ij},
\]
for all $i \in \mathcal{L}$. We note that these quantities are nonnegative, since the matrices $W_\mathcal{L}(k)$ and $W_\mathcal{M}(k)$ are nonnegative for all $k$. Looking at the malicious agents’ influence vector $\phi_\mathcal{M}(T_0, t)$ defined in (5), we see that
\[
|\phi_\mathcal{M}(T_0, t)_i| \leq \max_{j \in \mathcal{L}} \phi_j(T_0, t), \quad \forall i \in \mathcal{L}.
\]
Our next result provides a probabilistic bound on the maximal influence $\max_{j \in \mathcal{L}} \phi_j(T_0, t)$ of malicious agents inputs on the legitimate agents’ values. To do so, given $\delta > 0$, we define
\[
\tilde{g}_\mathcal{M}(T_0, \delta) = \frac{\eta |\mathcal{L}| |\mathcal{M}|}{\delta \kappa} \exp(-2T_0c^2) \exp(-2T_0c^2),
\]
where $\eta \geq \sup_{i \in \mathcal{L} \cup \mathcal{M}, t \in \mathcal{L}} |x_\mathcal{L}(t)|$.

Proposition 5. Given an error tolerance $\delta > 0$, for the deviations $\phi_i(T_0, t), i \in \mathcal{L}$, as defined in (27), we have
\[
P \left( \max_{i \in \mathcal{L}} \limsup_{t \to \infty} \phi_i(T_0, t) > \tilde{g}_\mathcal{M}(T_0, \delta) \right) < \delta,
\]
where $\tilde{g}_\mathcal{M}(T_0, \delta)$ is given by (28).

Proof: First, observe that
\[
P \left( \max_{i \in \mathcal{L}} \limsup_{t \to \infty} \phi_i(T_0, t) > \tilde{g}_\mathcal{M}(T_0, \delta) \right) \leq \sum_{i \in \mathcal{L}} \mathbb{P} \left( \limsup_{t \to \infty} \phi_i(T_0, t) > \tilde{g}_\mathcal{M}(T_0, \delta) \right).
\]
It follows by Markov’s inequality and the linearity of the expectation operator that
\[
P \left( \limsup_{t \to \infty} \phi_i(T_0, t) > \tilde{g}_\mathcal{M}(T_0, \delta) \right) \leq E \left( \limsup_{t \to \infty} \phi_i(T_0, t) \right).
\]

In view of the definition of \( \tilde{g}_\mathcal{M}(T_0, \delta) \), to complete the proof, it suffices to show that
\[
E \left( \limsup_{t \to \infty} \phi_i(T_0, t) \right) \leq \frac{\eta|\mathcal{M}|}{\kappa} \exp(-2T_0c^2) \frac{1}{1 - \exp(-2\kappa^2)}.
\]

The rest of the proof deals with establishing relation \( (30) \).

By the definition of \( \phi_i(T_0, t), i \in \mathcal{L} \) in \( (27) \), we have that
\[
\phi_i(T_0, t) = \eta \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} \left[ \left( \prod_{l=k}^{t-1} W_{\mathcal{L}}(l) \right) W_{\mathcal{M}}(k) \right]_{ij}
\]
\[
\leq \eta \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} W_{\mathcal{M}}(k)_{ij},
\]
where the inequality is obtained by using the fact that \( W_{\mathcal{M}}(k) \) is a row-stochastic matrix. Moreover, we have that \( W_{\mathcal{M}}(k)_{ij} \leq 1/\kappa \) (see \( (8) \)), implying that
\[
\phi_i(T_0, t) \leq \eta \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} \frac{1}{\kappa} = \frac{\eta}{\kappa} \phi_i(T_0, t) \leq \frac{\eta}{\kappa} \phi_i(T_0, t).
\]

Hence,
\[
E \left( \limsup_{t \to \infty} \phi_i(T_0, t) \right) \leq E \left( \limsup_{t \to \infty} \tilde{\phi}_i(T_0, t) \right).
\]

By the definition of \( \tilde{\phi}_i(T_0, t) \) in \( (31) \), the sequence \( \{\tilde{\phi}_i(T_0, t)\} \) is a nonnegative non-decreasing sequence with \( t \). Therefore, by the Monotone Convergence Theorem [60],
\[
E \left( \limsup_{t \to \infty} \tilde{\phi}_i(T_0, t) \right) = E \left( \lim_{t \to \infty} \tilde{\phi}_i(T_0, t) \right) \leq E \left( \tilde{\phi}_i(T_0, t) \right).
\]

We further have that
\[
\lim_{t \to \infty} E \left( \tilde{\phi}_i(T_0, t) \right) = \lim_{t \to \infty} E \left( \eta \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} \mathbb{1}_{\{\beta_{ij}(k) \geq 0\}} \right)
\]
\[
= \frac{\eta}{\kappa} \lim_{t \to \infty} \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} E \left( \mathbb{1}_{\{\beta_{ij}(k) \geq 0\}} \right). 
\]

Since \( E \left( \mathbb{1}_{\{\beta_{ij}(k) \geq 0\}} \right) = P(\beta_{ij}(k) \geq 0) \), we obtain that
\[
\lim_{t \to \infty} E \left( \tilde{\phi}_i(T_0, t) \right) = \frac{\eta}{\kappa} \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} P(\beta_{ij}(k) \geq 0).
\]

\[
\leq \frac{\eta}{\kappa} \lim_{t \to \infty} \sum_{k=T_0-1}^{t-1} \sum_{j \in \mathcal{N}_i \cap \mathcal{M}} \exp(-2(k+1)c^2),
\]
where the last inequality follows by Lemma [2]. Hence,
\[
\lim_{t \to \infty} E \left( \tilde{\phi}_i(T_0, t) \right) \leq \frac{\eta|\mathcal{M}|}{\kappa} \lim_{t \to \infty} \sum_{k=T_0-1}^{t-1} \exp(-2(k+1)c^2)
\]
\[
= \frac{\eta|\mathcal{M}|}{\kappa} \cdot \exp(-2T_0c^2) \frac{1}{1 - \exp(-2\kappa^2)}.
\]

By combining estimates in \( (32)-(34) \), the desired relation in \( (30) \) follows.

We are now ready to prove our main deviation result which states that the maximum deviation of the converged consensus value following the dynamics in Eq. \( (2) \) with the weights described in Eq. \( (9) \) can be bounded by a finite value \( \Delta_{\max}(T_0, \delta) \) that we characterize.

**Theorem 2** (Deviation from Nominal Average Consensus). Consider the consensus system described by the dynamics in Eq. \( (2) \) with an arbitrary start time of \( t = T_0 \geq 0 \) and weights as described by Eq. \( (8) \). Given that 1) \( \mathcal{G} \) is a sufficiently connected network and that 2) bounds on the expected values of the observations \( \alpha_{ij}(k) \) (see Defn. II.2) satisfy
\[
c = E(\alpha_{ij}(k)) - 1/2 < 0 \quad \text{for all } i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M}, k \geq 0,
\]
and \( d = E(\alpha_{ij}(k)) - 1/2 > 0 \quad \text{for all } i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{L}, k \geq 0, \)
for a given error probability \( \delta > 0 \), we have the following result
\[
P \left( \max_{i \in \mathcal{L}} \limsup_{t \to \infty} |x_{\mathcal{L}}(T_0, t) - 1v'x_{\mathcal{L}}(0)| \right) \leq \Delta_{\max}(T_0, \delta)
\]
\[
\geq 1 - \delta,
\]
where
\[
\Delta_{\max}(T_0, \delta) = 2 \left[ \tilde{g}_\mathcal{L}(T_0, \delta) + \tilde{g}_\mathcal{M}(T_0, \delta) \right],
\]
with \( \tilde{g}_\mathcal{L}(T_0, \delta) \) and \( \tilde{g}_\mathcal{M}(T_0, \delta) \) respectively given by
\[
\tilde{g}_\mathcal{L}(T_0, \delta) = \frac{\eta|\mathcal{L}|}{\delta} \cdot \exp(-2T_0d^2) + \frac{\eta|\mathcal{L}||\mathcal{M}|}{\delta} \cdot \exp(-2T_0d^2) \\
\tilde{g}_\mathcal{M}(T_0, \delta) = \frac{\eta|\mathcal{L}|}{\delta} \cdot \exp(-2T_0d^2)
\]

and
\[
\eta|\mathcal{L}| = \frac{\eta|\mathcal{L}|}{\delta} \cdot \frac{\exp(-2T_0d^2)}{1 - \exp(-2\kappa^2)}.
\]

**Proof:** First, note that by the triangle inequality
\[
| | x_{\mathcal{L}}(T_0, t) - 1v'x_{\mathcal{L}}(0) | | \leq \left| x_{\mathcal{L}}(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_{\mathcal{L}} \right) x_{\mathcal{L}}(0) \right| \\
+ \left| \left( \prod_{k=T_0}^{t-1} W_{\mathcal{L}} \right) x_{\mathcal{L}}(0) - 1v'x_{\mathcal{L}}(0) \right|
\]
Since \( W_{\mathcal{L}} \) is a primitive stochastic matrix, by the Perron-
Frobenius Theorem it follows that
\[
\lim_{t \to \infty} \left| \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) - 1 v'x_L(0) \right| = 0.
\]
Thus,
\[
\mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |x_L(T_0, t) - 1 v'x_L(0)| \geq \delta \right) \leq \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} \left| x_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| \right.
\]
\[
\geq \delta \mathbb{P} \left( \phi_M(T_0, t) \right) \geq \tilde{g}_M(T_0, \delta/2). \]

Now, since \(x_L(T_0, t) = \tilde{x}_L(T_0, t) + \phi_M(T_0, t)\), by the triangle inequality we have
\[
\left| x_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| \leq \left| \tilde{x}_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| + |\phi_M(T_0, t)|.
\]
It follows by the definition of \(\Delta_{\max}(T_0, \delta)\) in (35) and the union bound that
\[
\mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |x_L(T_0, t) - 1 v'x_L(0)| \geq \Delta_{\max}(T_0, \delta) \right)
\]
\[
\leq \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} \left| \tilde{x}_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| \right.
\]
\[
\geq 2 \tilde{g}_L(T_0, \delta/2) \right) + \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |\phi_M(T_0, t)| \right)
\]
\[
\geq 2 \tilde{g}_M(T_0, \delta/2).
\]

Since \(2 \tilde{g}_L(T_0, \delta) = \tilde{g}_L(T_0, \delta/2)\) and \(2 \tilde{g}_M(T_0, \delta) = \tilde{g}_M(T_0, \delta/2)\), it follows that
\[
\mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |x_L(T_0, t) - 1 v'x_L(0)| \geq \Delta_{\max}(T_0, \delta) \right)
\]
\[
\leq \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} \left| \tilde{x}_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| \right.
\]
\[
\geq \tilde{g}_L(T_0, \delta/2) \right) + \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |\phi_M(T_0, t)| \right) \geq \tilde{g}_M(T_0, \delta/2).
\]
Finally, by Proposition 4 we have
\[
\frac{\delta}{2} \geq \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} \left| \tilde{x}_L(T_0, t) - \left( \prod_{k=T_0}^{t-1} W_L \right) x_L(0) \right| \right.
\]
\[
\geq \tilde{g}_L(T_0, \delta/2) \right) \geq \tilde{g}_M(T_0, \delta/2),\]
while, similarly, by Proposition 5 we have
\[
\frac{\delta}{2} \geq \mathbb{P} \left( \max_{i \in L} \limsup_{t \to \infty} |\phi_M(T_0, t)| \right) \geq \tilde{g}_M(T_0, \delta/2) \right) \geq \tilde{g}_M(T_0, \delta/2).
\]

Thus, we have solved Problem 2 by showing that beyond achieving convergence of consensus in the face of a malicious attack, the deviation of the converged value from the nominal average consensus value (in the case of no malicious agents) can be characterized as a function of user-defined parameters \(T_0\) and failure probability \(\delta\). In the next section we characterize convergence rate of our consensus protocol.

C. Convergence Rate of Resilient Consensus

In this section we discuss convergence rate for the consensus protocol in Eq. (2) using the weights as defined in Eq. (8). Let
\[
\|z\|_v = \sqrt{\sum_{i=1}^{n_L} v_i z_i^2}.
\]
We start by presenting a useful lemma showing a contraction property for a consensus step using a primitive matrix that we will later employ.

Lemma 5. Assume that \(j \in N_i\) if and only if \(i \in N_j\) for every \(i, j \in L\) (i.e., bidirectional links between legitimate agents). Let \(p_2\) denote the second largest eigenvalue modulus of \(W_L\) and let \(v > 0\) be the stochastic Perron vector satisfying \(v'W_L = v'\). Then,
\[
\|W_L\delta x(0) - 1 v'x(0)\|_v \leq p_2 \cdot \|x(0) - 1 v'x(0)\|_v.
\]

Proof: It can be seen that the Perron vector \(v\) has entries \(v_i = \max_{\|i\| = j \in L} \frac{\|i\| + |j|}{2}\) for all \(i\), implying that \((W_L, v)\) is reversible Markov chain. Thus, the result follows from the convergence rate of reversible Markov chains (see [61]).

We are now ready to prove our convergence rate results. Namely, we show that the rate of convergence of legitimate agents to their limit value is governed by the second largest eigenvalue modulus of the ideal weight matrix \(W_L\) with a probability that approaches 1 exponentially, as \(T_0\) increases.

Theorem 3 (Convergence Rate of Resilient Consensus). Consider the consensus system described by the dynamics in Eq. (2) with a arbitrary start time of \(T_0 \geq 0\) and weights as described by Eq. (8). Assume that 1) \(G\) is a sufficiently connected network and that 2) bounds on the expected values of the observations \(\alpha_{ij}\) (see Defn. (12)) satisfy \(c = E(\alpha_{ij}) - 1/2 < 0\) for all \(i \in L, j \in N_i \cap M\), for all \(k \geq 0\), and \(d = E(\alpha_{ij}) - 1/2 > 0\) for all \(i \in L, j \in N_i \cap L\), for all \(k \geq 0\). Let \(p_2\) denote the second largest eigenvalue modulus of \(W_L\) and let \(v > 0\) be the stochastic Perron vector satisfying \(v'W_L = v'\). Additionally, let \(m = \max_{i \in L, j \in M, i \in \mathbb{N}} |x_i(t)|\) and assume that \(j \in N_i\) if and only if \(i \in N_j\) for every \(i, j \in L\). For every \(T_0 > 0\), if we start the consensus protocol from time \(T_0\), then for every \(t \geq T_0 - 1\) and \(m = T_0 - 1, \ldots, t\), we have
\[
\|x_L(T_0, t) - z(T_0)1\|_v \leq 2(m - T_0 + 1)p_2^{-m-\eta}, \tag{36}
\]
\[2^{-}\text{The second largest absolute value of the eigenvalues.}\]
with a probability greater than
\[ 1 - |L|^2 \cdot \frac{\exp(-2(m + 1)d^2)}{1 - \exp(-2d^2)} - |L||M| \cdot \frac{\exp(-2(m + 1)c^2)}{1 - \exp(-2c^2)}. \] (37)

Proof: Recall that \( \lim_{t \to \infty} x_L(T_0, t) = z(T_0) \) almost surely. Since \( x_L(T_0, t) = \tilde{x}_L(T_0, t) + \phi_M(T_0, t) \), it follows by the triangle inequality that
\[ \|x_L(T_0, t) - z(T_0)1\|_T = \|x_L(T_0, t) - \lim_{\tau \to \infty} x_L(T_0, \tau)\|_T \]
\[ \leq \|\tilde{x}_L(T_0, t) - \lim_{\tau \to \infty} \tilde{x}_L(T_0, \tau)\|_T + \|\phi_M(T_0, t) - \lim_{\tau \to \infty} \phi_M(T_0, \tau)\|_T. \]

Now, assume that for all \( k \geq m \), \( W_L(k) = W_L \) and \( W_M(k) = 0 \). Then, by the Perron-Frobenius Theorem and the definition of \( W \),
\[ W_{L} \leq \left( \frac{1}{1 - \exp(-2d^2)} \right)^2 \]
\[ \text{for } t \geq T_0. \]

As an immediate consequence of Theorem 3, we have the probabilistic bounds can be made tighter and the resulting performance guarantees such as bounds on the deviation can be made stronger. This observation is substantiated with analysis in the current section.

Lemma 2 lays the foundation for deriving the convergence of consensus of legitimate agents presented in Theorem 1 and the probabilistic bound presented in Theorem 2 and the probabilistic upper bound on the convergence rate presented in Theorem 3. Lemma 2 considers a large family of probability measures for the random variables \( \alpha_{ij} \) where only the expectation of the variables are fixed to known values. While the bounds derived in Lemma 2 decay exponentially over time, they are somewhat loose when higher moments of the random variables \( \alpha_{ij} \) are known. In the special case when the random variables \( \alpha_{ij} \) belong to a family of probability measures of known variance values, we can use the concentration inequalities that consider both the first and second moments, instead of the Chernoff-Hoeffding concentration inequality used in the proof of Lemma 2. Prominent examples of such inequalities are the Berstein inequality, Bennett’s inequality and the improved Bennett’s inequality, providing more refined results than the Chernoff-Hoeffding inequality, are included here for completeness.

Theorem 4 (Berstein inequality). Assume that \( x_1, \ldots, x_n \) are independent random variables and \( E(x_i) = 0, \ E(x_i^2) = \sigma_i^2 \) and \( |x_i| \leq M \) almost surely. Then,
\[ \mathbb{P} \left( \sum_{i=1}^{n} x_i \geq b \right) \leq \exp \left( \frac{-\frac{b^2}{2M^2} + \frac{1}{3} \frac{b^2}{M^2}}{\frac{1}{2} \frac{\sigma_i^2}{M^2}} \right) \cdot \exp \left( \frac{-c^2}{2} \right). \]

Theorem 5 (Bennett’s inequality). Assume that \( x_1, \ldots, x_n \) are independent random variables and \( E(x_i) = 0, \ E(x_i^2) = \sigma_i^2 \) and \( |x_i| \leq M \) almost surely. Then, for any \( 0 \leq t < nM \),
\[ \mathbb{P} \left( \sum_{i=1}^{n} x_i \geq b \right) \leq \exp \left( -\frac{n\sigma_i^2}{M^2} h \left( \frac{bM}{\sigma_i^2} \right) \right), \]
where \( h(x) = (1 + x) \ln(1 + x) - x \) and \( n\sigma_i^2 = \sum_{i=1}^{n} \sigma_i^2. \)
Consider the random variables \( x_1, \ldots, x_n \) to be independent random variables and \( E(x_i) = 0 \), \( E(x_i^2) = \sigma_i^2 \) and \( |x_i| \leq M \) almost surely. Additionally, let
\[
A = \frac{M^2}{\sigma^2} + \frac{nM}{b} - 1 \quad \text{and} \quad B = \frac{nM}{b} - 1,
\]
and \( \Lambda = A - W(Be^A) \), where \( W(\cdot) \) is the Lambert W function. Let \( \sigma_i = \frac{1}{b} \sum_{i=1}^{n} \sigma_i^2 \), then for any \( 0 \leq b < nM \),
\[
P \left( \sum_{i=1}^{n} x_i \geq b \right) \leq \exp \left[ \frac{\Lambda b}{M} + n \ln \left( 1 + \frac{\sigma_i^2}{M^2} \left( e^A - 1 - \Lambda \right) \right) \right].
\]

As stated in [64], Bennet’s inequality yields strictly tighter approximation than Bernstein’s inequality. Bennet’s inequality, in turn, is tightened by the improved Bennet’s inequality. For this reason, we present in Lemma 6 refined concentration inequalities that improve those presented in Lemma 2 using the improved Bennet’s inequality. Recall Assumption 1.2 and Assumption 1.3, and additionally assume that
\[
\text{var}(\alpha_{ij}(t) - 1/2) = \sigma_i^2 \quad \text{for all } t \geq 0, \ i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{L},
\]
and that
\[
\text{var}(\alpha_{ij}(t) - 1/2) = \sigma_i^2 \quad \text{for all } t \geq 0, \ i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{M}.
\]

Applying these tighter concentration inequalities to the history of observations \( \beta_{ij}(t) = \sum_{k=0}^t (\alpha_{ij}(k) - 1/2) \), we can obtain a faster exponential rate result for the decay of misclassification probabilities (misclassifying a legitimate agent as malicious or a malicious agent as legitimate) than that of Lemma 2 as seen in the following.

**Lemma 6.** Consider the random variables \( \beta_{ij}(t) \) defined in Equation (6) for every legitimate node \( i \in \mathcal{L} \) and every neighbor \( j \in \mathcal{N}_i \). Then, for every \( t \geq 0 \) and every \( i \in \mathcal{L} \), \( j \in \mathcal{N}_i \cap \mathcal{L} \),
\[
P (\beta_{ij}(t) < 0) \leq \max \left\{ 1, \frac{c}{(d + 1)^2} \right\},
\]
\[
\exp \left[ -d(t + 1)A_d + (t + 1) \ln \left( 1 + \frac{\sigma_i^2}{d^2} \left( e^{A_d} - 1 - A_d \right) \right) \right],
\]
where \( d = E(\alpha_{ij}(t) - 1/2) \) for \( i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{L} \), and
\[
\Lambda_d = A_d - W(Bc^A_d),
\]
with \( W(\cdot) \) being the Lambert W function, and
\[
A_d = \frac{1}{\sigma_d} + \frac{1}{d} - 1, \quad B_d = \frac{1}{d} - 1.
\]
Moreover, for every \( t \geq 0 \) and every \( i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{M} \),
\[
P (\beta_{ij}(t) \geq 0) \leq \max \left\{ 1, \frac{c}{(e + 1)^2} \right\},
\]
\[
\exp \left[ -c(t + 1)A_c + (t + 1) \ln \left( 1 + \frac{\sigma_i^2}{e^2} \left( e^{A_c} - 1 - A_c \right) \right) \right],
\]
where \( c = E(\alpha_{ij}(t) - 1/2) \) for \( i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{M} \), and
\[
A_c = A_c - W(Bc^A_c), \quad \text{with scalars } A_c \text{ and } B_c \text{ given by}
\]
\[
A_c = \frac{1}{\sigma_c^2} - \frac{1}{e} - 1, \quad B_c = \frac{1}{e} - 1.
\]

**Proof:** Let \( i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{L} \), and note that the given probabilistic bound holds trivially when \( d < 0 \), so assume that \( d > 0 \) and define
\[
\gamma_{ij}(t) = d + 1/2 - \alpha_{ij}(t).
\]
Since \( d = E(\alpha_{ij}(t) - 1/2) \), for the random variable \( \gamma_{ij}(t) \) we have \( E(\gamma_{ij}(t)) = 0 \) and \( \text{var}(\gamma_{ij}(t)) = \sigma_i^2 \). Since \( \alpha_{ij}(t) \in [0, 1] \), it follows that \( d \leq 1/2 \) implying that \( \gamma_{ij}(t) \leq 1 \). Also, since \( d > 0 \) and \( \alpha_{ij}(t) \in [0, 1] \) it follows that \( \gamma_{ij} > -1/2 \). Thus, \( |\gamma_{ij}| \leq 1 \). Using the definition of \( \beta_{ij}(t) \) in (6), we have
\[
P (\beta_{ij}(t) < 0) = P \left( \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2) < 0 \right)
\]
\[
= P \left( \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2 - d) < -(t + 1)d \right)
\]
\[
= P \left( \sum_{k=0}^{t} \gamma_{ij}(t) < -(t + 1)d \right)
\]
\[
= P \left( \sum_{k=0}^{t} \gamma_{ij}(t) > (t + 1)d \right).
\]

We now invoke Theorem 6 for the sequence of random variables \( \gamma_{ij}(0), \ldots, \gamma_{ij}(t) \), with \( M = 1, n = t + 1 \) and \( b = (t + 1)d \), which yields the desired relation. Note that the condition \( b < nM \) in Theorem 6 is satisfied in our case here, since \( b = (t + 1)d \) with \( d \geq 1/2 \) and \( nM = t + 1 \).

Consider now the case \( i \in \mathcal{L}, \ j \in \mathcal{N}_i \cap \mathcal{M} \). If \( c > 0 \), then the result holds trivially. Thus, assume that \( c < 0 \), and consider
\[
\gamma_{ij}(t) = \alpha_{ij}(t) - 1/2 - c.
\]
Since \( c = E(\alpha_{ij}(t) - 1/2) \), we have that \( E(\gamma_{ij}(t)) = 0 \) and \( \text{var}(\gamma_{ij}(t)) = \sigma_i^2 \). Additionally, since \( \alpha_{ij}(t) \in [0, 1] \), we can see (similar to the preceding case) that \( |\gamma_{ij}(t)| \leq 1 \). Hence,
\[
P (\beta_{ij}(t) > 0) = P \left( \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2) \geq 0 \right)
\]
\[
= P \left( \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2 - c) \geq -(t + 1)c \right)
\]
\[
= P \left( \sum_{k=0}^{t} \gamma_{ij}(t) \geq -(t + 1)c \right).
\]

Recall that \( c < 0 \), so we can invoke Theorem 6 for the sequence of random variables \( \gamma_{ij}(0), \ldots, \gamma_{ij}(t) \), with \( \alpha = -(t + 1)c > 0 \), \( M = 1 \), and \( n = t + 1 \). Since \( c = E(\alpha_{ij}(t) - 1/2) \) and \( \alpha_{ij}(t) \in [0, 1] \), it follows that \( -1/2 \leq c < 0 \). Thus, in our case here, the condition \( b < nM \) in Theorem 6 reduces to \( -(t + 1)c < t + 1 \), which holds since \( -1/2 \leq c < 0 \).

Therefore, in the case that additional information on the observations \( \alpha_{ij} \) is available, in particular that a bound on the variance of the \( \alpha_{ij} \) is known, then the results of Lemma 2 can be replaced with Lemma 6, resulting in tighter deviation
and convergence rate results when applied to Theorems 2 and 3. This shows the generality of our framework and the potential to improve our derived performance guarantees as more information on the αi,j observations becomes available. Similarly, Lemma 5 can be extended to the case where var(βi,j(t)) varies over time and is link dependent.

IV. Numerical Results

In this section we evaluate the performance of the proposed scheme in this work using Monte Carlo simulations. We aim at investigating the effects, on several performance metrics, of the following system parameters: the number |ℳ| of malicious agents, the starting time T0 of the data passing stage and the variance of the random variables αi,j. Performance metrics that we study include deviation from the true consensus value and the convergence rate. We evaluate a system setup with 15 legitimate agents, and with κ = 10. We consider the following values for the starting time T0: 0, 25, 50, 100, 150. The adjacency matrix of the subgraph G≤ of links among legitimate agents is

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix},
\]  

(38)

and we assume that every malicious agent is connected to all the legitimate agents in the system which is the worst-case scenario. We depict the topology corresponding to the connectivity graph G in Figure 4 since every malicious agent is connected to all legitimate agents in Figure 4 by a single node. The vector of initial values of legitimate agents is

\[x_L(0) = (-2.59, -2.44, -4.23, -1.45, -1.46, 0.871, -0.51, -3.19, -0.59, -3.31, -2.25, 1.31, 1.87, 1.34, 1.76)'\]

where (·)’ denotes the transpose operator, and we use η = 5 as the upper bound for the agents’ initial values. Additionally, for every legitimate agent i, we have E(αi,j) = 0.55 if j ∈ N_i ∩ L, and E(αi,j) = 0.45 if j ∈ N_i ∩ M. The random variable αi,j is uniformly distributed on the interval \[E(αi,j) - \frac{1}{2}, E(αi,j) + \frac{1}{2}\]. We consider the following three values for ℓ: 0.2, 0.4, 0.6, for which the standard deviation values are 0.0577, 0.1155, 0.1732, respectively. Note that the larger the ℓ, the larger the variance of αi,j about its mean value. We consider the following input scenarios presented in Figures 5 and 6. Both attacks assume that the malicious agents are allowed to communicate and coordinate attacks.

a) Maximal deviation malicious input: To measure the maximal possible deviation from true consensus value we consider the case were the malicious agents inputs are the furthest from the true consensus value. That is, if the true consensus value is positive than the malicious input is −η, and if the true consensus value is negative than the malicious input is η. In our scenario, all the malicious agents choose the maximum possible input η = 5. Figures 5 and 7 measure the effect of the malicious agent inputs that lead to the maximum deviation from the true consensus value. Figure 5 depicts agent values over time when running consensus according to the protocol from Equation (1) and the weights defined by Equation (8). Specifically, Figure 5 depicts the value of the agent with the maximal deviation from the true consensus value for a single system realization with maximum deviation malicious input. Additionally, Figure 7 additionally show the differences in converged values and deviation respectively for different consensus start times T0.

The maximum deviation malicious input leads to an upper bound to all possible malicious attacks, and depicts a worst case scenario in terms of deviation from true consensus value. However, though this attack is useful to analyze the worst case effect of malicious inputs it is easy to detect using outlier rejection methods for example [8]. For this reason, we consider the following attack scenario that is harder to detect.

b) Drift malicious input: Figure 6 and 8 measure the effect of a less transparent malicious agents’ attack where every malicious agent adds a drift term to the average legitimate agent values. More specifically, let d_L(m, t) be the average value at time t of the legitimate agents that are neighbors of the malicious agent m. Denote d_L(t) = \[\frac{1}{|M|} \sum_m d_L(m, t)\]. Then, at time t, the malicious agent m chooses its state value \[x_m(t)\] according to the following rule:

\[x_m(t) = d_L(t - 1) + 0.15 \times d_M(m, t),\]
where \( x_m(t = T_0 - 1) \) is chosen randomly from the set \([-0.15 \times \eta, 0.15 \times \eta]\) and for every \( t \geq T_0 \)

\[
d_x(m, t) = -\text{sign}(d_L(1)) \times \eta \times (0.75)^{0.05(t-T_0)} u(m, t),
\]

\( u(m, t) \sim U[0, 1] \). Thus malicious agent inputs are time varying and strategically close to legitimate agent values to avoid easy detection in this case. Additionally, if \( x_m(t) > \eta \) then \( x_m(t) \) is generated randomly and uniformly from the interval \([\eta-0.05, \eta]\). Similarly if \( x_m(t) < -\eta \) then \( x_m(t) \) is generated randomly and uniformly from the interval \([-\eta, -\eta+0.05]\).

We note that since in our scenario every malicious agent is a neighbor of all the legitimate agents, \( d_x(m_1, t) = d_x(m_2, t) \) for all \( m_1, m_2 \in \mathcal{M} \). Figure 6 depicts the value of the agent with the maximal deviation from the true consensus value for a single system realization with drift malicious input. Additionally, Figure 8 depicts the resulting absolute value of the maximum deviation from the true consensus value for a drift malicious input averaged over 500 system realizations.

Figures 5 and 6 depict a single realization of the agent value with the maximal deviation from true consensus value for a system with 15 malicious agents and \( \ell = 0.4 \). Figure 5 depicts the agent value with the maximal deviation from true consensus value for the maximal deviation malicious inputs where Figure 6 depicts the agent value with the maximal deviation from true consensus value for the drift malicious inputs.

In addition to Figures 5 and 6 that depict the legitimate agent values for a single system realization we depict in Figures 7 and 8 the average deviation of the legitimate agent values from true consensus value. The average deviation is calculated using 500 system realizations. Figures 7 and 8 show that the deviation from true consensus value grows both as \( \ell \) grows and as the number \( |\mathcal{M}| \) of malicious agents grows. In each figure we label the plot by (R#,C#) where R# denotes the row number and represents the increase in the number of malicious agents, and C# denotes the column number and represents the growth of \( \ell \). Figures 7 and 8 show the significance of introducing the variable \( T_0 \) that sets the starting time of the data passing between agents. The figures indicate that increasing \( T_0 \) can significantly reduce the deviation from true consensus value. Figures 7 and 8 also show that our scheme can combat an attack of a large number of malicious agents, even when \( |\mathcal{M}| = 30 \), that is, there is twice the number of malicious agents than legitimate agents in the system. This is a significant improvement over classical results that depend on the network connectivity [1], [5], [6]. Using the upper bound \( \min_{i \in \mathcal{G}(\mathcal{M})} |\mathcal{N}_i| \) on the network connectivity we can conclude that classical results will fail to detect all the malicious agents when their number is larger than the number of legitimate agents; in this case the undetected malicious agents can control the data values of the legitimate agents. More specifically, we can see from the choice of adjacency matrix (38) that for our setup, the network connectivity is at most \( 3 + |\mathcal{M}| \) and thus classical results can tolerate less than \( \frac{2}{3+|\mathcal{M}|} \) malicious agents. Since there are \( |\mathcal{M}| \) malicious agents, the consensus algorithm can fail and be controlled by the malicious agents. Furthermore, for the connectivity graph that we consider in the numerical results, the maximal number of malicious agents leading to a graph connectivity guaranteeing their detection and preventing them from controlling the network is 2, since we must fulfill the condition that \( |\mathcal{M}| < \frac{2}{3+|\mathcal{M}|} \). Finally, our numerical results show that there exists a finite time \( T_0 \) such that the probability to classify all the user in the network correctly for every \( t > T_0 \) is sufficiently high. We can see from Figures 7 and 8 that the minimal value of \( T_0 \) that yields a negligible deviation from true consensus value depends on probability of miss-classifying users, i.e., classifying a legitimate agent as malicious and vice-versa. This probability is increased, for example, as the number of malicious agents grows and as the variance of \( \alpha_{ij} \) grows. Figures 7 and 8 show that as we increase the number of malicious agents and the deviation \( \ell \) that governs the variance of \( \alpha_{ij} \) we should increase \( T_0 \) to limit the effect of the malicious agents on the consensus value. Additionally, Figures 7 and 8 show that the for \( T_0 = 150 \) the deviation from true consensus value for all considered scenarios is negligible. Finally, for \( T_0 = 100 \) the deviation from true consensus value is sufficiently small and the effect of malicious agents is limited even for \( |\mathcal{M}| = 30 \).

Figure 7 shows that as \( \ell \) becomes larger (and therefore also
Figure 7. Maximal deviation input the variance), the effect of malicious agents on deviation gets worse. However, even in this case increasing $T_0$ can recover the nominal consensus value with deviation approaching zero. The exact relationship between the deviation $\Delta_{\text{max}}(T_0, \delta)$ and $T_0$ is given by Theorem 2.

In the case of a time varying malicious agent input as with the drift malicious input model described in part (b) of this section, nominal average consensus can also be recovered by making $T_0$ larger, as depicted in Figure 8. This is true even with variable variance on the $\alpha_{ij}$ values.

V. DISCUSSION

In this section we discuss the significance of some of our characterizations on the trust observations $\alpha_{ij}$ and some of our assumptions. We also discuss directions for future work. Firstly, the results of this paper assume a fixed network topology. However, we believe that it is possible to relax Assumption 1.1 to include the case where the graph over legitimate agents is not always connected. In such a case Assumption 1.1 can be replaced with the assumption that the topology of the legitimate agents is $B$-connected. We believe that the convergence results from [65, Proposition 1] can be adapted to this case along the same vein as in Theorem 3. We now comment on independence of the $\alpha_{ij}$ values which is an important assumption for the application of concentration inequalities such as those used in our derivations.

3See definition in [29].
Conditioned on \( j \) being legitimate or spoofed, computation of the \( \alpha_{ij}(t) \) values will be subject to noise experienced over the wireless channel. This is widely assumed to be Gaussian white noise \([48]\) that is independent with respect to time. In addition, for uniform scattering environments signal paths decorrelate spatially \([47]\). However, it is possible that since the \( \alpha_{ij} \) terms also have a dependence on the environment, in the uncommon case that the environment and all robots stay perfectly static, then complete independence may no longer hold. Relaxing these assumptions may also be an interesting avenue for future work. Finally we discuss a few interesting observations about the trust values \( \alpha_{ij} \). Our results show that the more informed our \( \alpha_{ij} \) values are, the tighter the performance guarantees achievable become (cf. Lemma \([6]\)). This also opens a future avenue of investigation of finding \( \alpha_{ij} \) observations from physical channels in multi-agent systems which can be characterized as fully as possible. The theory in this paper provides the mathematical framework for understanding which characteristics of \( \alpha_{ij} \) are most critical for attaining performance guarantees that are important for multi-agent coordination. We hope that this can guide continued investigation into trust-based resilience, particularly for cyber-physical systems where additional information on inter-agent trust can be attained.

VI. CONCLUSION

This paper presents a unified mathematical theory to treat distributed multi-agent consensus in the presence of malicious
agents when stochastic values of trust between agents are available. Under this model, we present three new performance guarantees for consensus systems in the presence of malicious agents: 1) convergence of consensus is possible with probability 1 even when the number of malicious agents is far larger than 1/2 of the network connectivity—in contrast to classical results in resilience, 2) the deviation of the converged value from average consensus can be bounded and further, can be driven arbitrarily close to zero by using our derived edge weights and by allowing an observation time $T_0$ of trust values over the network, and 3) convergence to the correct classification of agents can be attained in finite time with probability 1 with an exponential rate that we derive. Additionally, we show that our performance guarantees can be made stronger if more information on the trust observations, such as bounds on their variance, is available. Taken together, these results point to the inherent value of quantifying and exploiting trust between agents for consensus. We believe that the mathematical formulations and framework of this paper hold promise for achieving a new generation of resilient multi-agent systems.

References

[1] D. Dolev, “The byzantine generals strike again,” Journal of Algorithms, vol. 3, no. 1, pp. 14 – 50, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0196677482900049
[2] M. J. Fischer, N. A. Lynch, and M. S. Paterson, “Impossibility of distributed consensus with one faulty process,” J. ACM, vol. 32, no. 2, pp. 374–382, Apr. 1985.
[3] L. Lamport, “Paxos made simple,” ACM SIGACT News, vol. 32, no. 4, pp. 18–25, Dec. 2001.
[4] R. D. Prisco, B. Lampson, and N. Lynch, “Revisiting the paxos algorithm,” Theoretical Computer Science, vol. 243, no. 1, pp. 35 – 91, 2000.
[5] S. Sundaram and C. N. Hadjicostis, “Distributed function calculation via linear iterative strategies in the presence of malicious agents,” IEEE Transactions on Automatic Control, vol. 56, no. 7, pp. 1495–1508, 2011.
[6] F. Pasqualetti, A. Bicchi, and F. Bullo, “Consensus computation in unreliable networks: A system theoretic approach,” IEEE Transactions on Automatic Control, vol. 57, no. 1, pp. 90–104, Jan. 2012.
[7] F. Higgins et al., “Threats to the swarm: Security considerations for swarm robotics,” International Journal on Advances in Security, 2009.
[8] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, “Robust asymptotic consensus in robust networks,” IEEE Journal on Selected Areas in Communications, vol. 31, no. 4, pp. 766–781, April 2013.
[9] R. Olfati-Saber and R. M. Murray, “Consensus problems in networks of agents with switching topology and time-delays,” IEEE Transactions on Automatic Control, vol. 51, no. 3, pp. 401–415, 2006.
[10] J. N. Tsitsiklis, “Convergence of type-symmetric stochastic consensus algorithms,” SIAM Journal on Optimization, vol. 7, no. 4, pp. 918–937, 1997.
[11] L. Lamport, “Time, clocks, and the ordering of events in a distributed system,” Communications of the ACM, vol. 21, no. 7, pp. 558–565, 1978.
[12] R. Olfati-Saber and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” Proc. IEEE, vol. 95, no. 1, pp. 215–233, 2007.
[13] J. N. Tsitsiklis, “Convergence of type-symmetric stochastic consensus algorithms,” SIAM Journal on Optimization, vol. 17, no. 2, pp. 568–585, 2006.
[14] L. Moreau, “Stability of multi-agent systems with time-delays,” SIAM Journal on Control and Optimization, vol. 47, no. 3, pp. 1677–1691, 2008.
[15] J. M. Hendrickx and J. N. Tsitsiklis, “Convergence of type-symmetric and cut-balanced consensus seeking systems,” IEEE Transactions on Automatic Control, vol. 58, no. 1, pp. 245–249, 2013.
[16] M. M. Zavlanos, M. B. Egerstedt, and G. J. Pappas, “Graph-theoretic connectivity control of mobile robot networks,” Proceedings of the IEEE, vol. 99, no. 9, pp. 1550–1560, 2011.
[17] J. Xiong and K. Jamieson, “Securearray: Improving wifi security with fine-grained physical-layer information,” in Proceedings of the 19th Annual International Conference on Mobile Computing and Networking, ser. MobiCom, 2013.
[18] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, “Rf-compass: Robot object manipulation using rfid,” ser. MobiCom, 2013.
[19] J. Xiong and K. Jamieson, “Securearray: Improving security with fine-grained physical-layer information,” in Proceedings of the 19th Annual International Conference on Mobile Computing and Networking, ser. MobiCom, 2013.
[20] J. Fax and R. Murray, “Information flow and cooperative control of multi-agent networks,” Proc. IEEE, vol. 95, no. 1, pp. 215–233, 2007.
[21] J. Fax and R. Murray, “Information flow and cooperative control of multi-agent networks,” Proc. IEEE, vol. 95, no. 1, pp. 215–233, 2007.
[22] M. M. Zavlanos, M. B. Egerstedt, and G. J. Pappas, “Graph-theoretic connectivity control of mobile robot networks,” Proceedings of the IEEE, vol. 99, no. 9, pp. 1550–1560, 2011.
[23] M. S. Kumar, E. Hamed, D. Katabi, and L. Erran Li, “Lte radio analytics made easy and accessible,” in SIGCOMM, 2014.
[50] S. Kumar et al., “Accurate indoor localization with zero start-up cost,” ser. MobiCom, 2014.
[51] N. Jadhav, W. Wang, D. Zhang, O. Khatib, S. Kumar, and S. Gil, “WSR: A WiFi sensor for collaborative robotics,” in arXiv:2012.04174, 2020.
[52] K. Zeng, K. Govindan, and P. Mohapatra, “Non-cryptographic authentication and identification in wireless networks [security and privacy in emerging wireless networks],” IEEE Wireless Communications, vol. 17, no. 5, pp. 56–62, 2010.
[53] T. Wang and Y. Yang, “Analysis on perfect location spoofing attacks using beamforming,” in INFOCOM, 2013.
[54] D. Moser, P. Leu, V. Lenders, A. Ranganathan, F. Ricciato, and S. Capkun, “Investigation of multi-device location spoofing attacks on air traffic control and possible countermeasures,” in Proceedings of the 22Nd Annual International Conference on Mobile Computing and Networking, ser. MobiCom ’16. New York, NY, USA: ACM, 2016, pp. 375–386.
[55] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, “Adaptive communication in multi-robot systems using directionality of signal strength,” IJRR, 2015.
[56] S. Gil, C. Baykal, and D. Rus, “Resilient multi-agent consensus using wi-fi signals,” IEEE Control Systems Letters, vol. 3, no. 1, pp. 126–131, Jan 2019.
[57] T. Wheeler, E. Bharathi, and S. Gil, “Switching topology for resilient consensus using wi-fi signals,” in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 2018–2024.
[58] D. P. Dubhashi and A. Panconesi, “Concentration of measure for the analysis of randomized algorithms.” Cambridge University Press, 2009.
[59] S. U. Pillai, T. Suel, and Seunghun Cha, “The perron-frobenius theorem: some of its applications,” IEEE Signal Processing Magazine, vol. 22, no. 2, pp. 62–75, 2005.
[60] R. Durrett, Probability: Theory and Examples, 4th ed. Cambridge University Press, 2010.
[61] P. Brémaud, Gibbs Fields, Monte Carlo Simulation, and Queues. New York, USA: Springer-Verlag, 1999.
[62] G. Bennett, “Probability inequalities for the sum of independent random variables,” Journal of the American Statistical Association, vol. 57, no. 297, pp. 33–45, 1962.
[63] S. Zheng, “An improved bennett’s inequality,” Communications in Statistics - Theory and Methods, vol. 47, no. 17, pp. 4152–4159, 2018.
[64] T. Jebara, “A refinement of Bennett’s inequality with applications to portfolio optimization,” in arXiv:1804.05454, 2018.
[65] A. Nedić and A. Ozdaglar, “Convergence rate for consensus with delays,” Journal on Global Optimization, vol. 47, p. 437–456, 2010.

ACKNOWLEDGMENT

The authors gratefully acknowledge partial funding support from NSF CAREER grant #1845225.