The Role of Confidence for Trust-Based Resilient Consensus

Luca Ballotta[1], Stephanie Gil[2], and Michal Yemini[3]

***Abstract*— In this paper, we consider a multi-agent system where agents aim to achieve a consensus in spite of interactions with malicious agents that communicate misleading information. Physical channels supporting communication in cyberphysical systems offer attractive opportunities to detect malicious agents: however, trustworthiness indications coming from the channel are subject to uncertainty and need to be treated with this in mind. We propose a resilient consensus protocol that incorporates trust observations from the channel and weighs them with a parameter that accounts for how confident an agent is regarding its understanding of the legitimacy of other agents in the network, with no need for the initial observation window T_0 that has been utilized in previous works. Analytical and numerical results show that *(i)* our protocol achieves a resilient consensus in the presence of malicious agents and *(ii)* the steady-state deviation from nominal consensus can be minimized by a suitable choice of the confidence parameter that depends on the statistics of trust observations.**

I. Introduction

Consensus in multi-agent systems is an essential tool that arises in many applications, ranging from distributed control to multi-robot coordination. However, the classical consensus protocol is known to be fragile to outliers, and it easily fails in the presence of agents that do not behave according to the protocol — for example in the adversarial case.

To tame such *malicious agents* and recover a *resilient consensus* among the others, several strategies have been proposed in the literature. One common method to achieve this goal is the Weighted-Mean Subsequence Reduced (W-MSR) algorithm [1], which implements a filtering of the messages received by each agent based on a supposed number of malicious agents and has been adapted to many application domains [2], [3]. Other strategies have been recently proposed that use different rules to filter out suspicious data, such as the similarity between two agents' states [4],or implement different weighing schemes, for example anchoring the agents to their initial condition [5], or leverage enhanced network structure, such as secured agents [6].

Recovering a resilient consensus purely based on the data exchanged among agents is in general a challenging task. A notable limitation on the theoretical guarantees of W-MSR is that the communication graph needs to enjoy a connectivity property, called r-robustness, that ensures a pervasive information flow among agents. Unfortunately, a high enough r-robustness may require dense networks, and it cannot be verified in polynomial time w.r.t. the number of agents [7], [8]. Thus, in real applications and especially in large networks, W-MSR may not lead to a consensus. In general, purely data-driven mechanisms are inherently subject to the limitation that the source of information used to determine which agents are well-behaving (the exchanged data) is also the very quantity the agents try to optimize, so that it is not easy to infer whether an agent is sending wrong data or is just distant from the target value.

In contrast to data-centered approaches, recent works [9]–[14] have proposed to use *physical* information of transmissions to boost resilience in distributed cyberphysical systems, leveraging the fact that this source of information is independent from the exchanged data. Cyberphysical systems are widely adopted in applications, from robot teams to smart grids. In such systems, communication occurs over physical channels that can be used to extract information used to assess the validity of a transmission: for instance, wireless signals can be analyzed to detect manipulated messages [15], [16]. providing useful information for security [15], [16].

However, while using physical transmission channels as a source of information for legitimacy of received messages allows one to decouple the consensus task from the detection of potential adversaries within the network, this information is usually uncertain [15], partially hindering its usefulness if this is not properly accounted for. In particular, while an agent typically gains confidence in the classification of its neighbors as trustworthy or not as information is accrued, individual transmissions may not be reliably used for such a classification. This calls for attention in embedding the physical trustworthiness indications into the design of a resilient consensus protocol.

In this paper, webuild on two recently proposed approaches to resilient consensus and design a protocol that uses the information embedded in the physical channel without compromising the execution of the consensus task through erroneous classification of agents. We draw inspiration from the trust-based protocol in [10] and the competition-based approach in [5], and propose a novel algorithm that integrates the notion of *trust*, coming from the physical channel, and the concept of *confidence*, which counterbalances the uncertainty in agent classification and can be seen as prior knowledge on the quality of gained trust information. This integration allows us to circumvent two limitations of the previous algorithms: on the one hand, we do not need a time window $T_0 > 0$ of trust observations as in [10]; on the other hand, the agents achieve an asymptotic consensus, differently from the data-driven context in [5]. Specifically, the proposed protocol anchors

We gratefully acknowledge partial support through MIUR PRIN project 2017NS9FEY and through ONR grant N00014-21-1-2714.

[1]Department of Information Engineering, University of Padova, 35131 Padova, Italy ballotta@dei.unipd.it.

[2]Department of Computer Science, Harvard University, Boston, MA 02138 sgil@seas.harvard.edu.

[3]Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002 Israel michal.yemini@biu.ac.il.

the agents to their initial condition through a time-varying weight λ_t that reflects how confident an agent is about the trustworthiness of its neighbors: owing to the competition-based approach, this strategy avoids the agents to be misled through misclassification of neighbors and enhances resilience in the face of both unknown malicious agents and uncertain information from the physical channel. Moreover, we show that the confidence parameter can be tuned to optimize performance: analytical and numerical results indicate that λ_t should decay according to the average time the agents need to correctly classify their neighbors.

The rest of this paper is organized as follows. Section II-A presents the system model and the problem formulation, while Section II-B introduces the proposed resilient consensus protocol and mathematical models for trust and confidence. Then, Section III provides theoretical guarantees offered by the protocol, focusing on convergence (Section III-A) and asymptotic deviation from the nominal consensus (Section III-B). Finally, Section IV presents numerical simulation results that corroborate the analysis and prove our protocol effective.

II. Setup

A. System Model and Problem Formulation

Network. We consider a multi-agent system composed of N agents equipped with scalar-valued states: we denote the state of agent i at time t by $x_t^i \in \mathbb{R}$, with $i \in \mathcal{V} \doteq \{1, \dots, N\}$, and the vector with all stacked states by $x_t \in \mathbb{R}^N$. The agents can communicate and exchange their states through a fixed communication network, modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each element $e = (i, j) \in \mathcal{E}$ indicates communication edge between agents i and j: if $(i, j) \in \mathcal{E}$, it means that agent j can transmit data to agent i through a direct link.

In the network, L agents truthfully follow a designated protocol (*legitimate agents* $\mathcal{L} \subset \mathcal{V}$) while $M = N - L$ agents behave arbitrarily (*malicious agents* $\mathcal{M} \subset \mathcal{V}$), potentially disrupting the task executed by legitimate agents. We set the labels of legitimate and malicious agents as $\mathcal{L} = \{1, \dots, L\}$ and $\mathcal{M} = \{L+1, \dots, N\}$ and denote their collective states respectively by $x_t^{\mathcal{L}} \in \mathbb{R}^L$ and $x_t^{\mathcal{M}} \in \mathbb{R}^M$. We denote by $d_{\max}$ the maximal (in-)degree of legitimate agents, with $d_{\max} < N$.

Consensus Task. The legitimate agents aim to achieve a consensus. The nominal consensus value is determined by their initial states $x_0^{\mathcal{L}}$ and by the ideal communication network without malicious agents. Specifically, let $\mathcal{N}_i \in \mathcal{V}$ denote the neighbors of agent i in the communication network $\mathcal{G}$, *i.e.,* $\mathcal{N}_i \doteq \{j \in \mathcal{V} : (i, j) \in \mathcal{E}\}$, and consider the nominal matrix $\overline{W}^{\mathcal{L}} \in \mathbb{R}^{L \times L}$ with weights defined as follows for $i, j \in \mathcal{L}$:

$$\left[\overline{W}^{\mathcal{L}}\right]_{ij} = \begin{cases} \frac{1}{|\mathcal{N}_i \cap \mathcal{L}| + 1} & \text{if } j \in \mathcal{N}_i \cap \mathcal{L}, \\ 0 & \text{if } j \notin \mathcal{N}_i \cup \{i\}, \\ 1 - \sum_{i \in \mathcal{N}_j} \left[\overline{W}^{\mathcal{L}}\right]_{ij} & \text{if } j = i. \end{cases} \quad (1)$$

Ideally, the legitimate agents should disregard messages sent by malicious agents (*i.e.,* set their weights to zero) and run the following nominal consensus protocol starting from $x_0^{\mathcal{L}}$:

$$x_{t+1}^{\mathcal{L}} = \overline{W}^{\mathcal{L}} x_t^{\mathcal{L}}, \qquad t \geq 0. \qquad \text{(NOM)}$$

Unfortunately, the identity of malicious agents is unknown to legitimate agents, so that these cannot implement the weights (1) and the protocol (NOM). In the next section, we propose a *resilient consensus* protocol aimed at recovering the final outcome of (NOM) in the face of malicious agents.

B. Resilient Consensus Protocol

In this work, we propose the following resilient protocol to be implemented by each legitimate agent $i \in \mathcal{L}$ for $t \geq 0$:

$$x_{t+1}^i = \lambda_t x_0^i + (1 - \lambda_t) \sum_{j \in \mathcal{N}_i \cup \{i\}} w_{ij}(t) x_t^j. \qquad \text{(RES)}$$

Rule (RES) uses two key ingredients. The weights $w_{ij}(t) \in [0, 1]$ are computed online based on *trust information* that agent i collects about its neighbor j overtime. The time-varying parameter $\lambda_t \in [0, 1]$ accounts for how *confident* agent i feels about the trustworthiness of its neighbors. In the following, we describe these two features in detail. We note that the parameter λ_t is new w.r.t. to previous work [10] and a major objective in this work is to analytically characterize the impact of this "confidence" term on mitigating the effect of malicious agents when consensus protocol (RES) starts from time 0 (*i.e.,* no observation window as in [10] is present).

Trust. We are interested in the case where each transmission from agent j to agent i can be tagged with an observation $\alpha_{ij}(t) \in [0, 1]$ of a random variable α_{ij}.

Definition 1 (Trust variable α_{ij})**.** For every $i \in \mathcal{L}$ and $j \in \mathcal{N}_i$, the random variable α_{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 denote the expected value of α_{ij} by $E_{\mathcal{L}} \doteq \mathbb{E}[\alpha_{ij}] - 1/2$ for legitimate transmissions and by $E_{\mathcal{M}} \doteq \mathbb{E}[\alpha_{ij}] - 1/2$ for malicious ones. We assume the availability of observations $\alpha_{ij}(t)$ of α_{ij} through $t \geq 0$.

We refer to [17] for a concrete example of such an α_{ij} variable. Intuitively, a random realization $\alpha_{ij}(t)$ contains useful trust information if the legitimacy of the transmission can be thresholded. We assume that a value $\alpha_{ij}(t) > 1/2$ indicates a legitimate transmission and $\alpha_{ij}(t) < 1/2$ a malicious transmission in a stochastic sense (miscommunications are possible). The value $\alpha_{ij}(t) = 1/2$ means that the observation is completely ambiguous and contains no useful trust information for the transmission at time t.

Weights. The weights $w_{ij}(t)$ in (RES) are chosen according to the history of trust scores $\alpha_{ij}(t)$. By defining the aggregate trust of communications from agent j to agent i as

$$\beta_{ij}(t) = \sum_{s=0}^{t} \left(\alpha_{ij}(s) - \frac{1}{2}\right), \quad i \in \mathcal{L}, j \in \mathcal{N}_i, \qquad (2)$$

we define the *trusted neighborhood* of agent i at time t as

$$\mathcal{N}_i(t) \doteq \{j \in \mathcal{N}_i : \beta_{ij}(t) \geq 0\}. \qquad (3)$$

Then, the weights in (RES) are assigned online as follows:

$$w_{ij}(t) = \begin{cases} \frac{1}{|\mathcal{N}_i(t)|+1} & \text{if } j \in \mathcal{N}_i(t), \\ 0 & \text{if } j \notin \mathcal{N}_i(t) \cup \{i\}, \\ 1 - \sum_{i \in \mathcal{N}_j} w_{ij}(t) & \text{if } j = i. \end{cases} \quad (4)$$

The weighing rule above attempts to recover the nominal weights (1) as time proceeds. In particular, the trusted neighborhood $\mathcal{N}_i(t)$ is designed to reconstruct the set $\mathcal{N}_i \cap \mathcal{L}$ leveraging trust information collected by agent i overtime.

Confidence. Because trust observations $\alpha_{ij}(t)$ may misclassify transmissions, the weights computed as per (4) may not immediately recover the true weights: in fact, even assuming that a sufficient number of transmissions can give a clear indication about the trustworthiness of a neighbor, a legitimate agent needs to act cautiously as long as it is unsure about the trust information collected in order to not be misled by erroneous classifications. To this aim, we modify the standard consensus rule by adding the parameter λ_t in (RES) that anchors the legitimate agents to their initial condition and refrains them from fully relying on the neighbors' states.

Intuitively, agent i accrues knowledge about the trustworthiness of its neighbors as more trust-tagged transmissions have been received. This intuition can in fact be formalized by upper bounding the probability of misclassifying a neighbor.

Assumption 1 (Trust observations are informative)**.** Legitimate (malicious) transmissions are classified as legitimate (malicious) on average. Formally, $E_{\mathcal{L}} > 0$ and $E_{\mathcal{M}} < 0$.

Lemma 1 (Decaying misclassification probability [10])**.**

$$\begin{aligned} \mathbb{P}\left[\beta_{ij}(t) < 0\right] &\leq \mathrm{e}^{-2E_{\mathcal{L}}^2(t+1)} \quad \forall i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{L} \\ \mathbb{P}\left[\beta_{ij}(t) \geq 0\right] &\leq \mathrm{e}^{-2E_{\mathcal{M}}^2(t+1)} \quad \forall i \in \mathcal{L}, j \in \mathcal{N}_i \cap \mathcal{M}. \end{aligned} \quad (5)$$

Lemma 1 implies that, under Assumption 1 that trust values $\alpha_{ij}(t)$ are informative, the legitimate agents infer which neighbors are trustworthy with higher confidence overtime. On the other hand, the early iterations of the protocol have higher chance of misclassifications. To counterbalance this fact and make updates resilient, we design the parameter λ_t as *decreasing* with time. This way, early updates are conservative and not much sensitive to misclassifications ($\lambda_t \lesssim 1$ for small t), while late updates rely almost totally on the neighbors confidently classified as legitimate ($\lambda_t \gtrsim 0$ for large t).

Discussion - Trust and Confidence. The update rule (RES) leverages the two fundamental concepts of *trust* and *confidence*, which are used together in an intertwined manner.

The works [10], [12], [13] show how to utilize physics-based trust observations to help a legitimate agent decide which neighbors it should rely on as it runs the protocol. Nonetheless, at each step, the agent can either trust a neighbor or not and it does not scale the weights given to trusted neighbors relatively by how confident it is on the decision. Furthermore, in the work [10] the deviation from the nominal consensus value is strongly tied to an initial observation window T_0 where the agents do not trust any of their neighbors and only collect trust observations to choose wisely what neighbors to trust in the first data update round. This length T_0 value is not straightforward to choose when the number of overall rounds varies and is not guaranteed in advance. In contrast, this work introduces the parameter λ_t to capture the confidence that an agent has about the legitimacy of its neighbors, propose a softer approach to the clear-cut observation window used in [10] where agents do not trust one another, and explores the role of such a confidence parameter to opportunistically tune the weights assigned to the neighbors. In particular, the formulation (RES) highlights that the agent tunes the weights given to trusted neighbors scaling them by $(1 - \lambda_t)$.

The use of λ_t draws inspiration from previous work [5], [18] where the Friedkin-Johnsen model [19] is used to achieve resilient average consensus, intended as the minimization of the mean square deviation. Contrarily to the trust-based works mentioned above, the latter references do not use information derived from physical transmissions but study a robust update rule within a data-based context. The updates in [5], [18] use a constant parameter λ (interpreted as *competition* among agents) that mitigates the influence of malicious agents by forcefully anchoring the legitimate agents to the initial condition, ruling out the possibility of getting arbitrarily close to the nominal consensus. In this work, we use a source of information independent of the data (because it derives from physical transmissions) to make the competition-based rule more flexible and able to recover a consensus.

III. Performance Analysis

Let $W_t \in \mathbb{R}^{L \times N}$ denote the matrix with weights (4), *i.e.*, $[W_t]_{ij} = w_{ij}(t)$, and consider the following partition:

$$W_t = \left[\begin{array}{c|c} W_t^{\mathcal{L}} & W_t^{\mathcal{M}} \end{array}\right], \quad W_t^{\mathcal{L}} \in \mathbb{R}^{L \times L}. \quad (6)$$

The protocol (RES) can be rewritten as follows:

$$\begin{aligned} x_{t+1}^{\mathcal{L}} &= \lambda_t x_0^{\mathcal{L}} + (1-\lambda_t)\begin{bmatrix} W_t^{\mathcal{L}} & W_t^{\mathcal{M}} \end{bmatrix}\begin{bmatrix} x_t^{\mathcal{L}} \\ x_t^{\mathcal{M}} \end{bmatrix} \\ &= \bar{x}_t^{\mathcal{L}} + \bar{x}_t^{\mathcal{M}} \end{aligned} \quad (7)$$

where we define the state contributions due to legitimate and malicious agents respectively as

$$\bar{x}_t^{\mathcal{L}} \doteq \prod_{k=0}^{t}(1-\lambda_k)W_k^{\mathcal{L}} x_0^{\mathcal{L}} + \sum_{k=0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k x_0^{\mathcal{L}} \quad (8a)$$

$$\bar{x}_t^{\mathcal{M}} \doteq \sum_{k=0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)(1-\lambda_k)W_k^{\mathcal{M}} x_k^{\mathcal{M}}. \quad (8b)$$

In the following, we assume the parameter λ_t has expression

$$\lambda_t = c\mathrm{e}^{-\gamma t}, \qquad 0 < c < 1, \ \gamma > 0. \quad (9)$$

According to the discussion below Lemma 1, we impose that λ_t decays overtime to enable resilient updates at the beginning. The expression (9) is chosen mainly to make the analysis tractable: specifically, we will be mostly concerned with how the coefficient γ, which determines how fast λ_t decays to zero, affects the steady-state deviation. Nonetheless, we argue that the insights suggested by our analysis apply also to other

choices of λ_t, which will be numerically explored in a future extension. Moreover, we note that also the misclassification probabilities (5) decay exponentially, suggesting that the choice (9) could be a good match with the trust statistics.

A. Convergence to Consensus

Lemma 1 implies that there exists a.s. a finite time $T_{\rm f} \geq 0$ such that the estimated legitimate weights $W_t^{\mathcal{L}}$ equal the true weights $\overline{W}^{\mathcal{L}}$ for all $t \geq T_{\rm f}$ [10, Proposition 1]. Moreover, under the mild assumption that the subgraph induced by the legitimate agents is connected, the following fact holds.

Lemma 2 ([10, Lemma 1]). *The matrix $\overline{W}^{\mathcal{L}}$ is primitive and there exists a stochastic vector v such that $\left(\overline{W}^{\mathcal{L}}\right)^{\infty} = \mathbb{1}v^\top$.*

Let $a \vee b \doteq \max\{a, b\}$. For every finite $k_0 \geq 0$, we have

$$\prod_{k=k_0}^{\infty} (1-\lambda_k) W_k^{\mathcal{L}} = \mathbb{1}v^\top \pi_{k_0} \Pi_{k_0} \tag{10}$$

where v is the Perron eigenvector of $\overline{W}^{\mathcal{L}}$ (see Lemma 2) and

$$\pi_{k_0} \doteq \prod_{k=k_0 \vee T_{\rm f}}^{\infty} (1-\lambda_k), \quad \Pi_{k_0} \doteq \prod_{k=k_0}^{k_0 \vee (T_{\rm f}-1)} (1-\lambda_k) W_k^{\mathcal{L}}. \tag{11}$$

It can be verified that $\pi_{k_0} > 0$ if and only if $\sum_{k=k_0 \vee T_{\rm f}}^{\infty} \lambda_k$ converges, which is the case under (9) for every $\gamma > 0$.

Contribution by Legitimate Agents. From the definition (8a) and (10), at the limit it holds

$$\begin{aligned} \bar{x}_{\infty}^{\mathcal{L}} &= \prod_{k=0}^{\infty} (1-\lambda_k) W_k^{\mathcal{L}} x_0^{\mathcal{L}} + \sum_{k=0}^{\infty} \left(\prod_{s=k+1}^{\infty} (1-\lambda_s) W_s^{\mathcal{L}} \right) \lambda_k x_0^{\mathcal{L}} \\ &= \mathbb{1}v^\top \underbrace{\left(\Pi_0 x_0^{\mathcal{L}} + \sum_{k=0}^{\infty} \pi_{k+1} \Pi_{k+1} \lambda_k x_0^{\mathcal{L}} \right)}_{\doteq y^{\mathcal{L}}}. \end{aligned} \tag{12}$$

In view of (9), the coordinates of $y^{\mathcal{L}}$ are finite because so are the coordinates of $x_0^{\mathcal{L}}$ and the matrices Π_k are sub-stochastic.

Contribution by Malicious Agents. From the definition (8b) and (10), at the limit it holds

$$\begin{aligned} \bar{x}_{\infty}^{\mathcal{M}} &= \sum_{k=0}^{\infty} \left(\prod_{s=k+1}^{\infty} (1-\lambda_s) W_s^{\mathcal{L}} \right) (1-\lambda_k) W_k^{\mathcal{M}} x_k^{\mathcal{M}} \\ &\overset{\text{a.s.}}{=} \mathbb{1}v^\top \underbrace{\sum_{k=0}^{T_{\rm f}-1} \pi_{k+1} \Pi_{k+1} (1-\lambda_k) W_k^{\mathcal{M}} x_k^{\mathcal{M}}}_{\doteq y^{\mathcal{M}}} \end{aligned} \tag{13}$$

where $y^{\mathcal{M}}$ is the sum of a finite number of vectors a.s.

Combining (7) with (12)–(13), we conclude that the legitimate agents converge to the consensus $x_{\infty}^{\mathcal{L}} = \mathbb{1}v^\top (y^{\mathcal{L}} + y^{\mathcal{M}})$.

Remark 1 (Convergence *vs.* λ_t). Note that $x_t^{\mathcal{L}}$ converges for any convergent sequence of λ_t, not only under the choice (9).

B. Deviation from Nominal Consensus

After assessing that the legitimate agents asymptotically achieve a consensus, we wish to evaluate the steady-state deviation from the nominal consensus value, which is the one induced by the nominal weight matrix $\overline{W}^{\mathcal{L}}$. We quantify the deviation of agent $i \in \mathcal{L}$ at time t as follows:

$$\tilde{x}_t^i \doteq \left| x_t^i - x_{\rm ss}^{\mathcal{L},*} \right| = \left| \left[x_t^{\mathcal{L}} - \mathbb{1} x_{\rm ss}^{\mathcal{L},*} \right]_i \right| \tag{14}$$

where $x_{\rm ss}^{\mathcal{L},*} \doteq v^\top x_0^{\mathcal{L}}$ is the nominal consensus value of legitimate agents at steady state. In particular, we are interested in upper bounding the probability of the event

$$\max_{i \in \mathcal{L}} \limsup_{t \to \infty} \tilde{x}_t^i > \epsilon \tag{15}$$

to a threshold δ, for suitable $\epsilon > 0$ and $\delta > 0$. To this aim, we separately evaluate the state contributions of legitimate and malicious agents, and then combine their respective bounds to ultimately bound the probability of (15). In the derivation, we will use the following assumption without loss of generality.

Assumption 2 (State bound). It holds $\max_{i \in \mathcal{V}, t \geq 0} |x_i^t| \leq \eta$.

By virtue of the consensus reached at steady state by legitimate agents according to Section III-A, all state trajectories have a well-defined limit (*i.e.,* the consensus value) that is equal for all legitimate agents: this allows us to simplify (15) and formally consider the event $\lim_{t \to \infty} \tilde{x}_t^i > \epsilon$ for any $i \in \mathcal{L}$.

Evaluating the deviation from nominal is helpful to achieve analytical intuition that can help to design the parameter λ_t. Intuitively, small values of γ in (9) refrain the legitimate agents from collaborating with trusted neighbors for longer time, which should help when the trust scores $\alpha_{ij}(t)$ are rather uncertain, while large values of γ turn (RES) into the standard consensus protocol after a few iterations, and should suit cases when the true weights are quickly recovered.

Remark 2 (Design of λ_t). While we focus on the impact of γ for the sake of simplicity, all observations similarly hold for the coefficient c but with opposite monotonicity: for example, if the deviation increases with γ, it also decreases with c.

1) Legitimate Agents: In view of the setup in Section II-B, the only correct contribution to the state of any legitimate agent is the information coming from other legitimate agents, which ideally leads to the true consensus value $x_{\rm ss}^{\mathcal{L},*}$. Hence, we define the deviation term due to legitimate agents as

$$\tilde{x}_{t+1}^{i,\mathcal{L}} \doteq \left| \left[\bar{x}_{t+1}^{\mathcal{L}} - \mathbb{1} x_{\rm ss}^{\mathcal{L},*} \right]_i \right| = \left| \left[\widetilde{W}_t^{\mathcal{L}} x_0^{\mathcal{L}} \right]_i \right| \tag{16}$$

where

$$\widetilde{W}_t^{\mathcal{L}} \doteq \prod_{k=0}^{t} (1-\lambda_k) W_k^{\mathcal{L}} + \sum_{k=0}^{t} \left(\prod_{s=k+1}^{t} (1-\lambda_s) W_s^{\mathcal{L}} \right) \lambda_k - \left(\overline{W}^{\mathcal{L}} \right)^{\infty}. \tag{17}$$

We have the following result.

Lemma 3. *The deviation from nominal consensus due to legitimate agents' contribution can be bounded as*

$$\mathbb{P}\left[\lim_{t \to \infty} \tilde{x}_t^{i,\mathcal{L}} > \epsilon \right] < \eta u_{\mathcal{L}}(\epsilon), \qquad \forall i \in \mathcal{L} \tag{18}$$

where we define

$$u_{\mathcal{L}}(\epsilon) \doteq \frac{2}{\epsilon}\left(\mathrm{e}^{s(\gamma)}\left(1-\left(\frac{1}{d_{max}+1}\right)^{\mathbb{E}[T_f]}\right)+1-v_m\mathbb{E}[\ell]\right) \tag{19}$$

with $v_m \doteq \min_{i\in\mathcal{L}} v_i$, $\ell \doteq \min\{\ell_1,\ell_2\}$, *and*

$$s(\gamma) \doteq -\frac{1}{\gamma} - \frac{\ln(1-c\mathrm{e}^{-\gamma})}{\gamma}\cdot\frac{1-c\mathrm{e}^{-\gamma}}{c\mathrm{e}^{-\gamma}} \tag{20}$$

$$\ell_1 \doteq \left(1-c\mathrm{e}^{-\gamma(T_f\vee 1)}\right)^{\frac{1}{1-\mathrm{e}^{-\gamma}}}\left(c\mathrm{e}^{-\gamma((T_f-1)\vee 0)}\right. \tag{21}$$
$$\left.+c\left(\frac{1-c\mathrm{e}^{-\gamma}}{d_{max}+1}\right)^{T_f-1}\frac{1-\mathrm{e}^{-\gamma(T_f-1)}}{1-\mathrm{e}^{-\gamma}}\mathbb{1}_{\{T_f>1\}}\right)$$

$$\ell_2 \doteq 1-\mathrm{e}^{s(\gamma)}. \tag{22}$$

Proof. Let

$$\widetilde{W}_t^{\mathcal{L}} = \widetilde{W}_{t,1}^{\mathcal{L}} + \widetilde{W}_{t,2}^{\mathcal{L}} \tag{23}$$

where

$$\widetilde{W}_{t,1}^{\mathcal{L}} \doteq \prod_{k=0}^{t}(1-\lambda_k)W_k^{\mathcal{L}} - \left(\prod_{k=0}^{t}(1-\lambda_k)\right)\left(\overline{W}^{\mathcal{L}}\right)^{\infty} \tag{24}$$

expresses the mismatch with the nominal (true) weights and

$$\widetilde{W}_{t,2}^{\mathcal{L}} \doteq \sum_{k=0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k - \left(1-\prod_{k=0}^{t}(1-\lambda_k)\right)\left(\overline{W}^{\mathcal{L}}\right)^{\infty} \tag{25}$$

is associated with the input $\lambda_t x_0^{\mathcal{L}}$ that anchors the legitimate agents to their initial condition throughout. Similarly to the analysis in Section III-A, convergence to a consensus can be established for each of the two deviation terms respectively associated with $\widetilde{W}_{t,1}^{\mathcal{L}}$ and $\widetilde{W}_{t,2}^{\mathcal{L}}$. Applying the triangle inequality and Markov inequality, for all $t\geq 0$ and $i\in\mathcal{L}$ it holds

$$\begin{aligned}\mathbb{P}\left[\tilde{x}_{t+1}^{i,\mathcal{L}}>\epsilon\right] &= \mathbb{P}\left[\left|\left[\left(\widetilde{W}_{t,1}^{\mathcal{L}}+\widetilde{W}_{t,2}^{\mathcal{L}}\right)x_0^{\mathcal{L}}\right]_i\right|>\epsilon\right]\\ &\leq \mathbb{P}\left[\left|\left[\widetilde{W}_{t,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|+\left|\left[\widetilde{W}_{t,2}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|>\epsilon\right]\\ &\leq \frac{1}{\epsilon}\left(\mathbb{E}\left[\left|\left[\widetilde{W}_{t,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\right]+\mathbb{E}\left[\left|\left[\widetilde{W}_{t,2}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\right]\right).\end{aligned} \tag{26}$$

To retrieve the bound (18), we evaluate the expected values in (26) at steady state. To this aim, we use the following fact.

Lemma 4 ([10, Lemma 4]). *Let $\ell>0$ and $X,Y\in\mathbb{R}^{N\times N}$ be two sub-stochastic matrices such that $[X]_{ii}\geq\ell$ and $[Y]_{ii}\geq\ell$ for $i=1,\dots,N$. Then, it holds $[|X-Y|\mathbb{1}]_i\leq 2(1-\ell)$ for $i=1,\dots,N$ where $|A|$ is the matrix with elements $|[A]_{ij}|$.*

Bound on first term. Let $T(t)\leq t$ be the first instant such that the true weights are recovered through time t:

$$T(t) \doteq \min\left\{k\geq 0 : W_s^{\mathcal{L}} = \overline{W}^{\mathcal{L}}, s=k,\dots,t\right\}. \tag{27}$$

Note that $T(t)\leq T_{\mathrm{f}}$ for $t\geq 0$ and $\lim_{t\to\infty}T(t)\overset{\text{a.s.}}{=}T_{\mathrm{f}}$. Define

$$\Delta\widetilde{W}_{t,1}^{\mathcal{L}} \doteq \prod_{k=0}^{t}(1-\lambda_k)\left(\prod_{k=0}^{T(t)-1}W_k^{\mathcal{L}} - \prod_{k=0}^{T(t)-1}\overline{W}^{\mathcal{L}}\right). \tag{28}$$

It follows $\Delta\widetilde{W}_{\infty,1}^{\mathcal{L}} \doteq \lim_{t\to\infty}\Delta\widetilde{W}_{t,1}^{\mathcal{L}}$ with

$$\Delta\widetilde{W}_{\infty,1}^{\mathcal{L}} = \prod_{k=0}^{\infty}(1-\lambda_k)\left(\prod_{k=0}^{T_{\mathrm{f}}-1}W_k^{\mathcal{L}} - \prod_{k=0}^{T_{\mathrm{f}}-1}\overline{W}^{\mathcal{L}}\right). \tag{29}$$

In light of Lemma 2, it follows

$$\begin{aligned}\lim_{t\to\infty}\left|\left[\widetilde{W}_{t,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right| &= \left|\left[\left(\prod_{k=T_{\mathrm{f}}}^{\infty}\overline{W}^{\mathcal{L}}\right)\Delta\widetilde{W}_{\infty,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\\ &\overset{(i)}{\leq} \max_{i\in\mathcal{L}}\left|\left[\Delta\widetilde{W}_{\infty,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\\ &\overset{(ii)}{\leq} \eta\max_{i\in\mathcal{L}}\left[\left|\Delta\widetilde{W}_{\infty,1}^{\mathcal{L}}\right|\mathbb{1}\right]_i\\ &\overset{(iii)}{\leq} 2\eta\prod_{k=0}^{\infty}(1-\lambda_k)\left(1-\frac{1}{(d_{\max}+1)^{T_{\mathrm{f}}}}\right)\end{aligned} \tag{30}$$

where (i) is because $\overline{W}^{\mathcal{L}}$ is stochastic, (ii) follows from Assumption 2, and (iii) from Lemma 4 in view of (29) and the facts (see (1) and (4))

$$\left[W_t^{\mathcal{L}}\right]_{ii}\geq\frac{1}{d_{\max}+1},\qquad \left[\overline{W}^{\mathcal{L}}\right]_{ii}\geq\frac{1}{d_{\max}+1}. \tag{31}$$

Next, we find an upper bound to the infinite product in (30). The following relationship holds:

$$\begin{aligned}\prod_{k=0}^{\infty}(1-\lambda_k) &= \prod_{k=0}^{\infty}\left(1-c\mathrm{e}^{-\gamma k}\right)\\ &= \exp\left(\sum_{k=0}^{\infty}\ln\left(1-c\mathrm{e}^{-\gamma k}\right)\right)\\ &\leq \exp\left(\int_{k=0}^{\infty}\ln\left(1-c\mathrm{e}^{-\gamma(k+1)}\right)\mathrm{d}k\right).\end{aligned} \tag{32}$$

Define the dilogarithm function $\mathrm{Li}_2(z)\doteq\sum_{k=1}^{\infty}\frac{z^k}{k^2}$, then

$$\begin{aligned}\int_{k=0}^{\infty}\ln\left(1-c\mathrm{e}^{-\gamma(k+1)}\right)\mathrm{d}k &= -\frac{\mathrm{Li}_2\left(c\mathrm{e}^{-\gamma}\right)}{\gamma}\\ &= -\frac{1}{\gamma}\sum_{k=1}^{\infty}\frac{c^k\mathrm{e}^{-\gamma k}}{k^2}.\end{aligned} \tag{33}$$

Denote $\overline{s}(x)\doteq\frac{x-x\ln(1-x)+\ln(1-x)}{x}$. By recalling the identity $\sum_{k=1}^{\infty}\frac{x^k}{k(k+1)}=\overline{s}(x)$ for $|x|\leq 1$ and defining $s(\gamma)\doteq-\overline{s}(c\mathrm{e}^{-\gamma})$, it follows

$$-\frac{\mathrm{Li}_2\left(c\mathrm{e}^{-\gamma}\right)}{\gamma}\leq-\frac{1}{\gamma}\sum_{k=1}^{\infty}\frac{\left(c\mathrm{e}^{-\gamma}\right)^k}{k(k+1)}=s(\gamma). \tag{34}$$

Finally, from (30)–(32) and (34), the first expectation

in (26) can be upper bounded as follows:

$$\mathbb{E}\left[\lim_{t\to\infty}\left|\left[\widetilde{W}_{t,1}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\right] \le 2\eta \mathrm{e}^{s(\gamma)}\left(1-\mathbb{E}\left[\frac{1}{(d_{\max}+1)^{T_{\mathrm{f}}}}\right]\right) \le 2\eta \mathrm{e}^{s(\gamma)}\left(1-\left(\frac{1}{d_{\max}+1}\right)^{\mathbb{E}[T_{\mathrm{f}}]}\right) \tag{35}$$

where the second line follows from Jensen's inequality.

Bound on second term. We split the first matrix in $\widetilde{W}_{t,2}^{\mathcal{L}}$ as

$$\sum_{k=0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k = X_{t,1}+X_{t,2} \tag{36}$$

where

$$X_{t,1} \doteq \sum_{k=0}^{T(t)-2}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k \tag{37a}$$

$$X_{t,2} \doteq \sum_{k=(T(t)-1)\vee 0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)\overline{W}^{\mathcal{L}}\right)\lambda_k. \tag{37b}$$

Applying Assumption 2 and Lemmas 2 and 4 yields

$$\begin{aligned}\lim_{t\to\infty}\left|\left[\widetilde{W}_{t,2}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right| &\le \max_{i\in\mathcal{L}}\left|\left[\widetilde{W}_{\infty,2}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right| \\ &\le \eta\max_{i\in\mathcal{L}}\left[\left|\widetilde{W}_{\infty,2}^{\mathcal{L}}\right|\mathbb{1}\right]_i \\ &\le 2\eta(1-\ell)\end{aligned} \tag{38}$$

where we define $\widetilde{W}_{\infty,2}^{\mathcal{L}} \doteq \lim_{t\to\infty}\widetilde{W}_{t,2}^{\mathcal{L}}$ with

$$\widetilde{W}_{\infty,2}^{\mathcal{L}} = X_\infty - \left(1-\prod_{k=0}^{\infty}(1-\lambda_k)\right)\mathbb{1}v^\top \tag{39}$$

$$X_\infty \doteq X_{\infty,1}+X_{\infty,2} \tag{40}$$

$$X_{\infty,1} = \mathbb{1}v^\top\prod_{k=T_{\mathrm{f}}}^{\infty}(1-\lambda_k)\sum_{k=0}^{T_{\mathrm{f}}-2}\left(\prod_{s=k+1}^{T_{\mathrm{f}}-1}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k \tag{41}$$

$$X_{\infty,2} = \sum_{k=(T_{\mathrm{f}}-1)\vee 0}^{\infty}\left(\prod_{s=k+1}^{\infty}(1-\lambda_s)\overline{W}^{\mathcal{L}}\right)\lambda_k \tag{42}$$

and ℓ is a lower bound on the diagonal elements of each of the two matrices in (39). As for the second matrix, it holds:

$$\left[\left(1-\prod_{k=0}^{\infty}(1-\lambda_k)\right)\mathbb{1}v^\top\right]_{ii} \ge \ell_2 v_{\mathrm{m}} \tag{43}$$

where v_{m} and ℓ_2 are defined in Lemma 3.

Next, we separately bound the diagonal elements of the positive matrices $X_{\infty,1}$ and $X_{\infty,2}$. Consider the inequality

$$\prod_{s=k}^{K-1}(1-\lambda_s) > (1-\lambda_k)^{\frac{1}{\lambda_k}\sum_{s=k}^{K-1}\lambda_s} = (1-c\mathrm{e}^{-\gamma k})^{\frac{1-\mathrm{e}^{-\gamma(K-k)}}{1-\mathrm{e}^{-\gamma}}}. \tag{44}$$

From (44), The infinite product in (41) can be bounded as

$$\prod_{k=T_{\mathrm{f}}}^{\infty}(1-\lambda_k) > \left(1-c\mathrm{e}^{-\gamma T_{\mathrm{f}}}\right)^{\frac{1}{1-\mathrm{e}^{-\gamma}}}. \tag{45}$$

Consider now the matrix summation in $X_{\infty,1}$. It holds

$$\begin{aligned}&\left[\sum_{k=0}^{T_{\mathrm{f}}-2}\left(\prod_{s=k+1}^{T_{\mathrm{f}}-1}(1-\lambda_s)W_s^{\mathcal{L}}\right)\lambda_k\right]_{ii} \\ &\quad\ge \sum_{k=0}^{T_{\mathrm{f}}-2}\left(\prod_{s=k+1}^{T_{\mathrm{f}}-1}\frac{1-\lambda_s}{d_{\max}+1}\right)\lambda_k \\ &\quad\ge \sum_{k=0}^{T_{\mathrm{f}}-2}\left(\frac{1-\lambda_{k+1}}{d_{\max}+1}\right)^{T_{\mathrm{f}}-k-1}\lambda_k \\ &\quad\ge \left(\frac{1-\lambda_1}{d_{\max}+1}\right)^{T_{\mathrm{f}}-1}\sum_{k=0}^{T_{\mathrm{f}}-2}\lambda_k \\ &\quad\ge c\left(\frac{1-c\mathrm{e}^{-\gamma}}{d_{\max}+1}\right)^{T_{\mathrm{f}}-1}\frac{1-\mathrm{e}^{-\gamma(T_{\mathrm{f}}-1)}}{1-\mathrm{e}^{-\gamma}}\mathbb{1}_{\{T_{\mathrm{f}}>1\}}\end{aligned} \tag{46}$$

and the diagonal elements of $X_{\infty,1}$ are bounded as follows:

$$[X_{\infty,1}]_{ii} \ge v_{\mathrm{m}}\left(1-c\mathrm{e}^{-\gamma T_{\mathrm{f}}}\right)^{\frac{1}{1-\mathrm{e}^{-\gamma}}}c\left(\frac{1-c\mathrm{e}^{-\gamma}}{d_{\max}+1}\right)^{T_{\mathrm{f}}-1} \cdot\frac{1-\mathrm{e}^{-\gamma(T_{\mathrm{f}}-1)}}{1-\mathrm{e}^{-\gamma}}\mathbb{1}_{\{T_{\mathrm{f}}>1\}}. \tag{47}$$

Then, we bound the diagonal elements of $X_{\infty,2}$ as

$$\begin{aligned}[X_{\infty,2}]_{ii} &\ge \left[\left(\prod_{k=T_{\mathrm{f}}\vee 1}^{\infty}(1-\lambda_k)\overline{W}^{\mathcal{L}}\right)\lambda_{(T_{\mathrm{f}}-1)\vee 0}\right]_{ii} \\ &\ge v_{\mathrm{m}}\left(1-c\mathrm{e}^{-\gamma(T_{\mathrm{f}}\vee 1)}\right)^{\frac{1}{1-\mathrm{e}^{-\gamma}}}c\mathrm{e}^{-\gamma((T_{\mathrm{f}}-1)\vee 0)}\end{aligned} \tag{48}$$

and

$$[X_\infty]_{ii} = [X_{\infty,1}]_{ii}+[X_{\infty,2}]_{ii} \ge \ell_1 v_{\mathrm{m}} \tag{49}$$

where ℓ_1 is defined in (21). Finally, the two matrices in (39) have diagonal elements lower bounded by $\ell = \min\{\ell_1,\ell_2\}$, and we bound the second expectation in (26) through (38) as

$$\mathbb{E}\left[\lim_{t\to\infty}\left|\left[\widetilde{W}_{t,2}^{\mathcal{L}}x_0^{\mathcal{L}}\right]_i\right|\right] \le 2\eta\left(1-\mathbb{E}\left[\min\{\ell_1,\ell_2\}\right]\right). \tag{50}$$

Finally, the probability (18) can be bounded according to (26) by plugging in (35) and (50). □

A few remarks are in order to understand the meaning of bound (19) and how it behaves as γ varies. For convenience, we recall the expression of the bound below:

$$\mathbb{P}\left[\lim_{t\to\infty}\tilde{x}_t^{i,\mathcal{L}} > \epsilon\right] < \eta u_{\mathcal{L}}(\epsilon),\ u_{\mathcal{L}}(\epsilon) \propto \mathrm{e}^{s(\gamma)} - \mathbb{E}[\ell]. \tag{51}$$

The behavior of the term $u_{\mathcal{L}}(\epsilon)$ is mainly affected by two functions of γ, which are $\mathrm{e}^{s(\gamma)}$ and $\mathbb{E}[\ell]$.

The first function, ruled by $s(\gamma)$, expresses the deviation due to following the protocol (RES) with the learned weights (4) rather than with the (unknown) true weights (1), and it is increasing with γ. In words, this suggests that setting γ small (*i.e.*, making λ_t decay slowly overtime) is beneficial to performance because it lets legitimate agents learn the trustworthy neighbors while at the same keeping the weights balanced (thus avoiding biases caused by misclassification of legitimate neighbors) during this learning process. This

is reminiscent of the strategy proposed in [10], where the consensus protocol starts at T_0 and a larger value of T_0 helps to reduce the deviation term associated with data exchange among legitimate agents. Moreover, the coefficient that multiplies $\mathrm{e}^{s(\gamma)}$ increases with $\mathbb{E}[T_\mathrm{f}]$, so that, for any choice of γ, the deviation is larger for larger T_f.

The second function appearing in $u_{\mathcal{L}}(\epsilon)$ is proportional to the negative expectation of ℓ w.r.t. T_f and expresses the impact of the input term $\lambda_t x_0^i$ in (RES) that anchors the legitimate agents to their initial condition. It is not easy to analytically evaluate the minimum $\ell = \min\{\ell_1, \ell_2\}$, in general. Nonetheless, the following facts hold:

(1) the term ℓ_2 is strictly decreasing with γ;
(2) the term ℓ_1 is strictly increasing with γ for $T_\mathrm{f} \leq 1$, while for $T_\mathrm{f} > 1$ it is the product between an increasing and a decreasing function;
(3) the limits of the two terms evaluate

$$\ell_1^0 \doteq \lim_{\gamma\to 0} \ell_1 = 0 \tag{52a}$$

$$\ell_2^0 \doteq \lim_{\gamma\to 0} \ell_2 = v_\mathrm{m} \tag{52b}$$

$$\ell_1^\infty \doteq \lim_{\gamma\to \infty} \ell_1 = v_\mathrm{m} c \mathbb{1}_{\{T_\mathrm{f}\leq 1\}} + \frac{v_\mathrm{m} c \mathbb{1}_{\{T_\mathrm{f}>1\}}}{(d_{\max}+1)^{T_\mathrm{f}-1}} \tag{52c}$$

$$\ell_2^\infty \doteq \lim_{\gamma\to \infty} \ell_2 = c v_\mathrm{m} \tag{52d}$$

and it follows

$$\ell_1^0 < \ell_2^0, \qquad \ell_1^\infty \leq \ell_2^\infty \tag{53}$$

where the equality $\ell_1^\infty = \ell_2^\infty$ holds if and only if $T_\mathrm{f} \leq 1$;
(4) from (52c), it follows that ℓ_1^∞ decreases with T_f and $\lim_{T_\mathrm{f}\to\infty} \ell_1^\infty = 0$.

From the items (1)–(3) above and continuity of ℓ_1 and ℓ_2, we infer the following result, summarized as a lemma.

Lemma 5. *If $T_f \leq 1$, it holds $\ell = \ell_1$. If $T_f > 1$, there exist $\bar\gamma_1 > 0$ and $\bar\gamma_2 > 0$ such that $\ell = \ell_1$ for $\gamma < \bar\gamma_1$ and $\gamma > \bar\gamma_2$.*

Lemma 5 implies that, for $T_\mathrm{f} \leq 1$, $\ell = \ell_1$ and thus the term $-\ell$ is decreasing with γ.

On the other hand, it is analytically difficult to infer how the term ℓ_1 (and hence ℓ) behaves for $T_\mathrm{f} > 1$ and generic values of γ. Nonetheless, the fact (4) above suggests that, as T_f grows, ℓ_1 should have a non-monotonic behavior and admit a nontrivial maximum. Indeed, numerical tests show that $-\ell_1$ is minimized at a finite value of γ for every $T_\mathrm{f} > 1$, and that such a minimizer decreases as T_f increases, as illustrated in Fig. 1. Considering that the bound (19) is proportional to $-\ell$ in expectation, this further suggests that the deviation decreases for small values of γ and increases for large values, with the minimum point that shifts towards $\gamma = 0$ as the more likely values of T_f increase. This behavior means that, if the legitimate agents need time to learn which neighbors are trustworthy, they should act more cautiously and increase the parameter λ_t, especially at the beginning.

To summarize, the bound (19) on the deviation due to misclassifications of legitimate agents is numerically seen to be quasi-convex with γ, with the minimum point that approaches $\gamma = 0$ as T_f increases, *i.e.,* according to how difficult learning the true weights $\overline{W}^{\mathcal{L}}$ is.

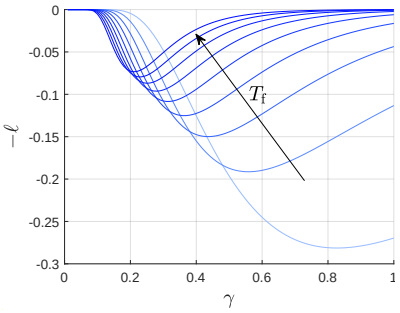

Fig. 1. Profile of $-\ell$ in (19) as a function of γ with $T_\mathrm{f} \in \{2, \dots, 10\}$ (the arrow indicates how the curve varies as T_f grows). Recall that the bound on (probability of) deviation due to legitimate agents is proportional to $-\mathbb{E}[\ell]$.

2) Malicious Agents: Because malicious agents cannot be trusted, the protocol (RES) should ideally annihilate their contribution to legitimate agents' states. Hence, we define the deviation term due to malicious agents as

$$\tilde{x}_{t+1}^{i,\mathcal{M}} \doteq \left|\left[\bar{x}_{t+1}^{\mathcal{M}}\right]_i\right|. \tag{54}$$

We have the following result that bounds their harmful effects on legitimate agents.

Lemma 6. *The deviation from nominal consensus due to malicious agents' contribution can be bounded as*

$$\mathbb{P}\left[\lim_{t\to\infty} \tilde{x}_t^{i,\mathcal{M}} > \epsilon\right] \leq \eta u_{\mathcal{M}}(\epsilon), \qquad \forall i \in \mathcal{L} \tag{55}$$

where

$$u_{\mathcal{M}}(\epsilon) \doteq \frac{\min\{d_{max}, M\}}{2\epsilon}\,\mathbb{E}[\xi] \tag{56}$$

and we define

$$\xi \doteq \frac{1-\mathrm{e}^{-2E_{\mathcal{M}}^2 T_f}}{\mathrm{e}^{2E_{\mathcal{M}}^2}-1} - \frac{c(1+\mathrm{e}^{-\gamma})\left(1-\mathrm{e}^{-(\gamma+2E_{\mathcal{M}}^2)T_f}\right)}{\mathrm{e}^{2E_{\mathcal{M}}^2}-\mathrm{e}^{-\gamma}} + \frac{c^2\mathrm{e}^{-\gamma}\left(1-\mathrm{e}^{-2(\gamma+E_{\mathcal{M}}^2)T_f}\right)}{\mathrm{e}^{2E_{\mathcal{M}}^2}-\mathrm{e}^{-2\gamma}}. \tag{57}$$

Proof. From (8b) and (54), it follows

$$\begin{aligned}
\tilde{x}_{t+1}^{i,\mathcal{M}} &= \left|\left[\sum_{k=0}^{t}\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)(1-\lambda_k)W_k^{\mathcal{M}}x_k^{\mathcal{M}}\right]_i\right| \\
&\overset{(i)}{\leq} \sum_{k=0}^{t}\left|\left[\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)(1-\lambda_k)W_k^{\mathcal{M}}x_k^{\mathcal{M}}\right]_i\right| \\
&\overset{(ii)}{\leq} \eta\sum_{k=0}^{t}\left[\left(\prod_{s=k+1}^{t}(1-\lambda_s)W_s^{\mathcal{L}}\right)(1-\lambda_k)W_k^{\mathcal{M}}\mathbb{1}\right]_i \\
&\overset{(iii)}{\leq} \eta\sum_{k=0}^{t}(1-\lambda_{k+1})(1-\lambda_k)\left[W_k^{\mathcal{M}}\mathbb{1}\right]_i
\end{aligned} \tag{58}$$

where (i) follows from the triangle inequality, (ii) from Assumption 2, and (iii) because $\{W_t^{\mathcal{L}}\}_{t\geq 0}$ are sub-stochastic

matrices and λ_t is a decreasing sequence with $0 < 1-\lambda_t < 1$. Then, the weights given to malicious agents are bounded as

$$\left[W_t^{\mathcal{M}}\mathbb{1}\right]_i = \sum_{j=1}^{M}\left[W_t^{\mathcal{M}}\right]_{ij} \leq \frac{1}{2}\sum_{j\in\mathcal{M}}\mathbb{1}_{\beta_{ij}(t)\geq 0}. \tag{59}$$

Recall that $\beta_{ij}(k) < 0$ for $k > T(t)$ and $j \in \mathcal{M}$. It follows

$$\begin{aligned}\mathbb{E}\left[\tilde{x}_{t+1}^{i,\mathcal{M}}\right] &\leq \eta\mathbb{E}\left[\sum_{k=0}^{t}(1-\lambda_{k+1})(1-\lambda_k)\frac{1}{2}\sum_{j\in\mathcal{M}}\mathbb{1}_{\beta_{ij}(k)\geq 0}\right]\\ &= \frac{\eta}{2}\sum_{k=0}^{T(t)-1}(1-\lambda_{k+1})(1-\lambda_k)\sum_{j\in\mathcal{M}}\mathbb{P}\left[\beta_{ij}(k)\geq 0\right]\\ &\leq \frac{\eta}{2}\sum_{k=0}^{T(t)-1}(1-\lambda_{k+1})(1-\lambda_k)\sum_{j\in\mathcal{M}\cap\mathcal{N}_i}\mathrm{e}^{-2(k+1)E_{\mathcal{M}}^2}\\ &\leq \frac{\min\{d_{\max},M\}\eta}{2}z_t\end{aligned} \tag{60}$$

where we define

$$z_t \doteq \sum_{k=0}^{T(t)-1}(1-\lambda_{k+1})(1-\lambda_k)\mathrm{e}^{-2(k+1)E_{\mathcal{M}}^2}. \tag{61}$$

From $\lim_{t\to\infty}T(t) \overset{\text{a.s.}}{=} T_{\mathrm{f}}$, it holds $\lim_{t\to\infty} z_t \overset{\text{a.s.}}{=} \xi$. It follows

$$\begin{aligned}\mathbb{E}\left[\lim_{t\to\infty}\tilde{x}_t^{i,\mathcal{M}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\lim_{t\to\infty}\tilde{x}_t^{i,\mathcal{M}}\middle|T_{\mathrm{f}}\right]\right]\\ &\leq \mathbb{E}\left[\frac{\min\{d_{\max},M\}\eta}{2}\mathbb{E}\left[\lim_{t\to\infty}z_t|T_{\mathrm{f}}\right]\right]\\ &\overset{\text{a.s.}}{=} \frac{\min\{d_{\max},M\}\eta}{2}\mathbb{E}\left[\xi\right]\end{aligned} \tag{62}$$

and (55) readily follows by applying Markov inequality. □

The bound $u_{\mathcal{M}}(\epsilon)$ in (55) increases with γ through the parameter ξ. This is intuitive: if λ_t is larger, the legitimate agents are less sensitive to their neighbors' states as per (RES), thus they are also more resilient against malicious transmissions and the corresponding deviation term $\tilde{x}_t^{i,\mathcal{M}}$ is smaller. Also, $u_{\mathcal{M}}(\epsilon)$ increases with $E_{\mathcal{M}}$ and T_{f}, suggesting that higher uncertainty in classification of malicious agents (represented by greater $E_{\mathcal{M}}$ or T_{f}) yields a larger deviation, on average.

3) Bound on Deviation: The overall bound on the deviation from nominal consensus can be computed by merging the two bounds obtained for legitimate and malicious agents' contributions. Applying the triangle inequality to (14), (16), and (54) yields

$$\tilde{x}_t^i \leq \tilde{x}_t^{i,\mathcal{L}} + \tilde{x}_t^{i,\mathcal{M}}. \tag{63}$$

We have the following result that quantifies how distant from the nominal consensus the legitimate agents eventually get.

Theorem 1 (Deviation from nominal consensus)**.** *The deviation from nominal consensus is upper bounded as*

$$\mathbb{P}\left[\lim_{t\to\infty}\tilde{x}_t^i > \epsilon\right] \leq \eta u(\epsilon), \qquad i\in\mathcal{L} \tag{64}$$

with

$$u(\epsilon) \doteq u_{\mathcal{L}}\left(\frac{\epsilon}{2}\right) + u_{\mathcal{M}}\left(\frac{\epsilon}{2}\right). \tag{65}$$

Proof. It follows by applying the union bound to (63) and then invoking Lemmas 3 and 6. □

We can assess the impact of a specific choice of λ_t by observing the overall deviation bound (64). Recall that, in light of the expression (9), larger values of γ correspond to faster decay of λ_t — *i.e.,* the standard consensus protocol is recovered more quickly. In view of what remarked for the two bounds $u_{\mathcal{L}}(\epsilon)$ and $u_{\mathcal{M}}(\epsilon)$, the bound $u(\epsilon)$ above suggests that the steady-state deviation from nominal consensus decreases for small values of γ and increases as γ is chosen larger. The presence of a nonzero point of minimum, which intuitively corresponds to an optimal design of γ, is caused by the input term $\lambda_t x_0^i$ added to the standard consensus in (RES) to enhance resilience, and represents a possible loss in performance due to forcing a suboptimal protocol for too long compared to the time needed for correct detection of adversaries. In particular, the term ℓ appearing in $u_{\mathcal{L}}(\epsilon)$ (see (19)) suggests that the optimal γ decreases as T_{f} increases, reflecting the need of legitimate agents to act more cautiously when the uncertainty in the trust variables $\alpha_{ij}(t)$ is higher. On the other hand, the term $u_{\mathcal{M}}(\epsilon)$ requires γ to be small (slow decay of λ_t) to annihilate the effect of malicious agents.

Remark 3 (Immediate detection). If $T_{\mathrm{f}} = 0$, the term $\tilde{x}_t^{i,\mathcal{M}}$ is zero and the deviation bound reduces to $u(\epsilon) = u_{\mathcal{L}}(\epsilon)$, which is strictly decreasing with γ for $T_{\mathrm{f}} = 0$. This confirms that, if the legitimate agents immediately identify their adversaries, the optimal choice is to remove λ_t and follow the standard consensus protocol with the learned (true) weights.

Remark 4 (Nominal scenario). In the ideal case with no malicious agents, the deviation term $\tilde{x}_t^{i,\mathcal{M}}$ is identically zero and the choice of γ affects the deviation only through the term $\tilde{x}_t^{i,\mathcal{L}}$ due to misclassifying legitimate agents.

IV. Numerical Simulations

To test the effectiveness of the proposed resilient consensus protocol and the design insight suggested by the bound proposed in Theorem 1, we run numerical simulations with a sparse network with 50 legitimate agents and 10 malicious agents. The communication links are modeled via a random geometric graph with communication radius equal to 0.2, the agents being spread across the ball $[0,1]\times[0,1]\in\mathbb{R}^2$. The initial states of legitimate agents are randomly drawn from the uniform distribution $\mathcal{U}(0,\eta)$ with $\eta = 1$, while the malicious agents follow an oscillatory trajectory about the mean value $2x_{\mathrm{ss}}^{\mathcal{L},*}$ (twice the nominal consensus value) under additive zero-mean Gaussian noise with standard deviation 0.05. Note that, in the absence of data-driven detection mechanisms (the malicious agents are classified based on the trust information $\alpha_{ij}(t)$ that comes from physical transmissions and not based on the states they transmit), this behavior is most harmful because it steadily drives the legitimate agents far away from the nominal consensus value. Moreover, it holds $0 < 2x_{\mathrm{ss}}^{\mathcal{L},*} < 1$ and the random oscillations of malicious agents are small compared to their mean value, which verifies Assumption 2.

We run the proposed protocol (RES) with λ_t according (9) with $c = 0.9$ for $T = 1000$ iterations and average all

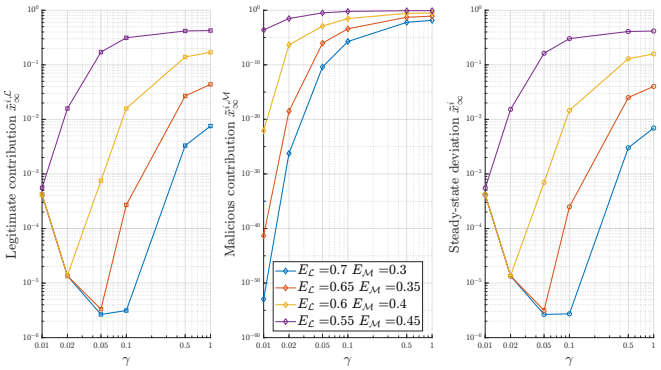

Fig. 2. Steady-state deviation from nominal consensus value (right) and contributions due to legitimate agents $\tilde{x}_\infty^{i,\mathcal{L}}$ (left) and to malicious agents $\tilde{x}_\infty^{i,\mathcal{M}}$ (middle) averaged over 1000 Monte Carlo runs. As predicted by the bound (19), the deviation term due to misclassification of legitimate agents is minimized by a (small) positive value of γ that decreases as the trust scores get more uncertain, while the deviation term due to malicious agents steadily increases as γ grows, according to bound (55).

results across 1000 Monte Carlo runs. We report four different setups with different values of $E_\mathcal{L}$ and $E_\mathcal{M}$ that respectively increase from 0.55 to 0.7 and decrease from 0.45 to 0.3. In all experiments, the trust observations of legitimate (resp., malicious) transmissions are drawn from the uniform distribution centered at $E_\mathcal{L}$ (resp., $E_\mathcal{M}$) with length equal to twice the minimum between $1 - E_\mathcal{L}$ and $E_\mathcal{M}$.

The outcomes are depicted in Fig. 2 that shows overall steady-state deviation $\tilde{x}_\infty^i$ (14) in the right box together with maximal deviation due to legitimate agents $\tilde{x}_\infty^{i,\mathcal{L}}$ (16) in the left box and maximal deviation due to malicious agents $\tilde{x}_\infty^{i,\mathcal{M}}$ (54) in the middle box. It can be seen that the simulated behavior agrees with the analytical bound in Theorem 1: indeed, the deviation term due to malicious agents steadily increases as γ grows, while the deviation associated with misclassification of legitimate neighbors is minimized by a nonzero value of γ that decreases with the uncertainty of trust variables. For example, when $E_\mathcal{L} = 0.7$ and $E_\mathcal{M} = 0.3$, the trust scores are very informative and the deviation is minimized at $\gamma = 0.05$, which dictates a relatively fast decay of the parameter λ_t. Conversely, in the case $E_\mathcal{L} = 0.55$ and $E_\mathcal{M} = 0.45$, the trust variables are more uncertain and the optimal choice is given by $\gamma = 0.01$, corresponding to a much slower decay of λ_t. Moreover, as the uncertainty in the trust variables increase, it is more difficult for legitimate agents to correctly classify their neighbors, which leads to the monotonic increase observed across all deviation terms for every choice of γ.

V. Conclusions

We propose a resilient consensus protocol that uses trustworthiness information derived from the physical transmission channel to progressively detect malicious agents, and complements this information with a time-varying scaling that accounts for how confident the agent is about its neighbors being malicious or not. Analytical results demonstrate that the proposed protocol leads to a consensus almost surely. Also, the asymptotic deviation is upper bounded by a non-monotonic function of the decay rate of the confidence parameter. Numerical results corroborate these findings, suggesting that the confidence parameter can be optimally tuned so as to minimize the steady-state deviation.

References

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

[2] J. Usevitch and D. Panagou, "Resilient Leader-Follower Consensus to Arbitrary Reference Values in Time-Varying Graphs," *IEEE Trans. Automat. Contr.*, vol. 65, no. 4, pp. 1755–1762, Apr. 2020.

[3] Y. Shang, "Resilient consensus in multi-agent systems with state constraints," *Automatica*, vol. 122, p. 109288, Dec. 2020.

[4] J. S. Baras and X. Liu, "Trust is the Cure to Distributed Consensus with Adversaries," in *Proc. Mediterr. Conf. Control Autom.*, Jul. 2019, pp. 195–202.

[5] L. Ballotta, G. Como, J. S. Shamma, and L. Schenato, "Competition-based resilience in distributed quadratic optimization," in *Proc. IEEE Conf. Decis. Control*, 2022.

[6] W. Abbas, A. Laszka, and X. Koutsoukos, "Improving Network Connectivity and Robustness Using Trusted Nodes With Application to Resilient Consensus," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 4, pp. 2036–2048, Dec. 2018.

[7] S. Sundaram and B. Gharesifard, "Consensus-based distributed optimization with malicious nodes," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2015, pp. 244–249.

[8] Y. Yi, Y. Wang, X. He, S. Patterson, and K. H. Johansson, "A sample-based algorithm for approximately testing r-robustness of a digraph," in *Proc. IEEE Conf. Decis. Control*, 2022, pp. 6478–6483.

[9] T. Wheeler, E. Bharathi, and S. Gil, "Switching topology for resilient consensus using wi-fi signals," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 2018–2024.

[10] 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, Feb. 2022.

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

[12] M. Yemini, A. Nedić, S. Gil, and A. Goldsmith, "Resilience to malicious activity in distributed optimization for cyberphysical systems," in *Proc. IEEE Conf. Decis. Control*, 2022.

[13] M. Yemini, A. Nedić, A. Goldsmith, and S. Gil, "Resilient distributed optimization for multi-agent cyberphysical systems," *arXiv:2212.02459*, 2022.

[14] C. N. Hadjicostis and A. D. Domínguez-García, "Trustworthy distributed average consensus," in *Proc. IEEE Conf. Decis. Control*. IEEE, 2022, pp. 7403–7408.

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

[16] A. Tsiamis, K. Gatsis, and G. J. Pappas, "State-secrecy codes for networked linear systems," *IEEE Trans. Autom. Control*, vol. 65, no. 5, pp. 2001–2015, 2020.

[17] 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.

[18] L. Ballotta, G. Como, J. S. Shamma, and L. Schenato, "Can competition outperform collaboration? The role of misbehaving agents," *arXiv e-prints*, 2022, arXiv:2207.01346 (pdf).

[19] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *J. Math. Sociol.*, vol. 15, no. 3-4, pp. 193–206, 1990.