2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Multi-Agent Resilient Consensus under Intermittent Faulty and Malicious Transmissions

Sarper Aydın, Orhan Eren Akgün, Stephanie Gil, and Angelia Nedić

Abstract— **In this work, we consider the consensus problem in which legitimate agents share their values over an undirected communication network in the presence of malicious or faulty agents. Different from the previous works, we characterize the conditions that generalize to several scenarios such as intermittent faulty or malicious transmissions, based on trust observations. As the standard trust aggregation approach based on a constant threshold fails to distinguish intermittent malicious/faulty activity, we propose a new detection algorithm utilizing time-varying thresholds and the random trust values available to legitimate agents. Under these conditions, legitimate agents almost surely determine their trusted neighborhood correctly with geometrically decaying misclassification probabilities. We further prove that the consensus process converges almost surely even in the presence of malicious agents. We also derive the probabilistic bounds on the deviation from the nominal consensus value that would have been achieved with no malicious agents in the system. Numerical results verify the convergence among agents and exemplify the deviation under different scenarios.**

I. Introduction

In this paper we are interested in the consensus problem [1], [2] in cyberphysical multi-agent systems under intermittent malicious attacks or failures. Agents need to reach an agreement over a set of variables using only local computation and communicating over a static undirected graph in the presence of malicious (non-cooperative) agents. Consensus algorithms constitute a basis for distributed decision-making in networked multi-agent systems [3], and are relevant for many multi-agent coordination applications, such as determining heading direction, rendezvous, and velocity agreement [4]–[6]. However, consensus algorithms that assume all agents are cooperative are known to be susceptible to malicious and faulty behaviors [7], [8]. Our goal in this work is to develop a resilient consensus algorithm utilizing "trust observations" for intermittent malicious and faulty behavior.

Achieving resilient consensus in the presence of malicious agents has been studied extensively in the literature. Earlier methods that only use the transmitted data to detect or eliminate untrustworthy information impose restrictions on the connectivity of the network and the number of tolerable malicious agents [8], [9]. As these fundamental limitations apply to other distributed computation algorithms [10], [11], researchers have explored leveraging additional information, that can be obtained from the physicality of the system, to assess the trustworthiness of the agents via stochastic trust observations $\alpha_{ij}(t) \in [0, 1]$ indicating the trustworthiness of a link (i, j) [12]–[16].

Previous work [17] shows that agents can detect untrustworthy agents with static behavior over time using trust observations with a predetermined threshold value and reach consensus even when malicious agents are in the majority. However, this ability breaks under *intermittent* attacks of the malicious agents, occurring infinitely many times with a constant positive probability at each time. In certain cases, malicious agents can inflict more damage to distributed systems by attacking randomly instead of attacking all the time [18], [19]. Notably, the detection ability with a constant threshold is compromised even for *unintentional* behavior such as intermittent failures due to noisy sensors leading to incorrect location reporting. The reason is that intermittent behavior results in a mixture of trustworthy and untrustworthy transmissions, precluding the ability to differentiate an attacker from a legitimate agent by using a constant threshold as was the case in previous works [17], [20]. Standard statistical tests necessitate the knowledge and certain forms of the distributions where samples are drawn, e.g. their moments and continuity [21], [22]. However, such properties may not be available to agents or may not hold with intermittent malicious transmissions, leading to the unavailability of convergence guarantees for standard statistical tests.

We address these challenges by proposing a new detection algorithm and a consensus method providing resilience against intermittent attacks and failures. Our detection method (Algorithm 1) utilizes the key point that legitimate agents' trust observations are sampled from the same distribution and that their expectations are higher than the malicious agents even when they act intermittently malicious. In the proposed algorithm, agents accumulate trust values from neighbors over time. Each round, they select their most trusted neighbor (the one with the highest aggregate trust value) as a reference and construct a trusted neighborhood by comparing other agents' aggregate trust values with the most trusted neighbor. Agents employ an adaptive threshold that grows over time, allowing them to exclude all malicious agents eventually, while still keeping their legitimate neighbors in their trusted neighborhood. Agents perform consensus updates using the values coming from their trusted neighbors only. Under the assumption that all legitimate agents have at least one legitimate neighbor, we demonstrate that the probability of agents misclassifying their neighbors decreases geometrically over time, result-

S. Aydın, O. E. Akgün and S. Gil are with the School of Engineering and Applied Sciences, Harvard University, Allston, MA 02134. E-mail: saydin@seas.harvard.edu, erenakgun@g.harvard.edu, sgil@seas.harvard.edu. A. Nedić are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281. E-mail: angelia.nedich@asu.edu. This work has been supported by the NSF awards CNS-2147641 and CNS-2147694.

ing in a period after which no classification errors occur (Lemmas 1-2). Moreover, we show that the legitimate agents reach consensus almost surely, and their deviation from the consensus value is bounded (Corollary 1). We derived the maximal deviation from the nominal consensus process for a predetermined error tolerance (Theorem 1). In summary, our contributions are two-fold, *i)* we propose a novel detection algorithm for removing faulty and malicious activity in finite time, and *ii)* we show the convergence of the consensus process and analyze the deviation from nominal consensus. Numerical results also corroborate our theoretical analysis in different scenarios.

II. Consensus Dynamics with Failures and Attacks

A. Notation

We use $|.|$ to denote absolute values of scalars and cardinalities of sets. We write $[.]_i$ and $[.]_{ij}$ for the i^{th} entry of a vector and the ij-th entry of a matrix, respectively. We also extend the notation $|.|$ to matrices/vectors to define the element-wise absolute value of matrices/vectors, e.g., $[|A|]_{ij} = |[A]_{ij}|$. For matrices A and B, we write $A > B$ (or $A \geqslant B$) when $[A]_{ij} > [B]_{ij}$ (or $[A]_{ij} \geqslant [B]_{ij}$) for all i, j. We use $\mathbf{0}$ and $\mathbf{1}$ to represent vectors/matrices whose entries are all 0 and 1, respectively.

We also use the backward matrix product of the matrices $H(k)$, defined as follows:

$$\prod_{k=\tau}^{t} H(k) := \begin{cases} H(t)\cdots H(\tau+1)H(\tau) & \text{if } t \geqslant \tau, \\ I & \text{otherwise,} \end{cases} \quad (1)$$

where I corresponds to the identity matrix.

B. Consensus in Presence of Untrustworthy Agents

We study the consensus dynamics among multiple agents defined by the set $\mathcal{N} := \{1, \ldots, N\}$. The agents send and receive information through a static undirected graph $G(\mathcal{N}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ represents the set of undirected edges among the agents. For each agents i, the set of neighboring agents is denoted by $\mathcal{N}_i := \{j \in \mathcal{N} : (i, j) \in \mathcal{E}\}$. The agent set $\mathcal{N}$ consists of legitimate agents who are always trustworthy and malicious agents who can be trustworthy or not. The set of legitimate agents is denoted by $\mathcal{L}$, while the set of malicious agents is denote by $\mathcal{M}$, with $\mathcal{L} \cup \mathcal{M} = \mathcal{N}$ and $\mathcal{L} \cap \mathcal{M} = \varnothing$. These sets are fixed over time and assumed to be unknown. The legitimate agents have associated nonnegative weights, subject to changes over time, for the existing communication links such that $w_{ij}(t) \in [0, 1]$ if $(i, j) \in \mathcal{E}$, otherwise $w_{ij}(t) = 0$. The consensus dynamics among the agents starts at some time $T_0 \geqslant 0$, and we model the dynamic for the legitimate agents, as follows: for all $i \in \mathcal{L}$ and for all $t \geqslant T_0 - 1$,

$$x_i(t+1) = w_{ii}(t)x_i(t) + \sum_{j \in \mathcal{N}_i} w_{ij}(t)x_j(t), \quad (2)$$

where $x_i(t) \in \mathbb{R}$ for all $i \in \mathcal{L}$. According to this update rule, each legitimate agent $i \in \mathcal{L}$ takes a convex combination of its value and its neighbors, i.e. $w_{ii}(t) > 0$, $w_{ij}(t) \geqslant 0$, and $w_{ii}(t) + \sum_{j \in \mathcal{N}_i} w_{ij}(t) = 1$. Since the consensus update starts at time T_0, we assume that $x_i(0) = x_i(t)$ for all $0 \leqslant t < T_0$. The dynamic of the malicious agent's values is assumed to be unknown even in the case they are not actively attacking, and it is not modeled.

We define $x(t) \in \mathbb{R}^N$ as a vector of agents' values at time t. Given the partition of the agents as legitimate and malicious, we partition the vector $x(t)$ accordingly, i.e., $x(t) = [x_{\mathcal{L}}(t), x_{\mathcal{M}}(t)]^T$. Then, the consensus dynamics (2) can be written in a vector notation:

$$x_{\mathcal{L}}(t+1) = \begin{bmatrix} W_{\mathcal{L}}(t) & W_{\mathcal{M}}(t) \end{bmatrix} \cdot \begin{bmatrix} x_{\mathcal{L}}(t) \\ x_{\mathcal{M}}(t) \end{bmatrix}, \quad (3)$$

where $W_{\mathcal{L}}(t) \in \mathbb{R}^{|\mathcal{L}| \times |\mathcal{L}|}$ and $W_{\mathcal{M}}(t) \in \mathbb{R}^{|\mathcal{L}| \times |\mathcal{M}|}$ are the weight matrices associated with legitimate and malicious agents. Hence, the consensus dynamics of legitimate agents can be written as a sum of two terms at any time $t \geqslant T_0$,

$$x_{\mathcal{L}}(T_0, t) = \tilde{x}_{\mathcal{L}}(T_0, t) + \phi_{\mathcal{M}}(T_0, t), \quad (4)$$

where

$$\tilde{x}_L(T_0, t) = \Big(\prod_{k=T_0-1}^{t-1} W_{\mathcal{L}}(k) \Big) x_{\mathcal{L}}(0), \quad (5)$$

$$\phi_{\mathcal{M}}(T_0, t) = \sum_{k=T_0-1}^{t-1} \Big(\prod_{\ell=k+1}^{t-1} W_{\mathcal{L}}(\ell) \Big) W_{\mathcal{M}}(k) x_{\mathcal{M}}(k). \quad (6)$$

Here, the term $\tilde{x}_L(T_0, t)$ represents the influence of legitimate agents on each other and the term $\phi_{\mathcal{M}}(T_0, t)$ represents the influence of malicious agents on the legitimate agents' values. These relations in (4)-(6) are the backbone of the subsequent analysis, as they capture the consensus dynamics of the legitimate agents in terms of the starting time T_0, the initial values $x(0)$, together with the malicious inputs $x_{\mathcal{M}}(k)$.

We assume that the values $x_i(t)$ of all agents are bounded, i.e., $|x_i(t)| \leqslant \eta$ by a scalar $\eta > 0$ for all $i \in \mathcal{N}$, and this value is known by all agents. Under this assumption, no malicious agent will ever send a value outside the interval $[-\eta, \eta]$ otherwise it will be immediately detected. This assumption is crucial for bounding the cumulative impact of malicious inputs, as captured by $\phi_{\mathcal{M}}(T_0, t)$ in (6).

C. Trusted Neighborhood Learning

Each legitimate agent $i \in \mathcal{L}$ aims to classify its legitimate neighbors $\mathcal{N}_i^{\mathcal{L}} := \mathcal{N}_i \cap \mathcal{L}$ and malicious neighbors $\mathcal{N}_i^{\mathcal{M}} := \mathcal{N}_i \cap \mathcal{M}$ correctly over time by gathering trust values $\alpha_{ij}(t)$ for each transmission from their neighbors $j \in \mathcal{N}_i$ (see [12] for more details on how to compute the trust values $\alpha_{ij}(t)$). The values $\alpha_{ij}(t)$, $t \geqslant 0$, are random with values in the unit interval, i.e., $\alpha_{ij}(t) \in [0, 1]$ for all $j \in \mathcal{N}$ and all $t \geqslant 0$, where higher $\alpha_{ij}(t)$ values ($\alpha_{ij}(t) \to 1$) indicate the event that a neighbor j is legitimate, is more likely.

The legitimate agents utilize the observed trust values $\{\alpha_{ij}(k)\}_{0 \leqslant k \leqslant t}$ to determine their trustworthy neighbors and select the weights $w_{ij}(t)$ at time t. Following the work in [17], we use the aggregate trust values, i.e.,

$$\beta_{ij}(t) = \sum_{k=0}^{t} (\alpha_{ij}(k) - 1/2), \text{ for all } i \in \mathcal{L} \text{ and } j \in \mathcal{N}_i. \quad (7)$$

We make the following assumption on the trust values $\alpha_{ij}(t)$.

Assumption 1 *Suppose that the following statements hold.*

(i) The expected value of malicious and legitimate transmissions sent from a neighbor agent $j \in \mathcal{N}_i$ and received by a legitimate agent $i \in \mathcal{L}$ are constant over time t and satisfy

$$c_j = \mathbb{E}(\alpha_{ij}(t)),\ j \in \mathcal{M}, \tag{8}$$
$$d = \mathbb{E}(\alpha_{ij}(t)),\ j \in \mathcal{L}, \tag{9}$$

where $d - c_j > 0$ for all $j \in \mathcal{M}$.

(ii) The random variables $\alpha_{ij}(t)$ observed by a legitimate agent $i \in \mathcal{L}$ are independent and identically distributed for a given agent $j \in \mathcal{N}_i$ at any time index $t \in \mathbb{N}$.

(iii) The subgraph $G_{\mathcal{L}} = (\mathcal{L}, \mathcal{E}_{\mathcal{L}})$ induced by the set of legitimate agents $\mathcal{L}$ is connected, where $\mathcal{E}_{\mathcal{L}} := \{(i,j) \in \mathcal{E} : (i,j) \in \mathcal{L} \times \mathcal{L}\}$.

Assumption 1-*(i)* captures the scenarios where each malicious agent $m \in \mathcal{M}$ transmits malicious information with a (nonzero) probability $p_m \in (0, 1]$ at each time t, and it send legitimate values with probability $1 - p_m$, whereas a legitimate agent $l \in \mathcal{L}$ never exhibits malicious behavior. It also holds when a malicious agent periodically sends malicious information in (deterministic) bounded time intervals. As a result, different and unknown rates of malicious transmissions correspond to mixed distributions with different expectations for the malicious agents' trust values. Assumption 1-*(ii)* requires independent trust samples of each neighbor from identical distributions; note that distributions of trust values given an agent $j \in \mathcal{N}_i$ are identical, while these distributions can be a mixture of several distributions. Assumption 1-*(iii)* imposes the connectivity among legitimate agents with a fixed topology. This assumption is consistent with the existing work leveraging trust values [12], [17], [20], [23] and is more relaxed than connectivity assumptions in other resilient multi-agent system works [7]–[10].

Unlike the work in [17], *we assume that the legitimate agents do not have any apriori threshold values to determine their trusted neighborhood*. This phenomenon can arise due to the dynamic behavior of the malicious agents. To handle the situation when an apriori threshold is unavailable, we propose a new learning method that legitimate agents implement to identify their trustworthy neighbors over time. The algorithm is built on three properties, (1) all legitimate agents have at least one legitimate neighbor, i.e., $\mathcal{N}_i^{\mathcal{L}} \neq \varnothing$, (Assumption 1-*(iii)*) (2) the legitimate agents have identical aggregate trust values in expectation, and (3) the legitimate agents have higher trust values compared to malicious agents in expectation (see Assumption 1-*(i)*). Based on property (3), in the algorithm, each legitimate agent chooses the highest aggregate trust value as a reference point. Then, it eliminates the malicious agents based on the unbounded (expected) difference of trust value aggregates between a legitimate and a malicious agent, as $t \to \infty$ (based on Assumption 1-*(i)*). Algorithm 1 is provided as below.

Algorithm 1 Trusted Neighborhood Learning

1: **Input:** Threshold value $\xi > 0$, $\gamma \in (0.5, 1)$.
2: Each agent $i \in \mathcal{L}$ finds $\bar{j}(t) = \arg\max_{j \in \mathcal{N}_i} \beta_{ij}(t)$.
3: Each agent $i \in \mathcal{L}$ checks if $\beta_{i\bar{j}(t)}(t) - \beta_{ij}(t) \leqslant \xi_t$, where $\xi_t = \xi(t+1)^{\gamma}$, for all $j \in \mathcal{N}_i$.
4: Each agent $i \in \mathcal{L}$ returns $\mathcal{N}_i(t) = \{j \in \mathcal{N}_i | \beta_{i\bar{j}(t)}(t) - \beta_{ij}(t) \leqslant \xi_t\}$.

In words, each legitimate agent $i \in \mathcal{L}$ first selects its most trusted neighbor (Step 2). Then, it compares others' trusted values with the most trusted agent (Step 3), and finally determines its trusted neighborhood with time-varying threshold values (Step 4). The chosen range for γ ensures the threshold grows slow enough to exclude malicious agents while maintaining a pace that retains legitimate agents over time. The rationale behind this selection will become more evident in Lemma 1 and Lemma 2 later on.

Next, we define the actual weights $w_{ij}(t) = [W(t)]_{ij}$ assigned by legitimate agents $i \in \mathcal{L}$ based on their learned trusted neighborhoods $\mathcal{N}_i(t)$, as below,

$$w_{ij}(t) = \begin{cases} \frac{1}{n_{w_i}(t)} & \text{if } j \in \mathcal{N}_i(t), \\ 1 - \sum_{\ell \in \mathcal{N}_i(t)} w_{i\ell}(t) & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $n_{w_i}(t) = \max\{|\mathcal{N}_i(t)| + 1, \kappa\} \geqslant 1$ and $\kappa > 0$ is a parameter that limits the influence of other agents on the values $x_i(t)$. Similarly, we define the matrix $\overline{W}_{\mathcal{L}}$ that would have been constructed if the legitimate agents have known their trusted neighbors, i.e., for the pairs of agents $(i,j) \in \mathcal{L} \times \mathcal{N}$,

$$[\overline{W}_{\mathcal{L}}]_{ij} = \begin{cases} \frac{1}{\max\{|\mathcal{N}_i^{\mathcal{L}}|+1, \kappa\}} & \text{if } j \in \mathcal{N}_i^{\mathcal{L}}, \\ 1 - \frac{1}{\max\{|\mathcal{N}_i^{\mathcal{L}}|+1, \kappa\}} & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The (nominal) matrix $\overline{W}_{\mathcal{L}}$ is the ideal and target case for each legitimate agent to eliminate the effect of malicious agents in the consensus process defined in Eqs. (4)-(6).

III. Analysis of Consensus Dynamics

A. Convergence of Consensus Dynamics

We start by analyzing the probability that a legitimate agent i misclassifies one of its neighbors at some time t in Algorithm 1. This misclassification can occur in two ways. A legitimate agent i can misclassify one of its legitimate neighbors $j \in \mathcal{N}_i^{\mathcal{L}}$ as malicious, resulting in agent j being excluded from the trusted neighborhood $\mathcal{N}_i(t)$. Conversely, agent i can misclassify one of its malicious neighbors $m \in \mathcal{N}_i^{\mathcal{M}}$ as legitimate, resulting in agent m being included in the trusted neighborhood $\mathcal{N}_i(t)$.

Lemma 1 *Suppose Assumption 1 holds. Let $\xi > 0$ and $\gamma \in (0.5, 1)$ be the parameters defined in Algorithm 1. Let j be an arbitrary legitimate neighbor of a legitimate agent i, i.e., $j \in \mathcal{N}_i^{\mathcal{L}}$ for agent $i \in \mathcal{L}$. Then, the misclassification probability of agent j by agent i has the following upper bound,*

$$\mathbb{P}(j \notin \mathcal{N}_i(t)) \leqslant |\mathcal{N}_i^{\mathcal{L}}| \exp(-\xi^2(t+1)^{2\gamma}/2(t+1)) + |\mathcal{N}_i^{\mathcal{M}}| \exp(-(\xi(t+1)^{\gamma} + \lambda(t+1))^2/2(t+1)),$$

$\lambda := \min_{m\in\mathcal{M}}(d - c_m)$.

Proof: By Algorithm 1, a legitimate neighbor j is mislassified when the event $\{\beta_{i\bar{j}(t)}(t) - \beta_{ij}(t) > \xi_t\}$ occurs. First, notice that this event is equivalent to the union of the events $\bigcup_{n\in\mathcal{N}_i}\{\beta_{in}(t) - \beta_{ij}(t) > \xi_t\}$. We check this equivalence in both directions. The event $\beta_{i\bar{j}(t)}(t) > \beta_{ij}(t) + \xi_t$ implies that there exist an element $j' \in \mathcal{N}_i$ such that $\beta_{ij'}(t) > \beta_{ij}(t) + \xi_t$, as we can simply choose $j' = \bar{j}(t)$. The converse is also true. The existence of a $j' \in \mathcal{N}_i$ with $\beta_{ij'}(t) > \beta_{ij}(t) + \xi_t$ implies that $\beta_{i\bar{j}(t)}(t) > \beta_{ij}(t) + \xi_t$ since $\beta_{i\bar{j}(t)}(t) \geqslant \beta_{ij'}(t)$ by definition given in Algorithm 1 Step 2. Therefore, we have that the event $\{\beta_{i\bar{j}(t)}(t) - \beta_{ij}(t) > \xi_t\}$ is equivalent to the union of events $\bigcup_{n\in\mathcal{N}_i}\{\beta_{in}(t) - \beta_{ij}(t) > \xi_t\}$. Using this equality, we have,

$$\begin{aligned}
\mathbb{P}(j \notin \mathcal{N}_i(t)) =& \mathbb{P}(\beta_{i\bar{j}(t)}(t) - \beta_{ij}(t) > \xi_t) && (12)\\
=& \mathbb{P}\Big(\bigcup_{n\in\mathcal{N}_i}\{\beta_{in}(t) - \beta_{ij}(t) > \xi_t\}\Big) && (13)\\
\leqslant& \sum_{l\in\mathcal{N}_i^{\mathcal{L}}\setminus\{j\}} \mathbb{P}(\beta_{il}(t) - \beta_{ij}(t) > \xi_t) && (14)\\
&+ \sum_{m\in\mathcal{N}_i^{\mathcal{M}}} \mathbb{P}(\beta_{im}(t) - \beta_{ij}(t) > \xi_t), && (15)
\end{aligned}$$

where the last step follows from the union bound. First, we focus on bounding the probability $\mathbb{P}(\beta_{il}(t) - \beta_{ij}(t) \geqslant \xi_t)$. Notice that $\beta_{il}(t) - \beta_{ij}(t) = \sum_{s=0}^{t}(\alpha_{il}(s) - \alpha_{ij}(s))$ is the sum of independent random variables $(\alpha_{il}(s) - \alpha_{ij}(s))$ with expectation $\mathbb{E}(\alpha_{il}(s) - \alpha_{ij}(s)) = 0$. Therefore, we directly apply the Chernoff-Hoeffding inequality to obtain

$$\begin{aligned}
\mathbb{P}(\beta_{il}(t) - \beta_{ij}(t) > \xi_t) &\leqslant \exp(-\xi_t^2/2(t+1))\\
&= \exp(-\xi^2(t+1)^{2\gamma}/2(t+1)),
\end{aligned}$$

where in the last step we used the definition of ξ_t. Using a similar line of reasoning, we bound the probability $\mathbb{P}(\beta_{im}(t) - \beta_{ij}(t) > \xi_t)$, as follows:

$$\begin{aligned}
&\mathbb{P}(\beta_{im}(t) - \beta_{ij}(t) > \xi_t)\\
&\overset{(a)}{=} \mathbb{P}(\beta_{im}(t) - \beta_{ij}(t) - \mathbb{E}(\beta_{im}(t) - \beta_{ij}(t))\\
&\qquad > \xi(t+1)^{\gamma} + (t+1)(d - c_m))\\
&\overset{(b)}{\leqslant} \exp(-(\xi(t+1)^{\gamma} + (t+1)(d - c_m))^2/2(t+1)),
\end{aligned}$$

where in (a) we embed the expected difference of trust values into both sides, and in (b) we apply the Chernoff-Hoeffding inequality since $(d - c_m) > 0$ by Assumption 1-*(i)*. The rest of the proof follows from combining the bounds with Eq. (15). ∎

Next, we analyze the probability of misclassifying a malicious neighbor $m \in \mathcal{N}_i$. Such misclassification happens if the gap between the maximum aggregate trust value $\beta_{i\bar{j}(t)}(t)$ and m's value $\beta_{im}(t)$ is at most ξ_t, i.e., $\beta_{i\bar{j}(t)}(t) - \beta_{im}(t) \leqslant \xi_t$.

Lemma 2 *Suppose Assumption 1 holds. Let $\xi > 0$ and $\gamma \in (0.5, 1)$ be the parameters defined in Algorithm 1. Let m be an arbitrary malicious neighbor of a legitimate agent i, i.e., $m \in \mathcal{N}_i^{\mathcal{M}}$ for agent $i \in \mathcal{L}$. Then, for all $t > \left(\frac{\xi}{\lambda}\right)^{1/(1-\gamma)} - 1$, the misclassification probability of agent m by agent i has the following upper bound,*

$$\mathbb{P}(m \in \mathcal{N}_i(t)) \leqslant \exp(-(\xi(t+1)^{\gamma} - (t+1)\lambda)^2/2(t+1)).$$

Proof: By the definition of the trusted neighborhood in Algorithm 1, a malicious neighbor m is misclassified when we have $\beta_{i\bar{j}(t)}(t) - \beta_{im}(t) \leqslant \xi_t$. We note that this event is equivalent to the intersection of events $\bigcap_{j\in\mathcal{N}_i}\{\beta_{ij}(t) - \beta_{im}(t) \leqslant \xi_t\}$. This equivalence holds as we validate it from both directions. The event that the maximum element $\beta_{i\bar{j}(t)}(t) \leqslant \beta_{im}(t) + \xi_t$, which holds by the definition of $\mathcal{N}_i(t)$ (from Alg. 1) and $\beta_{i\bar{j}(t)}(t)$, implies that $\beta_{ij}(t) \leqslant \beta_{im}(t) + \xi_t$ for all $j \in \mathcal{N}_i$ (see Fig. 1.b). Conversely, if $\beta_{ij}(t) \leqslant \beta_{im}(t) + \xi_t$ for all $j \in \mathcal{N}_i$, then we have $\beta_{i\bar{j}(t)}(t) \leqslant \beta_{ij}(t) + \xi_t$ since $\bar{j}(t)$ is also chosen from the set of all $j \in \mathcal{N}_i$ (see Step 2 in Algorithm 1). Thus we have that the event $\{\beta_{i\bar{j}(t)}(t) - \beta_{im}(t) \leqslant \xi_t\}$ is equivalent to the intersection of events $\bigcap_{j\in\mathcal{N}_i}\{\beta_{ij}(t) - \beta_{im}(t) \leqslant \xi_t\}$. Using this equality, we get

$$\begin{aligned}
\mathbb{P}(m \in \mathcal{N}_i(t)) &= \mathbb{P}(\beta_{i\bar{j}(t)}(t) - \beta_{im}(t) \leqslant \xi_t)\\
&= \mathbb{P}\Big(\bigcap_{n\in\mathcal{N}_i}\{\beta_{in}(t) - \beta_{im}(t) \leqslant \xi_t\}\Big)\\
&\leqslant \min_{n\in\mathcal{N}_i} \mathbb{P}(\beta_{in}(t) - \beta_{im}(t) \leqslant \xi_t).
\end{aligned}$$

Consider an arbitrary legitimate neighbor of an agent i, $l \in \mathcal{N}_i^{\mathcal{L}}$. We know that such a neighbor must exist due to Assumption 1-*(iii)*. Then, we have

$$\begin{aligned}
&\min_{n\in\mathcal{N}_i} \mathbb{P}(\beta_{in}(t) - \beta_{im}(t) \leqslant \xi_t) \leqslant \mathbb{P}(\beta_{il}(t) - \beta_{im}(t) \leqslant \xi_t)\\
&= \mathbb{P}(\beta_{il}(t) - \beta_{im}(t) - \mathbb{E}(\beta_{il}(t) - \beta_{im}(t))\\
&\qquad \leqslant \xi(t+1)^{\gamma} - (t+1)(d - c_m)).
\end{aligned}$$

Then, for $t > \left(\frac{\xi}{d-c_m}\right)^{1/(1-\gamma)} - 1$, we have $\xi(t+1)^{\gamma} - (t+1)(d - c_m) < 0$. Therefore, we apply the Chernoff-Hoeffding inequality to obtain the desired result. ∎

Lemmas 1 and 2 show the misclassification probabilities go to 0, as $t \to \infty$, due to the Chernoff-Hoeffding bound. The next lemma states almost sure convergence of weights matrices.

Lemma 3 *Suppose Assumption 1 holds. Let $\xi > 0$ and $\gamma \in (0.5, 1)$ be the parameters defined in Algorithm 1.There exists a (random) finite time $T_f > 0$ such that $W_{\mathcal{L}}(t) = \bar{W}_{\mathcal{L}}$ for all $t \geqslant T_f$. Furthermore, it holds almost surely*

$$\prod_{t=T_0-1}^{\infty} W_{\mathcal{L}}(t) = \mathbf{1}\nu^T\Bigg(\prod_{t=T_0-1}^{\max\{T_f,T_0\}-1} W_{\mathcal{L}}(t)\Bigg), \qquad (16)$$

where the matrix product $\prod_{t=T_0-1}^{\infty} W_{\mathcal{L}}(t) > \mathbf{0}$ for any $T_0 \geqslant 0$ almost surely, and $\nu > \mathbf{0}$ is a stochastic vector.

Proof: Using Lemmas 1 and 2, legitimate agents have geometrically decaying misclassification probabilities. The infinite sums of misclassification probabilities satisfy $\sum_{t=0}^{\infty}\mathbb{P}(j \notin \mathcal{N}_i(t)) = \sum_{t=0}^{\infty} O(\exp(-\xi^2(t+1)^{2\gamma}/2(t+1)) < \infty$ for legitimate neighbors $j \in \mathcal{N}_i^{\mathcal{L}}$, and $\sum_{t=0}^{\infty}\mathbb{P}(j \in \mathcal{N}_i(t)) = \sum_{t=0}^{T'-1}\mathbb{P}(j \in \mathcal{N}_i(t)) + \sum_{t=T'}^{\infty} O(\exp(-(\xi(t +$

$1)^\gamma - (t+1)\lambda)^2/2(t+1))) < \infty$ for all $t \geqslant T'$, where $T' \geqslant \left(\frac{\xi}{d-c_m}\right)^{1/(1-\gamma)} - 1$. Hence, there exists a finite time T_f such that we have $W_{\mathcal{L}}(t) = \overline{W}_{\mathcal{L}}$ for all $t \geqslant T_f$.

As $W_{\mathcal{L}}(t) = \overline{W}_{\mathcal{L}}$ for all $t \geqslant T_f$, for the product of the matrices $W_{\mathcal{L}}(t)$, we have,

$$\begin{aligned}\prod_{t=T_0-1}^{\infty} W_{\mathcal{L}}(t) &= \prod_{t=\max\{T_0,T_f\}}^{\infty} W_{\mathcal{L}}(t) \prod_{t=T_0-1}^{\max\{T_0,T_f\}-1} W_{\mathcal{L}}(t) \\ &= \prod_{t=\max\{T_0,T_f\}}^{\infty} \overline{W}_{\mathcal{L}} \prod_{t=T_0-1}^{\max\{T_0,T_f\}-1} W_{\mathcal{L}}(t) \\ &= \lim_{t\to\infty} \overline{W}_{\mathcal{L}}^{t-\max\{T_0,T_f\}} \prod_{t=T_0-1}^{\max\{T_0,T_f\}-1} W_{\mathcal{L}}(t).\end{aligned}$$

By Assumption 1, the subgraph induced by the legitimate agents is connected and, by the definition of $\overline{W}_{\mathcal{L}}$, it follows that $\overline{W}_{\mathcal{L}}$ implies is a primitive stochastic matrix. Therefore,by the Perron-Frobenius Theorem, we have that $\lim_{t\to\infty} \overline{W}_{\mathcal{L}}^{t-\max\{T_0,T_f\}} = \mathbf{1}\nu^T$, with $\nu > \mathbf{0}$, and

$$\prod_{t=T_0-1}^{\infty} W_{\mathcal{L}}(t) = \mathbf{1}\nu^T \left(\prod_{t=T_0-1}^{\max\{T_f,T_0\}-1} W_{\mathcal{L}}(t) \right).$$

In addition, note that as ν is a stochastic and that diagonal entries of $W_{\mathcal{L}}(t)$ are positive. Thus, $\prod_{t=T_0-1}^{\infty} W_{\mathcal{L}}(t) > \mathbf{0}$ almost surely and $\nu^T \left(\prod_{t=T_0-1}^{\max\{T_f,T_0\}-1} W_{\mathcal{L}}(t) \right) > \mathbf{0}$. ■

The following two lemmas are the direct results of Lemma 3 and the proofs are along the lines of Propositions 2-3 in [17].

Lemma 4 *Suppose Assumption 1 holds. In Algorithm 1, let $\xi > 0$ and $\gamma \in (0.5, 1)$. Let $x_{\mathcal{L}}(0)$ be the initial values of legitimate agents. Then, $\tilde{x}_{\mathcal{L}}(T_0, t)$ converges almost surely, i.e., almost surely*

$$\lim_{t\to\infty} \tilde{x}_{\mathcal{L}}(T_0, t) = \left(\prod_{k=T_0-1}^{\infty} W_{\mathcal{L}}(k) \right) x_{\mathcal{L}}(0) = y\mathbf{1},$$

where $y \in \mathbb{R}$ is a random variable depending on T_f and T_0.

Lemma 5 *Suppose Assumption 1 holds. In Algorithm 1, let $\xi > 0$ and $\gamma \in (0.5, 1)$. Then, the influence $\phi_{\mathcal{M}}(T_0, t)$ from malicious agents converges almost surely, i.e., we have almost surely*

$$\begin{aligned}\lim_{t\to\infty} \phi_{\mathcal{M}}(T_0, t) &= \sum_{k=T_0-1}^{\infty} \left(\prod_{\ell=k+1}^{\infty} W_{\mathcal{L}}(\ell) \right) W_{\mathcal{M}}(k) x_{\mathcal{M}}(k) \\ &= f\mathbf{1},\end{aligned}$$

where $f \in \mathbb{R}$ is a random variable depending on T_f and T_0.

We now state that legitimate agents reach a common value asymptotically.

Corollary 1 *Suppose Assumption 1 holds, and let $\xi > 0$ and $\gamma \in (0.5, 1)$ in Algorithm 1. Then, the consensus protocol* (3) *among the legitimate agents converges almost surely, i.e.,*

$$\lim_{t\to\infty} x_{\mathcal{L}}(T_0, t) = z\mathbf{1} \quad \textit{almost surely}, \tag{17}$$

where $z \in \mathbb{R}$ is a random variable given by $z = y + f$, with y and f from Lemma 4 and Lemma 5, respectively.

Proof: The result follows by the relation $x_{\mathcal{L}}(T_0, t) = \tilde{x}_{\mathcal{L}}(T_0, t) + \phi_{\mathcal{M}}(T_0, t)$ (see (4)) and Lemmas 4–5. ■

Corollary 1 states that the legitimate agents reach the same random scalar value z almost surely. However, the consensus value z can be outside the convex hull of the initial values $x_{\mathcal{L}}(0)$ of legitimate agents, unlike the result of the standard consensus process.

We conclude this section with the final theorem on the deviation from nominal consensus.

Theorem 1 *Suppose Assumption 1 holds. Let $\xi > 0$ and $\gamma \in (0.5, 1)$ be as given in Algorithm 1. For an error level $\delta > 0$, $T_0 > \left(\frac{\tilde{\xi}}{\lambda}\right)^{1/(1-\gamma)} - 1$ and any $\tilde{\xi} > \xi$, we have,*

$$\begin{aligned}\mathbb{P}(\max \limsup_{t\to\infty} |[x_{\mathcal{L}}(T_0, t) - \mathbf{1}\nu^T x_{\mathcal{L}}(0)]_i| < \Delta_{\max}(T_0, \delta)) \\ \geqslant 1 - \delta,\end{aligned}$$

where $\Delta_{\max}(T_0, \delta) = 2(\frac{2\eta}{\delta} g_{\mathcal{L}}(T_0) + \frac{\eta}{\kappa\delta} g_{\mathcal{M}}(T_0))$.

We examined the deviation from the nominal consensus process in the extended version [24]. The result shows that as agents wait longer to start the consensus process, *i.e.* with increasing T_0, they have tighter bounds on the probabilities.

IV. Numerical Studies

In this section, we assess the performance of our proposed algorithm against different type of malicious attacks in numerical studies. We consider a challenging scenario with 10 legitimate and 15 malicious agents where the malicious agents constitute the majority. We construct the communication graph as follows: first, we generate a cycle graph among the legitimate agents and then add 10 more random edges between them. The malicious agents form random connections to every other agent with probability 0.2 while we also ensure that they are connected to at least one legitimate agent. We sample agents' initial values from the uniform distribution $\mathcal{U}[-4, 4]$ once for both legitimate and malicious agents. Legitimate agents follow the consensus dynamics given in Eq. (2) with $\kappa = 10$. The legitimate neighbors' trust values are sampled from the uniform distribution $\mathcal{U}[0.3, 1]$ resulting in the expected value $d = 0.65$. To model the case where each malicious agent has a different expected value, we choose their expectations from the uniform distribution $\mathcal{U}[0, 0.45]$.

We consider two types of attack in our experiments: 1) Consistent attacks where malicious agents always send η (or $-\eta$) if the true consensus value is negative (positive). During each communication, a malicious agent m's trust value is sampled from the uniform distribution $\mathcal{U}[2c_m - 1, 1]$ with probability p_m and from $\mathcal{U}[0.3, 1]$ (the same distribution as the legitimate neighbors) with probability $1 - p_m$. 2) Intermittent failures where malicious nodes follow the same consensus update rule as the legitimate agents but send η (or $-\eta$) to their neighbors with probability p_m. During failures, malicious agents' trust values are sampled from the uniform distribution $\mathcal{U}[2c_m - 1, 1]$ and from $\mathcal{U}[0.3, 1]$ otherwise.

For both attacks, we assume that all malicious agents have the same p_m, and consider two cases with $p_m = 0.2$ and $p_m = 0.8$. We use $\xi = 0.15$ and $\gamma = 0.7$ as the parameters of our learning algorithm Algorithm 1. We use $T_0 = 60$ as it satisfies the largest theoretical lower bound on T_0 given in Theorem 1 for all cases. We track the maximum deviation from the nominal consensus value over time in Fig. 1. Note that these results are averaged over 100 trials for each setup, where the communication graph, the initial values and the expected values of agents are fixed across the trials. In all cases and trials, we observe that agents reach consensus, as predicted by Corollary 1. Moreover, we see that the probability of being observable, p_m, has the highest impact on the deviation, as it affects the misclassification probabilities (see Lemma 2 and Lemma 1). As expected, consistent attacks have more impact on the system when the attack probability is low ($p_m = 0.2$) since malicious agents are always inserting a constant value η (or $-\eta$) to the system, and they stay undetected for a longer time. When the attack probability is high ($p_m = 0.8$), malicious agents get detected quickly, and the errors mainly stem from misclassified legitimate agents.

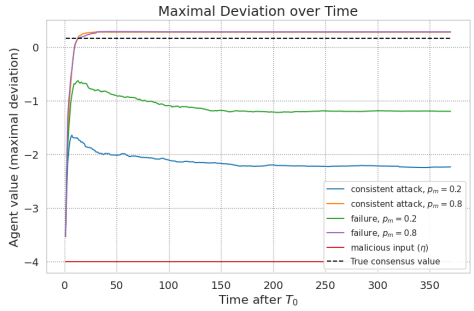

Fig. 1: The maximal deviation from the nominal consensus. Malicious input η is the maximum impact that malicious agents can have on the system.

V. Conclusion

In this paper, we studied the multi-agent resilient consensus problem in undirected and static communication networks. Assuming trust observations are available, we considered the scenarios with intermittent faulty or malicious transmissions. We developed a novel detection algorithm to let legitimate agents determine their neighbors' types correctly. We showed that misclassification probabilities go to zero in finite time and agents reach a consensus almost surely asymptotically. We characterized the maximal deviation in terms of error tolerance, expected trust values, and algorithmic parameters. Numerical experiments showed the convergence of the consensus process and the deviation under different scenarios.

References

[1] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical association*, vol. 69, no. 345, pp. 118–121, 1974.

[2] 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. 49, no. 9, pp. 1520–1533, 2004.

[3] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multiagent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, 2009.

[4] S. S. Kia, B. Van Scoy, J. Cortes, R. A. Freeman, K. M. Lynch, and S. Martinez, "Tutorial on dynamic average consensus: The problem, its applications, and the algorithms," *IEEE Control Systems Magazine*, vol. 39, no. 3, pp. 40–72, 2019.

[5] J. Cortes, S. Martinez, and F. Bullo, "Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions," *IEEE Transactions on Automatic Control*, vol. 51, no. 8, pp. 1289–1298, 2006.

[6] S. Martinez, "Distributed interpolation schemes for field estimation by mobile sensor networks," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 2, pp. 491–500, 2009.

[7] 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, 2011.

[8] 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, 2010.

[9] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, 2013.

[10] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1063–1076, 2019.

[11] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[12] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, "Guaranteeing spoof-resilient multi-robot networks," *Autonomous Robots*, vol. 41, pp. 1383–1400, 2017.

[13] M. Cavorsi, O. E. Akgün, M. Yemini, A. J. Goldsmith, and S. Gil, "Exploiting trust for resilient hypothesis testing with malicious robots," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 7663–7669.

[14] ——, "Exploiting trust for resilient hypothesis testing with malicious robots," *IEEE Transactions on Robotics*, vol. 40, pp. 3514–3536, 2024.

[15] A. Pierson and M. Schwager, *Adaptive Inter-Robot Trust for Robust Multi-Robot Sensor Coverage*. Cham: Springer International Publishing, 2016, pp. 167–183. [Online]. Available: https://doi.org/10.1007/978-3-319-28872-7_10

[16] 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 & Networking*, ser. MobiCom '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 441–452. [Online]. Available: https://doi.org/10.1145/2500423.2500444

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

[18] E. Nurellari, D. McLernon, and M. Ghogho, "A secure optimum distributed detection scheme in under-attack wireless sensor networks," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, no. 2, pp. 325–337, 2018.

[19] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Asymptotic analysis of distributed bayesian detection with byzantine data," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 608–612, 2015.

[20] O. E. Akgun, A. K. Dayi, S. Gil, and A. Nedich, "Learning trust over directed graphs in multiagent systems," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 142–154.

[21] S. M. Kay, *Fundamentals of Statistical Signal Processing Volume II Detection Theory*. New Jersey: Prentice Hall PTR, 1998.

[22] V. W. Berger and Y. Zhou, "Kolmogorov–smirnov test: Overview," *Wiley statsref: Statistics reference online*, 2014.

[23] M. Yemini, A. Nedić, S. Gil, and A. J. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 4185–4192.

[24] S. Aydin, O. E. Akgun, S. Gil, and A. Nedić, "Multi-agent resilient consensus under intermittent faulty and malicious transmissions (extended version)," *arXiv preprint*, 2024.