Strategic Attacks on Finite-Time Consensus

Orhan Eren Akgün, Thomas Kaminsky, Áron Vékássy, Angelia Nedić, Stephanie Gil

***Abstract*— This work addresses the linear consensus problem in multi-agent systems under adversarial attacks. We examine scenarios where legitimate agents utilize stochastic inter-agent trust observations to assess the likelihood of neighboring agents acting maliciously in order to mitigate their impact. Malicious agents, in turn, aim to strategically choose what values to send to maximize the disagreement among legitimate agents in a finite horizon. In contrast to prior studies that assume static adversarial behavior and focus on asymptotic consensus, this work investigates the impact of strategic attacks within a finite number of iterations. Specifically, we characterize the maximum disagreement that malicious agents can induce in finite time and the computational complexity of computing the best attack strategy. We compare the effectiveness of different attack strategies through numerical experiments.**

I. Introduction

In this work, we study the finite-time performance of linear consensus in multi-agent systems under adversarial attacks. We consider agents performing linear average consensus over undirected communication graphs during a fixed and known time horizon H. In our model, an adversarial attack occurs at some unknown time $\tau < H$, compromising an unknown subset of agents. These compromised agents, which we call "malicious," behave non-cooperatively from time τ until the end of the horizon, with the goal of maximizing disagreement between the agents in the rest of the network. Assuming the availability of stochastic observations that quantify the likelihood of neighboring agents acting maliciously, we formally define a finite-time consensus problem in the presence of such strategic malicious agents. We also highlight the analytical challenges posed by this problem and present initial steps toward addressing them.

While finite-time consensus [1]–[3] and consensus with malicious agents [4], [5] are individually well-studied, little is known about their combination. Classical methods relying on filtering extreme transmitted values are capable of mitigating malicious influence in finite time, but they also impose restrictive constraints on the tolerable number of malicious agents [6], [7]. More recent approaches are able to slacken these restrictions by leveraging sensor measurements available to embodied distributed systems. This enables agents to estimate agent trustworthiness [8], [9] rather than relying on information sent along the graph. However, work in this setting thus far has focused exclusively on asymptotic results rather than finite-time scenarios and assumes malicious agents always behave maliciously rather than strategically timing their attacks [10]–[12].

Like previous work, our approach augments linear average consensus with a concurrent detection algorithm utilizing trust observations. However, unlike previous works [11]–[13], our finite-time framework allows malicious agents to behave strategically—remaining hidden until launching their attack at time τ, making almost sure detection impossible. To address this challenge, in Section III-A, we introduce a sliding observation window that limits the set of considered trust observations, and characterize detection probabilities based on the window's size.

In Section III-B, building upon this detection capability, we analyze how malicious agents influence disagreement and provide bounds on their impact. We also characterize analytical challenges unique to the finite-time setting. We show that stepwise disagreement reduction cannot be guaranteed, even in the absence of malicious agents. We also prove that sufficiently strong malicious agents capable of predicting when they will get detected can calculate an optimal attack strategy in polynomial time.

Finally, in Section IV we provide simulations modeling malicious attacks, finding that the optimal attack is well-approximated by a simple heuristic for two graph topologies.

II. Problem Setup

We consider the case where a set of agents $\mathcal{V}$ communicate over a static undirected communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We denote the neighbors of agent i by $\mathcal{N}_i$. Let the set of malicious agents be $\mathcal{M}$ with $N_{\mathcal{M}} \triangleq |\mathcal{M}|$. The remaining agents, which always act cooperatively, are referred to as 'legitimate' agents and are denoted by $\mathcal{L}$ with $N_{\mathcal{L}} \triangleq |\mathcal{L}|$. These sets remain fixed over time, and each agent is either malicious or legitimate; i.e., $\mathcal{L} \cap \mathcal{M} = \emptyset$ and $\mathcal{L} \cup \mathcal{M} = \mathcal{V}$. Legitimate agents do not know the attack time τ.

Each agent $i \in \mathcal{V}$ has an initial value $x_i(0) \in \mathbb{R}$. Without loss of generality, we write the well-known average consensus dynamics $x(t+1) = W(t)x(t)$ with the assumption that the first $N_{\mathcal{L}}$ indices correspond to the legitimate agents and the remaining indices correspond to the malicious agents. Let $x_{\mathcal{L}}(t)$ and $x_{\mathcal{M}}(t)$ represent the values of the legitimate and malicious agents at time t, respectively. Then, for all $t \in \{0, 1, \ldots, \tau - 1\}$, the consensus process of all agents can be decomposed into block matrix form as:

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

O. E. Akgün, T. Kaminsky, Á. Vékássy, and S. Gil are with the School of Engineering and Applied Sciences, Harvard University, Allston, MA 02134. E-mail: erenakgun@g.harvard.edu, tkaminsky@g.harvard.edu, avekassy@g.harvard.edu, sgil@seas.harvard.edu. A. Nedić is 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, CNS-2147694, and AFOSR award FA9550-22-1-0223

where $W_{ij}(t)$ is the weight assigned by agent i to agent j at time t, satisfying $W_{ij}(t) \geq 0$, $W_{ii}(t) > 0$, $W_{ii}(t) + \sum_{j\in\mathcal{N}_i} W_{ij}(t) = 1$ for all $t \in \{0, 1, \ldots, H-1\}$, and $\mathcal{N}_i$ denotes the neighbors of agent i. As we elaborate later on, the matrices $W(t)$ are possibly time-varying and random. Once the attack begins, malicious agents may send arbitrary values. We assume that at all time steps, agent values remain within a bounded interval, i.e., $x_i(t) \in [-\eta, \eta]$ for all $i \in \mathcal{V}$ and all $t \in \{0, 1, \ldots, H\}$. Therefore, malicious agents send values $x_{\mathcal{M}}(t) \in [-\eta, \eta]^{N_{\mathcal{M}}}$ for $t \geq \tau$. [1] We consider a broadcast communication model where agents send the same value to all their neighbors. The malicious agents' goal is to maximize the expected disagreement among legitimate agents at the horizon, given the attack time τ. The 'disagreement'[2] at time t is:

$$\delta(t) = x_{\mathcal{L}}(t)^{\intercal} L x_{\mathcal{L}}(t), \tag{2}$$

where $L \triangleq I - \frac{1}{N_{\mathcal{L}}}\mathbf{1}\mathbf{1}^{\intercal}$ and $\mathbf{1}$ denotes a column vector of ones of the appropriate dimension. Then, the goal of malicious agents is defined as

$$\max_{x_{\mathcal{M}}(t)\in[-\eta,\eta]^{N_{\mathcal{M}}},\ \forall t\in\{\tau,\ldots,H-1\}} \mathbb{E}[\delta(H)], \tag{3}$$

where the randomness in expectation is over the weight matrices $W(t)$. Next, we define how malicious agents can have influence on the disagreement $\delta(H)$.

Let $W(s;t)$ denote the backward product of matrices—that is, for $t \geq s$, $W(s;t) \triangleq W(t)W(t-1)\cdots W(s)$, and $W(s;t) = I$ for $t < s$. Then the consensus process of legitimate agents after the start of the attack can be written as: for all $t = \tau, \ldots, H-1$,

$$\begin{aligned} x_{\mathcal{L}}(t+1) &= W_{\mathcal{L}\mathcal{L}}(\tau;t)x_{\mathcal{L}}(\tau) \\ &\quad + \sum_{k=\tau}^{t-1} W_{\mathcal{L}\mathcal{L}}(k+1;t-1)W_{\mathcal{L}\mathcal{M}}(k)x_{\mathcal{M}}(k). \end{aligned}$$

From the perspective of the malicious agents who wish to maximize the expected disagreement defined in (3), this is a constrained control problem. To see this, define the controllability matrix $B(\tau,t)$ and control input $u(\tau,t)$ as

$$\begin{aligned} B(\tau,t) \triangleq [&W_{\mathcal{L}\mathcal{L}}(\tau+1;t-1)W_{\mathcal{L}\mathcal{M}}(\tau) \mid \\ &W_{\mathcal{L}\mathcal{L}}(\tau+2;t-1)W_{\mathcal{L}\mathcal{M}}(\tau+1) \mid \cdots \mid W_{\mathcal{L}\mathcal{M}}(t-1)], \end{aligned} \tag{4}$$

$$u(\tau,t) \triangleq \left[x_{\mathcal{M}}(\tau)^T \mid x_{\mathcal{M}}(\tau+1)^T \mid \cdots \mid x_{\mathcal{M}}(t)^T\right]^T, \tag{5}$$

where $B(\tau,t) \in \mathbb{R}^{N_{\mathcal{L}}\times N_{\mathcal{M}}(t-\tau)}$ and $u(\tau,t) \in \mathbb{R}^{N_{\mathcal{M}}(t-\tau)}$ is the malicious agent's input. As in [5], malicious agents can control the system through $u(\tau,t)$. Then our update appears to be a constrained linear controller:

$$x_{\mathcal{L}}(t+1) = W_{\mathcal{L}\mathcal{L}}(\tau;t)x_{\mathcal{L}}(\tau) + B(\tau,t)u(\tau,t). \tag{6}$$

[1]It suffices to assume that the initial values of legitimate agents are in the interval $[-\eta, \eta]$ and all agents are aware of this. Since legitimate agents always perform linear averaging, their values remain within this range. Consequently, malicious agents cannot send values outside this range either; otherwise, they would be immediately detected.

[2]Note that disagreement we define corresponds to 'polarization' in opinion dynamics literature [14].

Next, we discuss how agents choose their weights $W_{ij}(t)$. In this work, we are interested in settings where agents may receive stochastic observations of trust from other agents that send them information. Following previous works [8], [11], [15], [16] which formalize this notion, we give the following definition and assumptions on stochastic trust observations:

Definition 1 (Stochastic Observation of Trust α_{ij}) *If agent $j \in \mathcal{V}$ sends information to a legitimate agent $i \in \mathcal{L}$ at time t (meaning $j \in \mathcal{N}_i$), then agent i receives a stochastic observation of trust $\alpha_{ij}(t) \in [0,1]$. Larger $\alpha_{ij}(t)$ values indicate higher levels of trust.*

Assumption 1 *Assume that the following statements hold:*
1) For all agents $i \in \mathcal{V}$ and their legitimate neighbors $l \in \mathcal{N}_i \cap \mathcal{L}$, at any time $t \in \{0, 1, \ldots, H-1\}$, trust observations $\alpha_{il}(t)$ have a known expected value $E_{\mathcal{L}} \triangleq \mathbf{E}[\alpha_{il}(t)]$.
2) Prior to the attack time τ, malicious agents always behave cooperatively, i.e., they follow the update rule in Equation (1). Consequently, for agent $i \in \mathcal{V}$ with malicious neighbor $m \in \mathcal{N}_i$, for any $t < \tau$, the trust observations $\alpha_{im}(t)$ satisfy $\mathbf{E}[\alpha_{im}(t)|t < \tau] = E_{\mathcal{L}}$, like legitimate agents. After the attack time, the expected trust observation changes to $\mathbf{E}[\alpha_{im}(t)|t \geq \tau] = E_{\mathcal{M}}$ where $E_{\mathcal{M}}$ satisfies $E_{\mathcal{L}} > E_{\mathcal{M}}$.
3) Given a fixed attack time τ, trust observations $\alpha_{ij}(t)$ are independent over time for all agents $i \in \mathcal{V}$ and their neighbors $j \in \mathcal{N}_i$.

Assumption 1.3 implies that the trust observations are independent of the values $x_i(t)$ exchanged between the agents, which is consistent with previous work that characterizes and utilizes stochastic trust observations [8], [11], [16]. Now, we explain how legitimate agents use trust observations to decide which neighbors to trust. We will consider a simple detection rule where legitimate agents estimate the expected trust value of a neighbor based on a window of n_o-many past observations. For ease of exposition, we will assume that agents will start with n_o-many observations at time 0. Then agent i estimates neighbor j's expected trust value as:

$$\hat{\alpha}_{ij}(t) = \frac{1}{n_o}\sum_{k=t-n_o}^{t-1} \alpha_{ij}(t). \tag{7}$$

Next, we define the trusted neighborhood of a legitimate agent i at time t:

$$\hat{\mathcal{N}}_i(t) \triangleq \{j \in \mathcal{N}_i \mid E_{\mathcal{L}} - \hat{\alpha}_{ij}(t) \leq \xi\}, \tag{8}$$

where $\xi > 0$ is a fixed threshold parameter that is chosen by legitimate agents. Our later results derive ranges for ξ to allow for detection of malicious agents. Agents in the trusted neighborhood are assigned positive weights, while all others receive zero weights. Thus, $W_{ij}(t) > 0$ if and only if $j \in \hat{\mathcal{N}}_i(t)$. Note that, before the attack time τ, malicious agents behave indistinguishably from legitimate agents and therefore adhere to the same trusted neighborhood detection rule. However, after τ, every malicious agent $m \in \mathcal{M}$ can broadcast its values to all of its neighbors in $\mathcal{N}_m$.

III. Analysis

In this section, we analyze the performance of the detection algorithm and characterize the impact of malicious agents on

disagreement, $\delta(H)$, at time H. Specifically, we first show upper bounds on the misclassification probabilities resulting from the detection rule in (8). Then, we characterize bounds on the disagreement as well as the computational complexity of finding the optimal control input $u(\tau, t)$ that maximizes the disagreement.

A. Detection Performance

We are interested in the probabilities of misclassifying a legitimate agent as malicious and vice versa after an attack begins. This is in contrast to previous work [11], [15], [16] which assumes malicious agents are adversarial from initial time $t = 0$. Instead, we assume malicious agents act legitimately until time τ, after which they behave maliciously. At τ, past trust observations become misleading, as malicious agents would have previously appeared legitimate. Thus, unlike [11], [15], [16], we restrict our analysis to a fixed number of past observations, denoted as n_o in (7). We observe that increasing n_o reduces the probability of misclassifying malicious agents (Proposition 1), but also leads to a higher initial misclassification probability at τ (Proposition 2).

Proposition 1 *Let i be a legitimate agent and $l \in \mathcal{N}_i \cap \mathcal{L}$ be its legitimate neighbor. Then, for any $\xi > 0$, we have*

$$\mathbb{P}(l \notin \hat{\mathcal{N}}_i(t)) \leq \exp(-2\xi^2 n_o).$$

Moreover, for any time $t < \tau$, for a malicious neighbor $m \in \mathcal{N}_i \cap \mathcal{M}$, we have $\mathbb{P}(m \notin \hat{\mathcal{N}}_i(t)) = \mathbb{P}(l \notin \hat{\mathcal{N}}_i(t))$.

Proof. We have $\mathbb{P}(l \notin \hat{\mathcal{N}}_i(t)) = \mathbb{P}(E_{\mathcal{L}} - \hat{\alpha}_{ij}(t) > \xi)$ by the detection rule (8). As $\alpha_{ij}(t) \in [0, 1]$, and are independent over time for a fixed τ, the results follow from the Chernoff-Hoeffding bound [17, Theorem 1.1, p. 6]. □

Proposition 2 *Let i be a legitimate agent and $m \in \mathcal{N}_i \cap \mathcal{M}$ be its malicious neighbor. Define c_l as the number of observations before the attack time τ included in the observation window: $c_l \triangleq \max\{\tau - t + n_o, 0\}$. Let $0 < \xi < \frac{n_o - c_l}{n_o}(E_{\mathcal{L}} - E_{\mathcal{M}})$ Then, we have*

$$\mathbb{P}(m \in \hat{\mathcal{N}}_i(t) \mid t \geq \tau) \leq \exp\left(-\frac{2}{n_o}\left((n_o - c_l)(E_{\mathcal{L}} - E_{\mathcal{M}}) - n_o\xi\right)^2\right).$$

Proof. Using the detection rule in Equation (8), we get $\mathbb{P}(m \in \hat{\mathcal{N}}_i(t) \mid t \geq \tau) = \mathbb{P}(\hat{\alpha}_{im}(t) \geq E_{\mathcal{L}} - \xi \mid t \geq \tau)$. We have $\mathbb{E}[\hat{\alpha}_{im}(t)] = \frac{c_l E_{\mathcal{L}} + (n_o - c_l)E_{\mathcal{M}}}{n_o}$. Using this expectation and multiplying everything by n_o we get

$$\mathbb{P}(\hat{\alpha}_{im}(t) \geq E_{\mathcal{L}} - \xi \mid t \geq \tau) = \mathbb{P}(n_o\hat{\alpha}_{im}(t) \geq n_o\mathbb{E}[\hat{\alpha}_{im}(t)] + (n_o - c_l)(E_{\mathcal{L}} - E_{\mathcal{M}}) - n_o\xi \mid t \geq \tau).$$

Since trust observations are independent over time given τ, the result follows by the Chernoff-Hoeffding bound [17, Theorem 1.1, p. 6] for any $0 < \xi < \frac{n_o - c_l}{n_o}(E_{\mathcal{L}} - E_{\mathcal{M}})$. □

These results indicate that misclassification probabilities are low for a sufficiently large observation window n_o. The neighborhood detection rule, in combination with stochastic trust observations, results in a sequence of time-varying stochastic weight matrices $\{W(t)\}_{t=0}^{H}$. We highlight several key observations regarding the properties of these matrices. First, since trust values received from different neighbors $j, l \in \mathcal{N}_i$ are not necessarily independent, the inclusion of edges may be correlated. This contrasts with commonly studied random graph models, such as Erdős–Rényi graphs, where edges are typically assumed to be independent. Second, the weight matrices are temporally correlated, since agents determine their neighborhoods based on n_o past trust observations. This dependence introduces a structured evolution in the weight matrices over time. Lastly, the matrix sequence $\{W(t), t = 0, \ldots, H-1\}$ is independent of the values $u(t; \tau)$ sent by the malicious agents to disturb the process, as the values sent by the agents are decoupled from the trust values given the attack time τ.

B. Bounding Disagreement

In this section, we turn our attention to establishing possible bounds on disagreement for a given sequence of matrices generated by the trusted-neighborhood detection process. Throughout this section, we will assume that the stochastic weight matrix sequence $\{W(t)\}_{t=0}^{H}$ is given. Our goal is to analyze the ways malicious agents can cause disagreement and characterize their impact. First, we study the system's behavior in the absence of malicious agents. It is well-known that a consensus protocol on a connected graph with row-stochastic weights leads to agreement. However, in the finite-time setting, short-term disagreement reduction depends on how agents assign their weights $W_{\mathcal{LL}}(t)$. Our next result shows that common averaging rules, where agents consider only their in-neighbors to design row-stochastic weight matrices, do not always guarantee a step-wise decrease in disagreement. However, if the weight matrix is column-stochastic, then disagreement is non-increasing.

Proposition 3 *At time step t, assume that no malicious agent is included in any legitimate neighborhood, i.e., $m \notin \hat{\mathcal{N}}_i(t)$ for all $i \in \mathcal{L}$ and $m \in \mathcal{N}_i \cap \mathcal{M}$. Then there exists a row stochastic matrix $W_{\mathcal{LL}}(t)$ corresponding to a strongly connected graph such that $\delta(t+1) > \delta(t)$ for some $x_{\mathcal{L}}(t)$. However, if $W_{\mathcal{LL}}(t)$ is doubly stochastic, then we have $\delta(t+1) \leq \delta(t)$.*

Proof. We show the first claim by constructing an example. Consider a line graph with 4 nodes, row stochastic weight matrix $W_{\mathcal{LL}}(t) = \begin{pmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.8 & 0.1 & 0.1 & 0 \\ 0 & 0.1 & 0.1 & 0.8 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$, and the vector $x_{\mathcal{L}}(t)^\intercal = \begin{pmatrix} 1 & 0 & 0 & -1 \end{pmatrix}$. In this example the disagreement increases after one consensus step.

For the second, by definition of disagreement given in (2), we have $\delta(t) = x_{\mathcal{L}}(t)^\intercal L x_{\mathcal{L}}(t)$. Since $L = I - \frac{1}{N_{\mathcal{L}}}\mathbf{1}\mathbf{1}^\intercal$, $\delta(t) = \|x_{\mathcal{L}}(t)\|^2 - (\mathbf{1}^\intercal x_{\mathcal{L}}(t))^2$, where $\|\cdot\|$ is the 2-norm. Similarly, we have $\delta(t+1) = \|W_{\mathcal{LL}}(t)x_{\mathcal{L}}(t)\|^2 - (\mathbf{1}^\intercal W_{\mathcal{LL}}(t)x_{\mathcal{L}}(t))^2$. When $W_{\mathcal{LL}}(t)$ is doubly stochastic, we have $\mathbf{1}^\intercal W_{\mathcal{LL}}(t)x_{\mathcal{L}}(t) = \mathbf{1}^\intercal x_{\mathcal{L}}(t)$. Therefore, it is sufficient to show that $\|W_{\mathcal{LL}}(t)x_{\mathcal{L}}(t)\| \leq \|x_{\mathcal{L}}(t)\|$. This is true as the largest singular value of a doubly stochastic matrix is 1. □

We next investigate the impact of malicious agents on disagreement. For this, we define δ^* to be the maximum disagreement that the malicious agents can achieve at time

H, i.e., $\delta^* = \max_{u(\tau,t)\in[-\eta,\eta]^{N_\mathcal{M}(H-\tau)}} \delta(H)$. This value is the optimal value of a quadratic function $\delta(H)$ over the box constraint $u(\tau,t) \in [-\eta,\eta]^{N_\mathcal{M}(H-\tau)}$, which is an NP-hard problem in general [18]. However, in the next theorem, we show that due to the rank deficiency of $B(\tau,t)$, its complexity is polynomial for fixed $N_\mathcal{L}$.

Theorem 1 *We denote the total number of times that a malicious agent is connected to at least one legitimate neighbor at any time after τ as $S \triangleq \sum_{t=\tau}^{H-1}\sum_{m\in\mathcal{M}} \mathbb{1}_{\{\sum_{l\in\mathcal{L}}[W_{\mathcal{LM}}(t)]_{lm}>0\}}$, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function. Then, given the sequence of weight matrices $\{W(t)\}_{k=\tau}^{H}$ and initial values $x_\mathcal{L}(\tau)$, the problem*

$$\max_{u(\tau,t)\in[-\eta,\eta]^{N_\mathcal{M}(H-\tau)}} \delta(H)$$

can be solved in $\mathcal{O}(S^{N_\mathcal{L}}\left(SN_L + N_L^2\right))$ time.

Proof. Denote the feasible set of $x_\mathcal{L}(H)$ by $\mathcal{Z}$. We have $x_\mathcal{L}(H) = W_{\mathcal{LL}}(\tau;H-1)x_\mathcal{L}(\tau) + B(H,\tau)u(H,\tau)$, where $B(H,\tau)$ and $u(H,\tau)$ are the controllability matrix and control input defined in (6). Malicious agents influence the system through $B(H,\tau)$, which has S nonzero columns. Thus, varying $u(H,\tau)$, we see $\mathcal{Z}$ is the affine image of a hyper-cube in $\mathbb{R}^S$, called a zonotope [19]. Since zonotopes are convex and $\delta(H)$ is a convex quadratic, we can compute the optimum by enumerating the vertices of $\mathcal{Z}$, achievable in $\mathcal{O}(S^{N_\mathcal{L}})$ [18, Theorem 3.2], and maximizing disagreement over this set. At each step we compute the corresponding candidate $u(H,\tau)$, adding S columns of $B(H,\tau)$ and evaluating $\delta(H)$, requiring $\mathcal{O}(SN_L + N_L^2)$. □

Note that S is always upper-bounded by $N_\mathcal{M}(H-\tau)$, and, despite being polynomial in S, the complexity still grows exponentially with $N_\mathcal{L}$. In our next result, we provide a bound on the maximum achievable disagreement δ^*.

Theorem 2 *Let σ be the largest singular value of the matrices in the sequence $\{W_{\mathcal{LL}}(t)\}_{t=\tau}^{H}$. Define the maximum total weight a legitimate agent assigns its malicious neighbors as $\Upsilon = \max_{i\in\mathcal{L},t\geq\tau}\sum_j [W_{\mathcal{LM}}(t)]_{ij}$. Then, we have $0 \leq \delta^* \leq \eta^2 N_\mathcal{L}$. Moreover, if $\sigma < 1$, we also have*

$$\delta^* \leq \eta^2 N_\mathcal{L} \min\left\{1, \left(\sigma^{H-\tau+1} + \Upsilon\frac{1-\sigma^{H-\tau-1}}{1-\sigma}\right)^2\right\}.$$

Proof. Let $\mathcal{Z}$ denote the feasible space of legitimate agents' final values $x_\mathcal{L}(H)$ as in the proof of Theorem 1. By the definition of δ^*, we have $\delta^* = \max_{x_\mathcal{L}(H)\in\mathcal{Z}} x_\mathcal{L}(H)^\intercal L x_\mathcal{L}(H)$. Since L is positive semi-definite, disagreement is always non-negative and so $0 \leq \delta^*$. For the upper bound, we consider an l_2 ball that encompasses the feasible space $\mathcal{Z}$. Our goal is to find a radius R such that $\|x_\mathcal{L}(H)\| \leq R$ and compute $\max_{\|x_\mathcal{L}(H)\|\leq R} x_\mathcal{L}(H)^\intercal L x_\mathcal{L}(H)$. The maximum value is achieved at $\lambda(L)R^2$, where $\lambda(L)$ is the largest eigenvalue of L, which is equal to 1. We will find both a general-case R, and one requiring additional assumptions. First, since legitimate agents' values are guaranteed to stay in range $[-\eta,\eta]$, we have $x_\mathcal{L}(t) \in [-\eta,\eta]^{N_\mathcal{L}}$ and so $\|x_\mathcal{L}(t)\| \leq \sqrt{N_\mathcal{L}}\eta$. Therefore, we can choose $R = \sqrt{N_\mathcal{L}}\eta$. Second, from the linear controller formulation of $x_\mathcal{L}(t)$ given in (6), we get $\|x_\mathcal{L}(H)\| = \|W_{\mathcal{LL}}(\tau;H)x_\mathcal{L}(\tau) + B(H,\tau)u(H,\tau)\|$. Assuming $\sigma < 1$ and using the triangle inequality and the definitions of $B(H,\tau)$ and $u(H,\tau)$ in (4), we obtain

$$\begin{aligned}\|x_\mathcal{L}(H)\| &\leq \|W_{\mathcal{LL}}(\tau;H)x_\mathcal{L}(\tau)\| \\ &\quad + \sum_{k=\tau}^{H-2}\|W_{\mathcal{LL}}(k+1;H-2)W_{\mathcal{LM}}(k)x_\mathcal{M}(k)\| \\ &\leq \eta\sqrt{N_\mathcal{L}}\sigma^{H-\tau+1} + \sum_{k=\tau}^{H-2}\sigma^{H-2-k}\|W_{\mathcal{LM}}(k)x_\mathcal{M}(k)\| \\ &\leq \eta\sqrt{N_\mathcal{L}}\sigma^{H-\tau+1} + \eta\Upsilon\sqrt{N_\mathcal{L}}\frac{1-\sigma^{H-\tau-1}}{1-\sigma}.\end{aligned}$$

Therefore, we can set $R = \sigma^{H-\tau} + \Upsilon\frac{1-\sigma^{H-\tau-1}}{1-\sigma}$. Taking the minimum over each value for R yields the result. □

This bound is loose, as it only considers the maximum norm of $x_\mathcal{L}(H)$ instead of the reachable space. To see this, assume that we have $H \gg \tau$ and malicious agents stay undetected after the attack time. If they always send η and the legitimate agents are sufficiently connected, we have $x_\mathcal{L}(t) \approx \eta\mathbf{1}$. In this case, the disagreement is almost 0, while $\|x_\mathcal{L}(t)\| \approx \eta\sqrt{N_\mathcal{L}}$. In numerical studies, we examine the emperical maximum reachable disagreement on two graphs.

IV. Numerical Studies

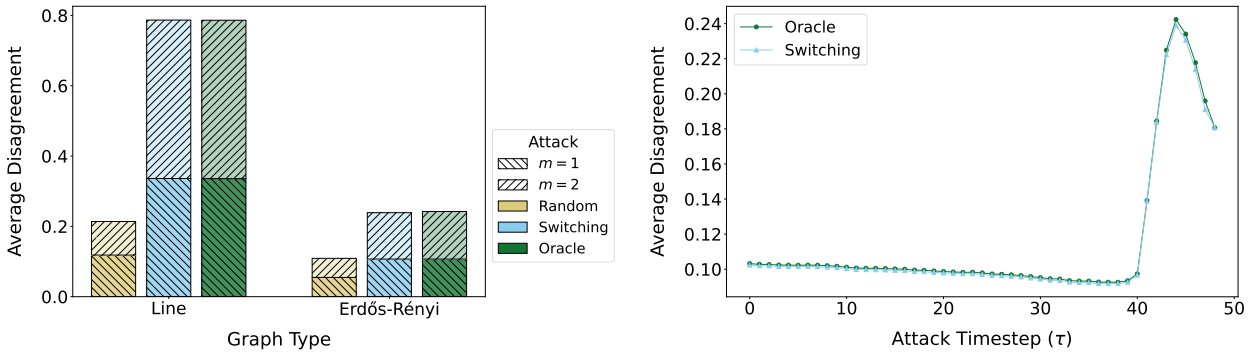
To validate our theoretical results, we conduct numerical experiments. Below we describe our setup in detail.

1) **Evaluation:** For all experiments, we use time horizon $H = 50$, window size $n_o = 10$, and randomly initialize each vector uniformly in $[-\eta,\eta]$. For each, we report the average disagreement. Without loss of generality, we let $\eta = 1$. We conduct 100 trials for each set of parameters, and likewise fix 100 random seeds—thus, the average disagreement for each attack is over the same sequence of weights $\{W(t)\}_{t=0}^{H}$.

2) **Graph Topologies:** We evaluate our model on two topologies: line and Erdős–Rényi (ER) graphs. Due to computational constraints, we consider graphs consisting of 6 legitimate agents, with $N_\mathcal{M} \in \{1,2\}$ malicious agents added either at the ends of the line or as nodes in the ER graph before edge generation. We add edges to the ER-graph with probability $p = 0.8$, and constrain the graph to be connected. We use the same line and ER graph across all trials.

3) **Trust Observations:** Following previous work [11], [15], we model trust observations as uniform distributions with distinguishable means. In particular, we define the legitimate and malicious distributions $\mathcal{D}_\mathcal{L} \triangleq \text{Unif}(0.375, 0.75)$ and $\mathcal{D}_\mathcal{M} \triangleq \text{Unif}(0.25, 0.625)$. We sample trust values following the specification given by Assumption 1, and use the detection rule from (8) with $\xi = 0.08$ to decide on edges. Weight matrices $\{W(t)\}_{t=0}^{H}$ are generated by normalizing each row to be uniform over in-neighbors.

4) **Malicious Strategies:** We consider three attack strategies: the `random` strategy has malicious agents take random actions in $\{-\eta,\eta\}$ at each timestep. The `switching` strategy uses the sequence $\{W(t)\}_{t=0}^{H}$ to find the disagreement-maximizing action constrained to switching exactly once between sending η or $-\eta$—for example, $[\eta,\eta,\eta,-\eta,-\eta,-\eta,-\eta]$. Finally, the `oracle` implements



(a) Disagreement induced by the best attack time τ under each attack.

(b) Average disagreement achieved by `oracle` and `switching` strategies on the Erdős–Rényi graph with 6 malicious agents, versus attack time.

Fig. 1. (a) For all attacks and graphs, the best attack times ranged from $\tau = 44$ to $\tau = 48$, and the `switching` and `oracle` strategies found identical optimal attack times and corresponding vectors. (b) The `switching` strategy heuristic performs similarly to the `oracle` across all attack times τ.

the poly-time algorithm in [18] to calculate the optimal malicious input for each $\{W(t)\}_{t=0}^{H}$. We find the best attack for every time $\tau \in \{0, ..., H-1\}$.

We find that the empirical average disagreement displayed in Figure 1(a) is far below the Theorem 2 bounds in all cases, with a maximum of 0.7864 incurred by the `oracle` and `switching` attacks with two malicious agents. Interestingly, we find the optimal `oracle` attack coincides exactly with the best attack found by `switching`. In general, the strategies only disagree for small τ, at which points the `oracle` can attain marginal improvements by sending extreme values whenever it is in a trusted neighborhood, shown in Figure 1(b). We note that the `switching` and `oracle` strategies are very strong adversaries, as they require knowledge of future trust values to compute their attacks. Thus, it is unlikely that stochastic matrices could be known *ex ante* in this way.

V. Conclusion

This paper studies finite-time consensus under adversarial attacks, where malicious agents aim to maximize disagreement and legitimate agents detect adversaries via stochastic trust observations. We propose a sliding-window trust-based detection algorithm and analyze its performance. We characterize the disagreement achievable by malicious agents, provide an algorithm to compute an optimal attack strategy in polynomial-time, and show numerically that a heuristic closely matches the optimal attack in concrete settings.

References

[1] S. Sundaram and C. N. Hadjicostis, "Finite-time distributed consensus in graphs with time-invariant topologies," in *2007 American Control Conference*, 2007, pp. 711–716.

[2] A. Sandryhaila, S. Kar, and J. M. F. Moura, "Finite-time distributed consensus through graph filters," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1080–1084.

[3] J. M. Hendrickx, G. Shi, and K. H. Johansson, "Finite-time consensus using stochastic matrices with positive diagonals," *IEEE Transactions on Automatic Control*, vol. 60, pp. 1070–1073, 2013.

[4] C. N. Hadjicostis and A. D. Domínguez-García, "Trustworthy distributed average consensus based on locally assessed trust evaluations," *IEEE Transactions on Automatic Control*, 2024.

[5] 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.

[6] 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.

[7] J. Usevitch and D. Panagou, "Resilient finite-time consensus: A discontinuous systems perspective," in *2020 American Control Conference (ACC)*, 2020, pp. 3285–3290.

[8] S. Gil, S. Kumar, M. Mazumder, D. Katabi, and D. Rus, "Guaranteeing spoof-resilient multi-robot networks," *Autonomous Robots*, vol. 41, no. 6, pp. 1383–1400, 2017.

[9] S. Gil, M. Yemini, A. Chorti, A. Nedić, H. V. Poor, and A. J. Goldsmith, "How physicality enables trust: A new era of trust-centered cyberphysical systems," 2023. [Online]. Available: https://arxiv.org/abs/2311.07492

[10] 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, 2019.

[11] M. Yemini, A. Nedić, A. J. Goldsmith, and S. Gil, "Characterizing trust and resilience in distributed consensus for cyberphysical systems," *IEEE Trans. Robot.*, vol. 38, no. 1, pp. 71–91, 2022.

[12] L. Ballotta and M. Yemini, "The role of confidence for trust-based resilient consensus," in *2024 American Control Conference (ACC)*. IEEE, 2024, pp. 2822–2829.

[13] S. Aydın, O. E. Akgün, S. Gil, and A. Nedić, "Multi-agent resilient consensus under intermittent faulty and malicious transmissions," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 6057–6062.

[14] J. Gaitonde, J. Kleinberg, and É. Tardos, "Adversarial perturbations of opinion dynamics in networks," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020, pp. 471–472.

[15] O. E. Akgün, A. K. Dayı, S. Gil, and A. Nedić, "Learning trust over directed graphs in multiagent systems," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, vol. 211. PMLR, 15–16 Jun 2023, pp. 142–154.

[16] M. Yemini, A. Nedić, A. J. Goldsmith, and S. Gil, "Resilient distributed optimization for multi-agent cyberphysical systems," *IEEE Transactions on Automatic Control*, 2025.

[17] D. P. Dubhashi and A. Panconesi, *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[18] J.-A. Ferrez, K. Fukuda, and T. M. Liebling, "Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm," *European Journal of Operational Research*, vol. 166, no. 1, pp. 35–50, 2005.

[19] G. M. Ziegler, *Lectures on polytopes*. Springer Science & Business Media, 2012, vol. 152.