Can Competition Outperform Collaboration? The Role of Malicious Agents

Luca Ballotta<sup>0</sup>, Graduate Student Member, IEEE, Giacomo Como<sup>0</sup>, Member, IEEE, Jeff S. Shamma<sup>0</sup>, Fellow, IEEE, and Luca Schenato<sup>0</sup>, Fellow, IEEE

Abstract—We investigate a novel approach to resilient distributed optimization with quadratic costs in a Networked Control System prone to exogenous attacks that make agents misbehave. In contrast with commonly adopted filtering strategies, we draw inspiration from a game-theoretic formulation of the consensus problem and argue that adding competition to the mix can improve resilience in the presence of malicious agents. Our intuition is corroborated by analytical and numerical results showing that (i) our strategy reveals a nontrivial performance trade-off between full collaboration and full competition, and (ii) such competition-based approach can outperform state-of-the-art algorithms based on Mean Subsequence Reduced. Finally, we study impact of communication topology and connectivity on performance, pointing out insights to robust network design.

I. INTRODUCTION

With great power comes great responsibility, and Networked Control Systems have great power indeed. From smart grids managing energy consumption [1], [2] to sensor networks able to monitor vast areas [3], to fleets of autonomous vehicles for intelligent transportation [4], [5], everyday life depends more and more on control of interacting devices.

While this brings numerous benefits, a major drawback is that malicious agents can locally intrude from any point in the system, and cause serious damage at global scale. Recently, Department of Energy secretary stated that enemies of the United States can shut down the U.S. power grid, and it is known that hacking groups around the world have high technological sophistication [6]. Cyberattacks hit Italian healthcare infrastructures during the COVID-19, disrupting services for weeks [7]. Another concern is accidental failures spreading from single source nodes. Cascading failure damages have notable examples, from electricity blackouts over large areas, to denial of service of web applications. Furthermore, as new frontiers through massively connected devices in Networked Control Systems and the Internet-of-Things are breached, thanks to powerful communication protocols such as 5G and 6G, this problem will only gain in importance.

This work has been partially supported by the Italian Ministry of Education, University and Research (MIUR) through the PRIN project no. 2017NS9F3EY entitled “Realtime Control of 5G Wireless Networks”, and through the initiative “Departments of Excellence” (Law 232/2016).

Luca Ballotta and Luca Schenato are with the Department of Information Engineering, University of Padova, 35131 Padova, Italy (e-mail: ballotta@dei.unipd.it; schenato@dei.unipd.it).

Giacomo Como is with the Department of Mathematical Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy (e-mail: giacomo.como@polito.it).

Jeff S. Shamma is with the Industrial and Enterprise Systems Engineering (ISE), University of Illinois Urbana-Champaign, Illinois, USA (e-mail: jshamma@illinois.edu).

The above problems have been addressed in literature for several years. A body of work develops or applies control techniques to overcome fragility in specific applications or scenarios. Examples are power outage in smart grids [8], [9], cascading failures in cyber-physical systems [10]–[13], denial of service [14], [15], robotic rendez-vous [16], distributed estimation [17], to mention a few. From a methodological perspective, related literature in control and optimization theory has mostly focused on robustness of distributed algorithms and control protocols to a fraction of misbehaving agents. This model can tailor either intentionally malicious agents, induced by cyber-attacks, or faulty agents, as a consequence of hardware faults or software bugs. A fundamental subclass of such approaches is resilient consensus, aimed to enforcing consensus of normally behaving agents in the face of unknown interacting adversaries. The consensus problem has been deeply studied in the past two decades [18], and underlies a plethora of application domains. In particular, average consensus is a crucial tool in, e.g., distributed estimation [17] and optimization [19], [20], management of power grids [21], distributed Federated Learning [22], [23], among others. However, the classical consensus algorithm is fragile, as misbehaving agents can easily deceive the rest of the network. To tame this issue, the most popular approach is based on the filtering technique referred as “Mean Subsequence Reduced” (MSR), whereby agents discard suspicious incoming data (extreme values) from local updates [24]. In particular, seminal work [25] introduced a weighted version (W-MSR) and the notion of $r$-robustness, a suitable measure enabling theoretical guarantees for resilient consensus. Among the many variants and adaptations of W-MSR, [26] studies resilient control for double integrators, [27] tackles mobile adversaries, [28] focuses on leader-follower framework, [29] targets nonlinear systems with state constraints, and [30]–[32] consider generic cost functions, aiming to achieve resilience in general distributed optimization.

Other approaches proposed in literature may not rely on active filtering of incoming neighbor information, but shift attention from update protocol to agent capabilities. For example, [33] employs a buffer to store all values received from other agents and replaces the thresholding mechanism with a voting strategy followed by dynamical updates. [34] proposes dynamically switching update rule for continuous-time double integrators, and [35]–[38] use stochastic or heuristic trust scores to filter out potentially malicious transmissions.
We draw inspiration from game theory. Distributed cooperative averaging is a well-studied problem in control and optimization. However, the cost of achieving average consensus is due to inter-agent competition, in an attempt to mitigate potential attacks from neighbors, but doing so prevents agents to reach consensus. The tunable parameter $\lambda \in [0, 1]$ allows normal agents to smoothly transition from full collaboration, where they trust equally all agents in the network, to full competition, where they trust only themselves, inducing a richer spectrum of behaviors at both local and global scale.

providing probabilistic bounds on detection, convergence, or deviation from average consensus. While such approaches may overcome limitations of MSR-based strategies, they usually require either stronger assumptions on the network (e.g., trusted agents) or burdening local computation or storage resources.

### B. Novel Contribution

Despite the success of MSR-based approaches both in literature and applications, their theoretical guarantees depend on a minimal level of network $r$-robustness that allows agents to reach resilient consensus. In fact, little can be said about their steady-state behavior when such robustness conditions do not hold. While algorithms might still work in some cases, the lack of theoretical guarantees may be undesired in some applications, where a more conservative but safer approach may be preferred. Moreover, while in some cases agents may agree upon a common value, other tasks require average consensus to succeed. Thus, we depart from classical MSR-based strategies, seeking algorithms that can offer theoretical guarantees in a broader sense – while ensuring some level of resilience.

Towards this goal, we set the stage with two key moves. Firstly, we replace the hard constraint of achieving consensus with the cost of a distributed optimization problem. Secondly, we draw inspiration from game theory. Distributed cooperative control and games, despite their apparent contrast, are linked from several perspectives which have been largely explored in literature [39]–[45]. In particular, in [46] the authors discuss connections between consensus and potential games. Stepping forward, we propose to use the celebrated Friedkin-Johnsen (FJ) dynamics [47] as an approach to achieve resilient distributed quadratic optimization. One key feature of this model is a tunable parameter $\lambda \in [0, 1]$ which allows to smoothly transition from full collaboration, where each normally behaving agent puts equal trust in all agents in the network (including itself), leading to standard average consensus, to full competition, whereby agents do not trust each other, namely they regard all other agents as adversaries. Such a mixed approach allows to explore the performance trade-off that arises from different choices of agents that may decide to trust their neighbors or not, which turns out to be crucial in the presence of adversaries. In fact, we observe a fundamental competition-collaboration trade-off: in general, the optimal choice to achieve resilience is a hybrid strategy that makes agents trust neighbors only partially, as illustrated in Fig. 1. In particular, the global cost function (blue) is the sum of two conflicting contributions, representing respectively deception due to collaboration with malicious agents (red) and inefficiency caused by competition against agent’s neighbors (yellow). To achieve analytical intuition, we discuss such competition-collaboration trade-off using tools from opinion dynamics, in particular social power and FJ model, that shed light on the role of malicious agents and parametrization of agent dynamics on optimization performance.

After characterizing and validating the proposed competition-based protocol, we fix agent update rule and shift attention to the network topology, in order to assess how the latter impacts performance of regular agents with respect to distributed optimization. In particular, we use regular graphs to numerically show that network connectivity can mitigate malicious attacks, and how performance varies when the topology gets sparser. In particular, we heuristically observe that not only high connectivity, but also degree balance across agents is useful to tame unknown adversaries, that could intuitively exploit highly connected agents to disrupt the optimization at network level.

Besides new results, this paper extends the preliminary conference version [48] in two ways. Firstly, we consider a more general distribution (non-i.i.d.) of nominal priors of agents. Secondly, we extend simulations and compare our proposed strategy with both classical W-MSR [25] and recently proposed SABA [33].

### C. Organization of the Paper

We motivate average consensus for distributed optimization tasks in Section II addressing a class of misbehaving agents in Section II-A. In Section III, we propose a competition-based strategy to enhance resilience of regular agents. In particular, we describe the link with game theory (Section III-A), characterize the cost function and its minimizer (Sections III-B–III-D), and perform numerical experiments (Section IV). Further, we offer analytical insight to interpret the competition-collaboration trade-off from a formal standpoint in Section IV-A. To assess effectiveness of our approach, we perform simulations on large-scale systems, showing that it can provide superior performance to MSR-based methods (Section V). Then, we study impact of communication topology on performance in Section VI. We conclude by addressing open questions and compelling avenues for future research in Section VII.

### II. SETUP AND PROBLEM FORMULATION

We consider a Networked Control System composed of $N$ agents in the set $\mathcal{V} = \{1, \ldots, N\}$. The state of agent $i \in \mathcal{V}$ is denoted by $x_i \in \mathbb{R}$, and all states are stacked in the column vector $\mathbf{x} = [x_1, \ldots, x_N]^T$. The system dynamics are governed by the following model:

$$\dot{x}_i = f_i(x_i, \mathbf{u}_i)$$

where $f_i$ is a function that depends on the state $x_i$ and the control input $\mathbf{u}_i$.
vector \( x \in \mathbb{R}^N \). At the beginning, each agent \( i \) carries local information encoded by prior \( \theta_i \in \mathbb{R} \).

**Assumption 1.** Priors are distributed as random variables with zero mean and covariance matrix \( \Sigma = \Sigma^T \in \mathbb{R}^{N \times N} \), \( \Sigma > 0 \), where \( \Sigma_{ii} = \sigma_i^2 > 0 \forall i \in \mathcal{V} \) and \( \Sigma_{ij} = \sigma_{ij} \forall i \neq j \).

Within the network, some agents behave according to the control task at hand, while others cannot be controlled by reasons and behave arbitrarily. We call the former regular and the latter malicious agents. Each regular agent \( i \) needs to minimize the mismatch \( f_{\text{local}} \) involving priors of regular agents, \( \Sigma \) (Priors of malicious agents)

\[
\Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{12} & \Sigma_{22}
\end{bmatrix}
\]

where \( \Sigma \) is the covariance matrix of noises and actual priors by \( \Sigma = + \), respectively. Without loss of generality, we re-label agents as \( \mathcal{R} = \{1, \ldots, R\} \) and \( \mathcal{M} = \{R + 1, \ldots, N\} \), so that matrix \( \Sigma \) can be conveniently partitioned as

\[
V = \begin{bmatrix}
0 & 0 \\
0 & V_M
\end{bmatrix}, \quad V_M = \text{diag}(d_{R+1}, \ldots, d_N)
\]

and, accordingly,

\[
\Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{12} & \Sigma_{22}
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{12} & \Sigma_{22}
\end{bmatrix}
\]

Remark 1 (Malicious agents disrupt optimization). Assumption 3 is a worst-case scenario for average consensus, whereby malicious agents constantly pull regular agents towards a value which is possibly very distant from the average of nominal priors. While this may still allow regular agents to reach consensus, it will drive them far off from average consensus, thus disrupting the distributed optimization tasks.

**Remark 2** (Intelligent malicious behavior). In resilient consensus literature, it is common to assume worst-case behavior of the malicious agents, irrespectively of the detection capabilities of regular agents. In practice, malicious agents may behave in less suspicious way than what stated in Assumption 3 to avoid being detected, which should be contrasted by suitable resilient strategies. This scenario goes beyond the scope of this paper, where we are interested in studying the competition-collaboration trade-off, and is left for future research.

### III. Resilient Average Consensus

#### A. The Consensus Problem and Game Theoretic Models

Because classical consensus is fragile to misleading agents \( 25 \), we look for alternative strategies to minimize (II.4).

To this aim, we shift to a game-theoretic interpretation of (II.1)–(II.3). In particular, we assume that each regular agent seeks to maximize the following utility (cognitive dissonance),

\[
u_i(x_i) = -\lambda_i (x_i - \theta_i)^2 - (1 - \lambda_i) \sum_{j \in V} P_{ij} (x_i - x_j)^2,
\]

where \( \lambda_i \in [0, 1] \) and \( P_{ij} \) weighs information exchange between \( i \) and \( j \), with \( P_{ij} = 0 \) if \( i \) and \( j \) do not communicate. In words, utility (III.1) makes the \( i \)th agent anchor to its prior proportionally to parameter \( \lambda_i \). Interestingly, for \( \lambda_i = 0 \), we
retrieve the utility function used in [46], where the authors analyze the classical consensus protocol from a game-theoretic perspective. In that case, agents had no incentive in retaining prior information, while the opposite is true with \((\text{III.1})\) as soon as \(\lambda > 0\). Greedily maximizing utility \((\text{III.1})\) at step \(k + 1\) yields the celebrated Friedrich-Johnsen (FJ) dynamics [47].

\[
x_i(k+1) = \lambda_i \theta_i + (1 - \lambda_i) \sum_{j \in V} W_{ij} x_j(k). \tag{\text{III.2}}
\]

where \(W_{ij}\) is obtained by normalizing \(P_{ij}\) w.r.t. weights of agent \(i\). For the sake of simplicity, we set \(\lambda_i \equiv \lambda\) for all agents in the following.

The intuition behind using \((\text{III.1}) - (\text{III.2})\) is that regular agents are likely to be closer to the nominal average consensus than values imposed by attackers, and thus they may prefer to act a bit selfishly rather than be misled by malicious neighbors. In the following, we refer to the FJ dynamics with \(\lambda = 0\) (standard consensus) as full collaboration, and to the case \(\lambda = 1\) as full competition. Tuning the parameter \(\lambda\) within the interval \([0, 1]\) originates a nontrivial competition-collaboration trade-off: should an agent fully collaborate, fully compete, or choose a hybrid strategy and trust neighbors only partially? Such performance trade-off in the presence of malicious agents will be the main matter of investigation in the rest of the paper.

**Assumption 4** (Network topology and weights). Weights \(W_{ij}\) in update \((\text{III.2})\) define an irreducible doubly-stochastic matrix \(W\). In view of the local optimization tasks, there are no self-loops, i.e., we set \(W_{ii} = 0\) \(\forall i \in V\).

**B. Full Competition vs. Full Collaboration**

A first remarkable result is that, in this scenario, letting \(\lambda = 1\) in \((\text{III.2})\) - which is equivalent to a totally unbalanced dynamics enforcing full competition among agents - may outperform the standard consensus protocol if noises are sufficiently intense.

**Proposition 1** (Full competition vs. full collaboration). In the presence of malicious agents, FJ dynamics \((\text{III.2})\) with \(\lambda = 1\) yields smaller error than with \(\lambda = 0\) if and only if

\[
\sum_{m \in M} d_m \geq \frac{M^2}{R} \text{Tr}(\Sigma_{11}) - \frac{2M^2}{R^2} B(\Sigma_{11}) + 2 \frac{M}{R} B(\Sigma_{12}) - B(\Sigma_{22}), \tag{\text{III.3}}
\]

where \(B(A)\) denotes summation of all elements of matrix \(A\).

**Sketch of proof:** The statement follows from manipulation of the consensus errors induced by the two considered protocol instantiations. Full derivation is reported in Appendix [III].

In words, Proposition [III] implies that, if malicious agents' noise variances \(d_m\) are sufficiently intense compared to the cross-correlations among regular and malicious agents (elements of \(\Sigma_{12}\)), a trivial fully competitive approach (agents keep priors constant overtime) yields better performance than the standard consensus protocol. Intuitively, the latter is still able to drive regular agents to a meaningful value when attacks are mild (small variance \(d_m\)), while the full-competitive strategy takes over as soon as attacks become sufficiently aggressive (large \(d_m\)) so that the drift error experienced by the consensus dynamics is larger than if agents just froze their priors.

In the case where priors and noises are i.i.d., we get the following simplified result.

**Corollary 1** ([48, Proposition 2]). If \(\Sigma = I\) and \(d_m \equiv d\), condition \((\text{III.3})\) becomes

\[
d > M \left(1 - \frac{2}{R}\right) - 1. \tag{\text{III.4}}
\]

**C. The Truth Lies in the Middle**

After assessing that competition-based approaches may be more resilient than standard consensus protocol, we now aim to characterize the "competitiveness" of the optimal strategy. In other words, we are interested in choosing \(\lambda\) so as to further reduce the consensus error. In particular, we aim to characterize analytically the optimal parameter, which we denote by

\[
\lambda^* = \arg \min_{\lambda} \varepsilon_R. \tag{\text{III.5}}
\]

Such optimal parameter always exists. Indeed, if \((1 - \lambda)W\) is Schur (i.e., \(\lambda > 0\)), the FJ dynamics induces the steady state

\[
x = L \hat{\theta}, \quad L = (I - (1 - \lambda)W)^{-1} \lambda. \tag{\text{III.6}}
\]

Matrix \(L\) can be interpreted as a generalization of the consensus matrix, and depends on both the update weights in \(W\) and the parameter \(\lambda\). In particular, \(L\) has a continuous extension at \(\lambda = 0\) given by \(\lim_{\lambda \to 0^+} L = \lim_{k \to +\infty} W^k\) [49], which in words means that, as \(\lambda\) gets close to zero, the steady-state achieved by the FJ dynamics actually tends to the one reached by the consensus protocol. Indeed, this is quite intuitive by looking at the update rule \((\text{III.2})\).

Clearly, the continuous extension of \(L\) also implies a corresponding continuous extension of \(\varepsilon_R\) at \(\lambda = 0\). Hence, extended continuity of \(\varepsilon_R\) in \([0, 1]\) and Weierstrass theorem ensure existence of a global minimum within the interval \([0, 1]\).

The next result characterizes the case when the optimal resilient strategy is nontrivial, meaning that regular agents should partially trust their neighbors in order to minimize error \((\text{III.4})\). Intuitively, this happens if some minimal interaction can do better than full competition (which translates into \(\Sigma\) being “non-degenerate” is some suitable sense), and if the attacks are sufficiently aggressive (i.e., variances of noises are large enough) so that the consensus protocol yields poor performance, akin to what remarked below Proposition [III].

**Theorem 1** (Nontrivial competition-collaboration trade-off). Let \(\Gamma = \lim_{\lambda \to 0^+} \frac{d\varepsilon_R}{d\lambda}\) with block partition

\[
\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & 0 \end{bmatrix}, \quad \Gamma_{11} \in \mathbb{R}^{R \times R}, \tag{\text{III.7}}
\]

and let \(C_R = \frac{\Gamma_{12} \Gamma_{21}}{R}, \quad C_{RM} = \frac{\Gamma_{12} \Gamma_{21}}{M}\). Then, the optimal parameter \(\lambda^*\) belongs to the open interval \((0, 1)\) if there exists one regular agent \(i \in \mathcal{R}\) such that \(\sigma_i^2 > \sigma_j^2 \forall j \neq i\) and

\[
\sum_{m \in M} \alpha_m d_m > \text{Tr} \left( -\Sigma_{11} \Gamma_{11}^T C_R - \Sigma_{12} \Gamma_{21}^T C_{RM} + \Sigma_{12} \Gamma_{21}^T C_{RM} + \Sigma_{22} \Gamma_{22}^T C_{RM} \right), \tag{\text{III.8}}
\]
where \( \alpha_m \geq 0 \) is the negative scalar product between the \( m \)th columns of \( \Gamma_2 \) and \( C_{RM} \).

**Sketch of proof:** Firstly, we show that \( \lambda^* < 1 \) under the first hypothesis. To do this, we compute the first derivative of \( e_{R} \) at \( \lambda = 1 \) and show that it is positive, hence the error function is strictly increasing in a left neighborhood of \( \lambda = 1 \). Secondly, we show that \( \lambda^* > 0 \) under the second hypothesis. To this aim, we compute the limit of the first derivative of \( e_{R} \) as \( \lambda \) vanishes and show that it is negative, hence the error function is strictly decreasing in a right neighborhood of \( \lambda = 0 \). The detailed calculations are reported in Appendix C.

**Corollary 2.** If \( \Sigma \) is diagonal, then \( \lambda^* \in (0, 1) \).

**Remark 3 (Explicit condition for \( \lambda^* > 0 \)).** Explicitly checking when condition (III.8) holds is hard, because it involves the spectrum of \( W \). However, our numerical tests show that indeed such a condition is always satisfied in meaningful cases.

**Remark 4 (Optimal parameter with zero noise).** Theorem 1 implies that \( \lambda^* \) may be positive even when noise variances are zero. This is actually consistent with high-level intuition: not only attackers steer regular agents far off from the nominal average consensus value (whereby \( d_m \) quantifies intensity of such deception), but also they behave against the prescribed update rule, ruling out full collaboration as an effective strategy – wunless cross-correlations between priors of regular and malicious agents are way larger than all other cross-correlations.

**Remark 5 (Optimal strategy with general matrices).** It is worth mentioning that, even though we assume no self-loops in the original matrix \( W \) to be consistent with the optimization tasks, Theorem 1 can be generalized to arbitrary doubly-stochastic matrices \( W \). Further, we will show via numerical experiments that all above result also holds for row-stochastic matrices \( W \).

**D. FJ Dynamics vs. Attack Aggressiveness**

We now study how performance of FJ dynamics varies with attack intensity, as quantified by the variances of noises \( v_m \).

We first show an intuitive result: more aggressive attacks (with larger noise variances) induce larger error for any \( \lambda^* \).

**Proposition 2 (Performance vs. attacks aggressiveness).** Error \( e_{R}(d_1, \ldots, d_M) \) is strictly increasing with \( d_m, m \in \mathcal{M} \).

**Proof:** See Appendix E

We next characterize what happens to the optimal parameter \( \lambda^* \). Intuitively, the more the nominal (prescribed) system behavior is disrupted by attacks, the more regular agents benefit from being competitive rather than collaborating with (potential) malicious neighbors. Formally speaking, this requires \( \lambda^* \) to grow proportionally to the noise intensities \( d_m \). However, such a claim is hard to prove analytically because of the structure of the cost function. In particular, studying its second derivative is complicated by the fact that the function \( e_{R} \) is expressed as the trace of a non-positive semidefinite matrix (in fact, not even symmetric), and similarly, uniqueness of the root of its first derivative cannot be proved, in general. In the face of such analytical difficulties, the next results contributes towards our intuition, which is numerically confirmed in Section IV.

**Proposition 3 (Optimal strategy vs. attack aggressiveness).** Let \( \lambda_{cr}(d_1, \ldots, d_M) \) a critical point of \( e_{R}(d_1, \ldots, d_M) \), then \( \lambda_{cr}(d_1, \ldots, d_M) \) is strictly increasing with \( d_m, m \in \mathcal{M} \).

**Proof:** See Appendix F

**Proposition 3** implies that all points of local minimum are strictly increasing with noise variances \( d_m \). An immediate consequence is that, if there is a unique critical point for one choice of \( \{d_m\}_{m \in \mathcal{M}} \), such a point is \( \lambda^* \), is unique for any choice of \( \{d_m\}_{m \in \mathcal{M}} \), and is strictly increasing with any \( d_m \). In words, more aggressive attacks force regular agents to progressively become more competitive, in order not to be deceived by malicious agents that can draw them away from nominal average consensus. The next proposition refines this result, describing the limit behavior with “extreme” attacks.

**Proposition 4 (Optimal strategy with extreme attacks).** Let \( \lambda_{cr}(d_1, \ldots, d_M) \) a critical point of \( e_{R}(d_1, \ldots, d_M) \), then \( \lim_{d_m \to +\infty} \lambda_{cr}(d_1, \ldots, d_M) = 1, m \in \mathcal{M} \).

**Proof:** See Appendix G

According to intuition, the (trivial) optimal strategy for regular agents is to fully compete against each other when adversarial attacks are too aggressive. However, numerical simulations in the next section show that \( \lambda^* \) is significantly smaller than 1 in most practical scenarios.

**IV. Numerical Experiments**

In this section, we perform numerical experiments on the consensus error \( e_{R} \) to achieve intuition about the behavior of FJ dynamics under different topologies and attack scenarios, and draw insight about effective choices of the parameter \( \lambda \).

In Fig. 2, we considered a 3-regular communication graph with 100 agents and (nominal) uniform weights \( W_{ij} = 1/3 \). Matrix \( \Sigma \) was chosen such that, for each agent \( i \), the cross-covariances obey an exponential decay, \( \sigma_{ij} = 10^{-0.2d_{(i,j)}} \), \( d_{(i,j)} \) being the length of a shortest path between \( i \) and \( j \), with \( \sigma_i^2 = 1 \). Further, we randomly selected one malicious agent and varied its noise intensity \( d \) within the range \([0, 100]\).

Figure 2a shows the error curve as \( d \) increases. All curves exhibit a unique point of minimum \( \lambda^* \), plotted in Fig. 2b. Further, both error curve and point of minimum are increasing.
with the noise intensity $d$, according to Propositions 2-3, showing that competition level needs to increase with $d$.

Figure 3 shows the same experiment but with a diagonal covariance matrix $\Sigma$. We observe the same monotonic behavior of $e_{\mathcal{R}}$ and $\lambda^*$. Further, we note that the error curve has a convex shape. In fact, even though it was not possible to prove it formally, all tests performed with diagonal covariance matrices resulted in strictly convex error functions.

We next studied what happens when increasing the number of malicious agents $M$. To better visualize changes in the behavior of the system, we fixed the set $\mathcal{R}$ to be a network composed of $R = 100$ regular agents, and added malicious agents across the network. Figure 4 shows the error curve when 10 such agents are progressively introduced. In particular, in this example, all malicious agents are selected so as to affect different portions of the network, which allows $\lambda^*$ to have relatively low values, see Fig. 4b. Conversely, we note that, in the opposite scenario, some regular agents may be forced to almost freeze their priors (large $\lambda$) to not drive the error too large. Figure 5 shows two cases where the added malicious agents are connected to the same regular agents. In particular, each consecutive couple is added to the neighborhood of one regular agent (e.g., the first two malicious agents added to the network are neighbors of agent $1 \in \mathcal{R}$). In this case, $\lambda^*$ increases faster than Fig. 4b, because the regular agents affected by multiple malicious need to keep their error small: in other words, they can hardly collaborate because of their misbehaving neighbors. We note that $\lambda^*$ grows faster when priors of regular agents are correlated (Fig. 5a), because such agents can trust that their states may be similar even before starting dynamical updates, and competing is less risky than collaborating.

Finally, it is interesting to notice that the error behavior observed above is also present when $W$ is only row stochastic, thus yielding nonzero consensus error even without external attacks. Figure 6 shows consensus error and $\lambda^*$ when each node in the graph has degree 3 or 4 and $W$ has uniform weights.

Many other numerical tests performed with different graphs, prior distributions, and choice of the malicious agents, show the same monotonic and quasi-convex behavior of the error function, and are omitted here in the interest of space. This reinforces and extends the scope of our formal analysis, showing that indeed the competition-collaboration trade-off emerges as a natural resilient mechanism in network systems.

**Remark 6 (Value of optimal $\lambda$).** A remarkable feature of the FJ dynamics, that emerges from the above numerical experiments, is that the optimal value of $\lambda$ is relatively small, within the range $[0.1, 0.2]$ for many relevant scenarios. In fact, $\lambda^*$ reaches 0.3 in Fig. 2 when the noise variance of the malicious agent is two order of magnitude larger than the variance of priors. This translates into the practical advantage that adding a little competition is sufficient to achieve substantial performance improvement compared to the standard consensus protocol, which may be attractive to achieve good level of resilience without forcing too conservative local agent updates.

### A. Competition-Collaboration Trade-off: Analytical Insight

An mentioned earlier, the consensus error function is hard to parse and an exhaustive analysis seems not possible.

Yet, some intuition can be achieved from an exact decomposition, which we analyze in this section. To keep notation light, we assume one malicious agent, i.e., $\mathcal{M} = \{m\}$, and...
diagonal covariance matrix \( \Sigma \). In this case, we can expand the consensus error as follows (cf. (C.1) in Appendix C).

\[
e_R = \left( \sigma_i^2 + d \right) \| L^{-m}_i \|^2 + \sum_{i \in R} \sigma_i^2 \| L^{-m}_i - \frac{1}{\| R \|} \|_2^2, \quad (IV.1)
\]

where \( L_i \in \mathbb{R}^N \) is the \( i \)th column of \( L \) and \( L^{-m}_i \in \mathbb{R}^R \) is obtained from \( L_i \) by removing its \( m \)th row (corresponding to the malicious agent). The error curves are shown in Fig. 7. Equation (IV.1) allows for an intuitive interpretation of the error, which leverages the notion of social power \([50], [51]\).

In opinion dynamics, the social power is used to quantify how much each agent’s opinion affects the opinion of all agents. In particular, when opinions evolve according to the FJ dynamics, the element \( L_{ij} \) quantifies the influence of agent \( j \) on agent \( i \) as \( L_{ij} \) increases, agent \( i \) is more affected by agent \( j \)’s initial opinion. The overall social power of agent \( j \) is a symmetric and increasing function of all elements \( \{ L_{ij} \}_{i \in V} \).

Borrowing such concepts from opinion dynamics allows to interpret the two contributions highlighted in (IV.1). The first contribution, \( e_{R, deception} \), quantifies the impact of the malicious agent on regular agents. The “social power” of \( m \), as quantified by the vector \( L^{-m}_m \), depends on the communication matrix \( W \) and on the parameter \( \lambda \). In particular, each coordinate of \( L^{-m}_m \) decreases with \( \lambda \), intuitively meaning that influence of the malicious agent diminishes as regular agents anchor more tightly to their priors, and becomes exactly zero when \( \lambda = 1 \), namely, when regular agents have no iterations with their neighbors (see Appendix F for formal analysis).

The second contribution, \( e_{R, consensus} \), measures “democracy” among regular agents, i.e., it is proportional to the mismatch between how much each regular agent affects the others and the ideal value \( 1/n \), which means that each agent affects all agents equally. This cost is zero if and only if the submatrix of \( L \) corresponding to the interactions among regular agent is the consensus matrix; this can happen only if regular agents do not interact with the malicious \([39]\), in which case, conversely, the vector \( L^{-m}_m \) is zero (i.e., attacks have no effect). In this special case, \( e_{R, consensus} \) is zero at \( \lambda = 0 \) and increases monotonically as the network shifts from a democratic system, where all agents fully collaborate (\( \lambda = 0 \)), to a disconnected system where agents fully compete (\( \lambda = 1 \)). In the case under consideration, when malicious agents affect regular ones, \( e_{R, consensus} \) exhibits a U-shape. For small \( \lambda \), the malicious agent rules the overall dynamics, and interactions among regular agents are negligible. As \( \lambda \) increases, the regular agents start collaborating and their interactions become closer to ideal democracy, making \( e_{R, consensus} \) decrease. However, as \( \lambda \) grows further, agents compete too aggressively, shifting away from the democratic system. In practice, numerical tests show that the point of minimum of \( e_{R, consensus} \) is small (see zoomed box in Fig. 7), which mean that the malicious agent barely affects the network error due to competition.

Overall, the error has two concurrent causes that generate a phase transition: the collaboration with the malicious agent is critical with small \( \lambda \), while for large \( \lambda \) the error is mainly due to agents competing against each other, rejecting possibly useful shared information. Indeed, this analysis matches intuition from (III.2), where \( \lambda \) measures conservatism in agent updates.

V. COMPARISON WITH EXISTING LITERATURE

In this section, we test our algorithm and compare its performance with other approaches found in literature.

Many techniques have been proposed to mitigate malicious attacks in optimization and consensus. However, they usually focus on reaching generic consensus, possibly while keeping regular agents’ states within a safe region (usually given by initial conditions), and do not consider robustness with respect to average consensus, which here is key to the distributed optimization task, as argued in Section II Indeed, most resilient consensus strategies aim to get the agents agree on, e.g., a common position (as robot rendez-vous) with some level of resilience, not necessarily tying consensus to initial conditions.

We compare two strategies: Weighted Mean Subsequence Reduced (W-MSR) \([25]\) and Secure Accepting and Broadcasting Algorithm (SABA) \([33]\). As noted in the introduction, many resilient algorithms consist in adaptations of W-MSR to various scenarios, and their core behavior and guarantees are the same. W-MSR suffers from two main limitations related to \( r \)-robustness, which is the cornerstone of all theoretical results. First, while sufficient conditions for resilient consensus are clear, often there is no clue about necessary conditions. This translates into unknown behavior when robustness requirements are not met. While \( r \)-robustness has proved a good characterization for this class of algorithms, such fact raises practical limitations. On the one hand, the communication network may be fixed. On the other hand, checking \( r \)-robustness is computationally intractable for large-scale graphs \([30]\). Thus, in some cases, for example with sparse architectures, more conservative behaviors with provable performance bounds may be preferred. Also, W-MSR requires to estimate the number of malicious agents affecting the network. This may also be an issue: if the estimate is low, regular agents may be deceived and average consensus disrupted, while, if it is high, the algorithm may be too conservative, possibly preventing convergence. Further, agent failures could happen in a time-varying fashion.

\[ \text{For example, [50], [51] use the arithmetic mean of } \{ L_{ij} \}_{i \in V} \text{.} \]
have degree three or four (Fig. 10a), with $W$ a row-stochastic matrix. In this case, one may question whether a doubly-stochastic matrix would improve performance of the standard consensus protocol, in light of its optimality under nominal conditions. However, in the presence of malicious agents, the consensus protocol converges to the centroid of their states (cf. Fig. 11) regardless of weights in $W$. Conversely, Fig. 10b shows that FJ dynamics is a robust strategy against misbehaving agents even though it cannot retrieve the optimal solution under nominal conditions.

**Remark 7 (Benefits of FJ dynamics).** The above experiments highlight some advantages of the proposed approach. Firstly, the presence of a tunable parameter makes the algorithm flexible, as it can smoothly adapt to different attack intensities while still providing decent performance bounds. Further, while the optimal parameterization requires exact knowledge of the adversary, which may not be reasonably assumed, yet our proposed approach proves pretty robust to the choice of a specific $\lambda$, as the plots in Section IV show. This also works with row-stochastic matrices, as shown in Fig. 10b, enabling simple weighting rules to be implemented locally. In contrast, in other approaches the cost function may be highly sensitive to some design parameters, e.g., the estimated number of malicious agents in W-MSR. Further, most results in literature do not describe system behavior when resilient consensus is not guaranteed. In fact, they usually either ensure that agent’s states remain inside the safety region (which in practice may not be better than setting $\lambda = 1$ in our approach), or let agents reach consensus but potentially be steered far away from initial conditions [56]. Finally, computational complexity and memory requirements are minimal, which may be desirable for resource-constrained devices or time-critical applications.

**VI. THE ROLE OF COMMUNICATION NETWORK**

In the previous sections, we discussed the benefits of using a competition-based approach (FJ dynamics) to tame malicious agents. We now shift attention to the communication network, in order to achieve intuition about the optimal topology. In Section VI-A, we introduce a second performance metric which we use to evaluate resilience to attacks. In Section VI-B, we observe how performance varies with connectivity.
A. Performance Metrics

Besides consensus error, we also aim to assess energy spent to conduct attacks. We assume again a single malicious agent, \( M = \{ m \} \). Let us define the following block partition of \( W \),
\[
W = \begin{bmatrix}
W_R & W_m \\
0 & 1
\end{bmatrix}, \quad W_R \in \mathbb{R}^{R \times R}, \tag{VI.1}
\]
where \( W_R \) corresponds to interactions among regular agents in \( R \) and \( W_m \) to neighbors of the malicious agent. Using (VI.1), we can rewrite (IX.2) as
\[
x_R(k + 1) = Ax_R(k) + Bx_m(k) + \vartheta \\
A = (1 - \lambda)W_R, \quad B = (1 - \lambda)W_m, \quad \vartheta = \lambda \theta_R. \tag{VI.2}
\]
We interpret (VI.2) as a controlled system where the state \( x_R \) stacks regular agents’ states, and the malicious agent commands the control input \( x_m(k) \). The controllability Gramian in \( K \) steps \( W_K \), defined as
\[
W_K = \sum_{k=0}^{K-1} A^k B B^T (A^T)^k, \tag{VI.3}
\]
can be used to quantify the control effort: indeed, the trace of \( W_K \) (controllability index) is inversely related to the control energy spent in \( K \) steps (averaged over the reachable subspace), as shown in literature [52]–[54]. Intuitively, a small controllability index means that the malicious agent consumes a lot of energy to steer \( x_R \) to some reachable configuration, which may be desired to drain out adversarial resources and possibly hamper the attack.

Notably, the controllability index can be equivalently written as
\[
\text{Tr} (W_K) = (1 - \lambda)^2 \sum_{k=0}^{K-1} \left\| (1 - \lambda)^k W_R^k W_m \right\|^2, \tag{VI.4}
\]
resembling the consensus error component \( e_{R, \text{deception}} \) in (IV.1),
\[
e_{R, \text{deception}} \propto \left\| \sum_{k=0}^{\infty} (1 - \lambda)^k \sum_{j=0}^{k-1} W_R^j W_m \right\|^2. \tag{VI.5}
\]
Both \( \text{Tr} (W_K) \) and \( e_{R, \text{deception}} \) are decreasing with \( \lambda \) (i.e., the more competition, the better), and depend on the vectors \( W_R^k W_m \) that describe how an attack spreads in \( k \) steps. The discount factor \( (1 - \lambda)^k \) makes the tail of the series in (VI.5) negligible, enhancing similarity between those two metrics.

Remark 8 (Power of adversary). The controllability Gramian (index) lets richer attack strategy than Assumption 3 allowing the malicious agent to drive regular agents to any configuration in the reachable subspace via a suitable input trajectory \( x_m \).

B. Network Connectivity vs. Resilience

We look at the following worst-case optimization problems,
\[
\min \max_{W \in \mathcal{V}} e_R, \quad \tag{VI.6}
\min \max_{W \in \mathcal{V}} \text{Tr} (W_K). \tag{VI.7}
\]
The internal maximization (worst case) selects the agent that either causes the largest consensus error (VI.6) or makes the smallest control effort (VI.7), \( K \) being the reachability index. The external minimization addresses the network design.

In the following, rather than solving (VI.6)–(VI.7), we examine performance of some simple, but significant, classes of graphs. In particular, we aim to achieve intuition about some core properties of the network, such as connectivity and degree balance. A broader optimization study is deferred to future work. As a first step, we restrict ourselves to regular graphs with 100 nodes and uniform weights. Note that regular graphs are commonly found in applications [55]–[59]. To assess the role played by connectivity, we evaluated the worst-case performance of \( \Delta \)-regular graphs, \( \Delta \in \{3, \ldots, 10\} \), by averaging \( \max_{m \in V} e_{R, \text{deception}} \) for (VI.6) and \( \max_{m \in V} \text{Tr} (W_K) \) for (VI.7) over 1000 random graphs. The “worst” malicious agent was found via brute force. Results are shown in Fig. 11.

A first insight is that increasing the graph connectivity mitigates attacks with respect to both metrics. Intuitively, this is because high degrees mean many interactions among regular agents that the malicious agent cannot control directly.

However, in real systems, the number of communication links is subject to practical constraints. To study how performance varies when the total amount of communication links is limited, we consider almost-regular graphs, namely, where nodes have degree either \( \Delta \) or \( \Delta - 1 \), for some \( \Delta \). Intuitively, this corresponds to “middle-ways” between \( \Delta \) and \( (\Delta - 1) \)-regular graphs, that, looking at Fig. 11, could be ideally placed between two consecutive ticks (degrees) \( \Delta \) and \( \Delta - 1 \) on the \( x \)-axis.

Hence, given a \( \Delta \)-regular graph as starting point, we progressively prune edges to observe performance variations. More specifically, we iteratively remove one edge at a time so as to minimize performance degradation at each removal. This corresponds to the following simplification of (VI.6)–(VI.7):
\[
\min \max_{e \in \mathcal{E}} e_{R, \text{deception}}(\mathcal{E}\setminus\{e\}), \tag{VI.8}
\min \max_{e \in \mathcal{E}} \text{Tr} (W_K(\mathcal{E}\setminus\{e\})), \tag{VI.9}
\]
where \( \mathcal{E} \) is the set of edges defining the communication network (i.e., the support of \( W \)), and \( W \) always has uniform weights (before and after removal of each edge). To always get almost-regular graphs, it suffices to impose that at most one edge be removed per node in minimization problems (VI.8)–(VI.9).

This additional constraint is also motivated by numerical tests showing that, in non-regular graphs, the “worst” malicious agent exploits highly connected agents to make more damage.\footnote{Note that an almost-regular graph implies a row-stochastic matrix \( W \).}
Fig. 12: Consensus error (left) and controllability index (right) with greedy edge removal for different topologies starting from a 4-regular graph, with \( \lambda = 0.7 \). In the plots, edge removal iterations (blue diamonds) proceed from right (initially, all 100 edges are present) towards left. At each iteration, one edge is removed so as to minimize performance degradation according to (VI.8)–(VI.9) while enforcing that no node has fewer than three neighbors (i.e., degree is either three or four for each node). At the last iteration (leftmost diamonds), most or all nodes have degree three, with possibly a few nodes left with degree four (because of the enforced degree-balance constraint). The red squares show the performance metrics for a 3-regular graph obtained by removing a perfect matching (set of edges) from the initial 4-regular graph.

Fig. 13: Consensus error (left) and controllability index (right) with greedy edge removal starting from a 4-regular graph (100 edges), with \( \lambda = 0.2 \).

Figures 12–13 show performances obtained starting from a 4-regular graph with 50 nodes (100 edges in total, corresponding to the rightmost point in the plots) and gradually pruning edges according to the above discussion (proceeding leftwards on the x-axis). Also, performances with a 3-regular graphs obtained by removing perfect matchings from the initial 4-regular graphs are shown for comparison. Remarkably, performance degrades (almost) monotonically for both indices as edges are removed. This may be explained by a combination of lower connectivity and degree unbalancedness, which allows the adversarial to exploit high-connected agents to make more effective damage against low-connected regular agents.

Interestingly, while the consensus error increases quite smoothly as more edges are removed, the controllability index exhibits sharp “jumps”. This is especially evident with large \( \lambda \), as Fig. 12 shows. Such behavior suggests the presence of critical subsets of edges, and may give indication about which links should be primarily kept or may be removed.

Further, in almost all tests (not shown here through space limitation), the 3-regular graph obtained by removing a perfect matching yielded a performance improvement compared to the last edge removal (leftmost point on blue curve). This shows that increasing connectivity may not be beneficial if it entails a loss in balance: in Fig. 12, the 3-regular graph reduces the controllability index by 22% w.r.t. the last graph obtained by greedily pruning edges (0.34 against 0.45), which has a single node with degree 4 and all others with degree 3, and has comparable performance with graphs having most nodes with degree 4. This suggests a phase transition in the network design, whereby it is not convenient to add edges until a certain degree balance is met. However, as shown in Fig. 13, a regular graph of degree \( \Delta - 1 \) obtained by removing a perfect matching (which is not related to performance metrics) from a \( \Delta \)-regular graph may yield much worse performance than even the more unbalanced graph obtained by greedy pruning edges. This gives further insight: removing edges arbitrarily may perform substantially worse compared to a performance-aware removal strategy.

VII. CONCLUSION AND FUTURE WORK

In this article, we have proposed a competition-based update protocol based on Friedkin-Johnsen dynamics to mitigate adversarial attacks disrupting a quadratic distributed optimization task. We have presented formal results and numerical experiments on performance and optimal parametrization, and showed that our approach can outperform state-of-the-art algorithms. Further, we have discussed the competition-collaboration trade-off with analytical arguments that are insightful in understanding the overall behavior of the system and the interaction among regular and malicious agents. Finally, we have addressed network design and studied how to improve performance with respect to regular graphs and link budget, looking at both global cost of the optimization problem and energy spent by the attacker.

This opens several avenues for future research. Firstly, it is desirable to address an effective design of parameters \( \lambda \)'s in the realistic case where knowledge about the attack is scarce. This may also involve online reweighing of protocol parameters, in the realm of recent work where the authors that build on the concept of trust [36], [38].

Secondly, the more general and challenging scenario of distributed optimization ought to be extensively studied. In this case, the standard approach is to alternate local descent steps to consensus updates to steer all agents towards a common solution [19]. Here, the additional descent steps may critically impact performance even if consensus steps are made resilient.

A third research avenue involves zero-sum games [60], [61] to alternatively model the system dynamics. In particular, asymmetric zero-sum games let one player have more knowledge than other, which may be a suitable model for worst-case attacks. In this case, a research challenge is determining the optimal game strategies for both players, ultimately to derive effective algorithms in the presence of intelligent adversaries.

Finally, it is interesting to deeply investigate optimization of communication topology. While graph robustness to node or edge failures has been addressed in various domains [62]–[65], the novel element given by competitive dynamics calls for studying that from a different perspective, as heuristically
motivated in Section VI. Also, in the spirit of a game-theoretic approach, a comparison between classical centrality measures and maximum-damage attacker nodes may be drawn to get insights about which agents deserve higher attention.

REFERENCES

[1] H. Farhangi, “The path of the smart grid,” IEEE Power Energy Mag., vol. 8, no. 1, pp. 18–28, 2010.
[2] F. Olivier, P. Aristidou, D. Ernst, and T. Van Cutsen, “Active management of low-voltage networks for mitigating overvoltages due to photovoltaic units,” IEEE Trans. Smart Grid, vol. 7, no. 2, pp. 926–936, 2016.
[3] X. Lu, P. Wang, D. Niyato, D. J. Kim, and Z. Han, “Wireless networks with RF energy harvesting: A contemporary survey,” IEEE Commun. Surveys Tuts., vol. 17, no. 2, pp. 757–789, 2015.
[4] Z. Niu, X. S. Shen, Q. Zhang, and Y. Tang, “Space-air-ground integrated vehicular network for connected and automated vehicles: Challenges and solutions,” Intelligent and Converged Networks, vol. 1, no. 2, pp. 142–169, 2020.
[5] R. Chen and C. G. Cassandras, “Optimal assignments in mobility-on-demand systems using event-driven receding horizon control,” IEEE Trans. Intell. Transp. Syst., pp. 1–15, 2020.
[6] S. Raskin. Energy secretary says enemies are capable of shutting down us power grid. [web]
[7] L. Borghese and S. Braithwaite. Hackers block italian covid-19 vaccination booking system in ‘most serious cyberattack ever’. [web]
[8] C. Huang, R. Zhang, and S. Cui, “Optimal power allocation for wireless sensor networks with outage constraint,” IEEE Trans. Commun. Lett., vol. 3, no. 2, pp. 209–212, 2014.
[9] Y. Ye, L. Shi, X. Chu, H. Zhang, and G. Lu, “On the outage performance of swipt-based three-step two-way df relay networks,” IEEE Trans. Veh. Technol., vol. 68, no. 3, pp. 3016–3021, 2019.
[10] S. Gupta, R. Kambli, S. Wagh, and F. Kazi, “Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework,” IEEE Trans. Ind. Electron., vol. 62, no. 4, pp. 2478–2486, 2015.
[11] J. Qi, J. Wang, and K. Sun, “Efficient estimation of component interactions for cascading failure analysis by em algorithm,” IEEE Trans. Power Syst., vol. 33, no. 3, pp. 3153–3161, 2018.
[12] M. Rahnamay-Naeini and M. M. Hayat, “Cascading failures in interdependent infrastructures: An interdependent markov-chain approach,” IEEE Trans. Smart Grid, vol. 7, no. 4, pp. 1997–2006, 2016.
[13] Z. Huang, C. Wang, M. Stojmenovic, and A. Nayak, “Characterization of cascading failures in interdependent cyber-physical systems,” IEEE Trans. Comput., vol. 64, no. 5, pp. 2158–2168, 2015.
[14] Q. Yan, F. R. Yu, Q. Gong, and J. Li, “Software-defined networking (sdn) and distributed denial of service (ddos) attacks in cloud computing environments: A survey, some research issues, and challenges,” IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 602–622, 2016.
[15] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, “Network intrusion detection for iot security based on learning techniques,” IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2671–2701, 2019.
[16] N. Agmon and D. Peleg, “Fault-tolerant gathering algorithms for autonomous mobile robots,” SIAM J. Comput., vol. 36, no. 1, pp. 56–82, 2006.
[17] I. D. Schizas, G. Mateos, and G. B. Giannakis, “Distributed lms for multi-agent systems in the presence of locally bounded faults,” Syst. Control. Lett., vol. 79, pp. 23–29, 2015.
[18] Y. Wang, H. Ishii, F. Bonnet, and X. Défago, “Resilient consensus against mobile malicious agents,” IFAC-PapersOnLine, vol. 53, no. 2, pp. 3409–3414, 2020, 21st IFAC World Congress.
[19] J. Usevitch and D. Panagou, “Resilient leader-follower consensus to arbitrary reference values in time-varying graphs,” IEEE Trans. Autom. Control, vol. 65, no. 4, pp. 1755–1762, 2020.
[20] Y. Shang, “Resilient consensus in multi-agent systems with state constraints,” Automatica, vol. 122, pp. 109288, 2020.
[21] S. Sundaram and B. Gharesifard, “Consensus-based distributed optimization with malicious nodes,” in 53rd Annu. Allerton Conf. Commun. Control Comput., Sep. 2015, pp. 244–249.
[22] L. Su and N. H. Vaidya, “Byzantine-resilient multiagent optimization,” IEEE Trans. Autom. Control, vol. 66, no. 5, pp. 2227–2233, 2021.
[23] S. Sundaram and C. N. Hadjicostis, “Distributed function calculation via linear iterative strategies in the presence of malicious agents,” IEEE Trans. Autom. Control, vol. 56, no. 7, pp. 1495–1508, 2011.
[24] S. M. Dibaji, M. Safi, and H. Ishii, “Resilient distributed averaging,” in American Control Conf., 2019, pp. 96–101.
[25] W. Abbas, A. Laszka, and X. Koutsoukos, “Improving network connectivity and robustness using trusted nodes with application to resilient consensus,” IEEE Control Netw. Syst., vol. 5, no. 4, pp. 2036–2048, 2018.
[26] Y. Zhai, Z.-W. Liu, M.-F. Ge, G. Wen, X. Yu, and Y. Qin, “Trust-enabled subsequence reduction for designing resilient consensus algorithms,” IEEE Trans. Netw. Sci. Eng., vol. 8, no. 1, pp. 259–268, 2021.
[27] J. S. Baras and X. Liu, “Trust is the cure to distributed consensus with adversaries,” in 27th Mediters. Conf. Control Autom. (MED), 2019, pp. 195–202.
[28] D. G. Mikulski, F. L. Lewis, E. Y. Gu, and G. R. Hudas, “Trust method for multi-agent consensus,” in Unmanned Systems Technology XIV, R. E. Karlsen, D. W. Gage, C. M. Shoemaker, and G. R. Gerhart, Eds., vol. 8387, International Society for Optics and Photonics. SPIE, 2012, pp. 146 – 159.
[29] 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.
[30] Z. Junhui, Y. Tao, G. Yi, W. Jiao, and F. Lei, “Power control algorithm of cognitive radio based on non-cooperative game theory,” China Communications, vol. 10, no. 11, pp. 143–154, 2013.
[31] J. Barreiro-Gomez, H. Tentbline, L. Stella, D. Bauso, and P. Colaneri, “Risk-aware control and games in engineering,” in Proc. IEEE Conf. Decis. Control, 2020, pp. 3860–3870.
[32] F. J. Muros, J. M. Maestre, E. Alpga, T. Alamo, and E. F. Camacho, “Networked control design for coalitional schemes using game-theoretic methods,” Automatica, vol. 78, pp. 320–332, 2017.
[33] D. Monatte, G. Como, and F. Fagnani, “Systemic risk and network intervention,” IFAC-PapersOnLine, vol. 53, no. 2, pp. 2856–2861, 2020, 21st IFAC World Congress.
[34] J. R. Marden and J. S. Shamma, “Chapter 16 - game theory and distributed control,” in Handbook of Game Theory with Economic Applications, H. P. Young and S. Zamir, Eds. Elsevier, 2015, vol. 4, pp. 861–899.
[35] N. Li and J. R. Marden, “Designing games for distributed optimization,” IEEE J. Sel. Topics Signal Process., vol. 7, no. 2, pp. 230–242, 2013.
[36] A. V. Proskurnikov and R. Tempo, “A tutorial on modeling and analysis of dynamic social networks. part i,” Annual Reviews in Control, vol. 43, pp. 65–79, 2017.
[37] J. R. Marden, G. Arslan, and J. S. Shamma, “Cooperative control and potential games,” IEEE Trans. Syst., Man, Cybern., Part B (Cybern.), vol. 39, no. 6, pp. 1393–1407, 2009.
[38] N. E. Friedkin and E. C. Johnsen, “Social influence and opinions,” J. Math. Sociol., vol. 15, no. 3-4, pp. 193–206, 1990.
Lemma A.2. Let $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{n \times n}$ invertible and differentiable function of $\alpha$, then the derivative of $A^{-1}$ is
\[
\frac{dA^{-1}}{d\alpha} = -A^{-1} \frac{dA}{d\alpha} A^{-1}.
\] (A.2)

Lemma A.3. Let $A \in \mathbb{R}^{n \times n}$ invertible with eigenvalue-eigenvector couple $(\lambda, v)$, then $A^{-1}$ has eigenvalue-eigenvector couple $(\lambda^{-1}, v)$.

Corollary A.1. If $A \in \mathbb{R}^{n \times n}$ is diagonalizable, then $A$ and $A^{-1}$ are simultaneously diagonalizable.

Lemma A.4. Let $A \in \mathbb{R}^{n \times n}$ with eigenvalue-eigenvector couple $(\lambda, v)$, then $(I - \alpha A)$ has eigenvalue-eigenvector couple $((1 - \alpha)\lambda, v)$.

B. Proof of Proposition 4

We first compute the consensus error induced by (??). In virtue of Assumption 3 the steady-state consensus value induced by $W$ is the average of malicious agents’ priors, $\bar{\theta}_M = \frac{1}{M} \sum_{\theta \in M} \theta$, $M = |M|$. The consensus error becomes
\[
e^C = E \left( \left\| I \frac{1}{M} \bar{\theta} - I \frac{1}{R} \bar{\theta} \right\|^2 \right)
= E \left( \left\| I \frac{1}{M} \bar{\theta} \right\|^2 \right) + E \left( \left\| I \frac{1}{R} \bar{\theta} \right\|^2 \right) - 2E \left( \frac{\bar{\theta}^\top}{R} I \frac{1}{M} \bar{\theta} \right)
= R^2 \text{Tr} \left( \frac{1}{M} \Sigma_1 \frac{1}{M} \right) + \text{Tr} \left( \frac{1}{R} \Sigma_1 \frac{1}{M} \right) - 2 \text{Tr} \left( \frac{1}{M} \Sigma_1 \lambda M \right)
= \frac{R}{M^2} \sum_{m \in M} d_m + \frac{R}{M^2} \sum_{m \in M} \left( \sum_{\theta \in M} \sigma_{m}^2 + \sum_{\theta \not\in M} \sigma_{m}^2 \right)
+ \frac{1}{R} \sum_{i \in R} \left( \sum_{j \in R} \sigma_i^2 + \sum_{j \not\in i} \sigma_i \right) - \frac{2}{M} \sum_{i \in R} \sum_{m \in M} \sigma_{i,m}
\] (B.1)

where $1_M \in \mathbb{R}^N$ and $1_R \in \mathbb{R}^N$ are the indicator vectors of sets $M$ and $R$, respectively. On the other hand, the FJ dynamics with $\lambda = 1$ simply freezes all regular agents’ priors, yielding consensus error
\[
e^C_{FJ} = E \left( \left\| \theta_R - C_R \theta_R \right\|^2 \right)
= \text{Tr} \left( E \left( \theta_R \theta_R^\top \right) (I - C_R) \right)
= \frac{R^2 - 1}{R} \sum_{i \in R} \sigma_{i}^2 - \frac{1}{R} \sum_{i \in R} \sum_{j \not\in i} \sigma_{i,j},
\] (B.2)

which depends only on the nominal covariance matrix of priors $\Sigma$. By comparing the final expressions in (B.1) and (B.2), it
follows that $e^C > e^J_1$ is equivalent to the following inequality,

$$
\sum_{m \in M} d_m - \sum_{m \in M} \left( \sigma_m^2 + \sum_{n \in M, n \neq m} \sigma_{mn} \right) + \frac{M^2}{R} \sum_{i \in R} \sigma_i^2 \\
- \frac{2M^2}{R^2} \sum_{i \in R} \left( \sigma_i^2 + \sum_{j \in R} \sigma_{ij} \right) + \frac{2M}{R} \sum_{i \in R} \sum_{m \in M} \sigma_{im},
$$

which leads to condition (III.3).

### C. Proof of Theorem 1

#### Part One: $\lambda^* < 1$

Let us define the following matrices: $S_R \in \mathbb{R}^{R \times N}$ maps $x$ to $x_{R^*}$, and $C_R \equiv \frac{1}{M} \mathbf{1}_R \mathbf{1}_R^\top$. According to the labeling discussed in Section II-A, we set $S_R = [I_R | 0]$. Then, the error (II.4) can be written as

$$
e_{R^*} = \text{Tr} \left( \Sigma E^{\top} E \right), \quad E = S_R L - C_R S_R,
$$

and its derivative with respect to $\lambda$ is (up to constants)

$$
\frac{\text{d}e_{R^*}}{\text{d}\lambda} = \frac{1}{\lambda} \text{Tr} \left( \Sigma L^\top (I - W^\top L^\top) S_R^\top E \right),
$$

where Lemmas A.1-A.2 were used, and at $\lambda = 1$ takes value

$$
\frac{\text{d}e_{R^*}}{\text{d}\lambda} |_{\lambda = 1} = \text{Tr} \left( \Sigma (I - W^\top) S_R^\top (S_R - C_R S_R) \right).
$$

Straightforward computations show that the matrix argument of the trace in (C.5) takes the form

$$
\begin{bmatrix}
A & 0 \\
0 & 0
\end{bmatrix}, \quad A \in \mathbb{R}^{R \times R},
$$

and the $i$th diagonal element of $A$, associated with $i \in R$, is

$$
a_i = \sigma_i^2 + \frac{1}{R} \sum_{m \in M} \sigma_{im} \left( 1 - \sum_{m' \in M, m' \neq m} W_m^{o} \right) - \frac{1}{R} \sum_{m \in M} \sigma_{im} \sum_{m \in M} W_m^{o} \\
- \sum_{j \in R, j \neq i} \sigma_{ij} W_{ij} - \frac{1}{R} \sum_{j \in R, j \neq i} \sum_{m \in M} W_m^{o} - \sum_{m \in M} \sigma_{im} W_m^{o},
$$

where $W_{ij}^{o}$ is the weight of the directed edge from $j$ to $i$ in the original weight matrix $W^{o}$ (without malicious agents), and $W_{mj} = \delta_{mj}$ according to Assumption 3 where $\delta_{mj}$ is the Kronecker delta. Further, $W_{ij}^{o} = W_{ij}$ for $i, j \in R$. Being $W^{o}$ doubly stochastic, it holds

$$
\frac{1}{R} \sum_{m \in M} L_{mj} W_m^{o} = \sigma_{im} W_m^{o},
$$

which leads to condition (III.3).

#### Part Two: $\lambda^* > 0$

In virtue of continuity of the error derivative in $L$, we can compute the limit of (C.2) as

$$
\lim_{\lambda \to 0^+} \frac{\text{d}e_{R^*}}{\text{d}\lambda} = \text{Tr} \left( \Sigma \frac{\text{d}L^\top}{\text{d}\lambda} S_R^\top \frac{\text{d}L}{\text{d}\lambda} E \right)
$$

where the steady-state consensus matrix $\bar{W} = \lim_{\lambda \to 0^+} W$ has block partition (cf. Assumption 5 for the value of $\bar{W}$)

$$
\bar{W} = \begin{bmatrix}
C_{R_{11}} & 0 \\
0 & C_{R_{22}}
\end{bmatrix},
$$

where $\Gamma_1, \Sigma_{11} \in \mathbb{R}^{R \times R}$ and $\Gamma_2, \Sigma_{22} \in \mathbb{R}^{M \times M}$. Matrix $\Gamma$ can be computed exactly from the spectral decomposition of $W$. In particular, its elements are finite, $\Gamma_1$ is nonnegative, and $\Gamma_2$ is nonpositive (see detailed characterization in Appendix D). Hence, limit (C.9) is negative if and only if the following inequality holds,

$$
\text{Tr} \left( V_{M} (\Gamma_2^\top C_{R_{22}}) \right) > \text{Tr} \left( (-\Sigma_{11} \Gamma_1^\top C_R - \Sigma_{12} \Gamma_2^\top C_R + \Sigma_{21} \Gamma_1^\top C_{R_{11}} + \Sigma_{22} \Gamma_2^\top C_{R_{22}}) \right),
$$

which leads to condition (III.3). It follows that, if (C.11) holds, $e_{R^*}$ is strictly decreasing in a right neighborhood of $\lambda = 0$. By virtue of continuity, we conclude that $\lambda^* > 0$.

### D. Computation of Matrix $\Gamma$

In this appendix, we show how matrix $\Gamma$ can be derived from $W$ and discuss the sign of its elements. For the sake of simplicity, we consider the case when the original weight matrix $W$ (i.e., with weights not corrupted by malicious agents) is symmetric, which immediately implies that $W$ is diagonalizable also when malicious agents modify their weights. In the case
that $W$ is not diagonalizable, a similar derivation, but more with involved calculations, can be carried out by considering its Jordan canonical form. This is because a straightforward extension of Lemma \ref{lem:Jordan} shows that $W$ and $\Gamma$ share the same (chain of) generalized eigenvectors.

**Computation of $\Gamma$.** The derivative of $L$ is (Lemma \ref{lem:derivative_L})

$$
\frac{dL}{d\lambda} = \tilde{L} - \lambda \tilde{L}^{-1} \bar{L} = \tilde{L} - \lambda \tilde{L} \bar{L}_W \bar{L},
$$

where $\tilde{L} \equiv (I - (1 - \lambda)W)^{-1}$. Let $\lambda_W$ and $v_W$ an eigenvalue of $W$ and its associated eigenvector, respectively, from Lemmas \ref{lem:Jordan}, \ref{lem:Jordan2}, \ref{lem:Jordan3}; it follows that $\tilde{L}$ has eigenvalue $(1 - (1 - \lambda)\lambda_W)$ with associated eigenvector $v_W$. Hence, straightforward computations yield

$$
\frac{dL}{d\lambda} v_W = \frac{1}{1 - (1 - \lambda)\lambda_W} \lambda_W v_W.
$$

In particular, the dominant eigenvector $v_W = \mathbb{I}$ (associated with $\lambda_W = 1$) is in the kernel of $dL/d\lambda$ for any $\lambda$. As for the eigenvectors, by letting $\lambda$ go to zero in (D.2), one gets

$$
\Gamma v_W = (1 - \lambda W)^{-1} v_W.
$$

Finally, the eigendecomposition of $\Gamma$ is obtained from eigenvectors $v_W$ and eigenvalues $(1 - \lambda W)^{-1}$, plus the kernel.

**Sign of $\Gamma_1$ and $\Gamma_2$.** As regards $\Gamma_1$, note that the upper-left block in $\bar{W}$ is identically zero, and that $L$ is a stochastic matrix for any value of $\lambda$: hence, as $\lambda$ becomes larger than zero, (some) elements in $L_1$ become positive, and thus their derivative at $\lambda = 0^+$ is also positive.

As for $\Gamma_2$, define the following block partitions,

$$
L = \begin{bmatrix} L_1 & L_2 \\ 0 & I_M \end{bmatrix},
$$

with $W_1, L_1 \in \mathbb{R}^{R \times R}$ and $W_2, L_2 \in \mathbb{R}^{R \times M}$. Then, it holds

$$
\frac{dL}{d\lambda} = \frac{1}{\lambda} L (I - WL) = \begin{bmatrix} * & -L_1 W_1 L_2 - L_1 \\ 0 & 0 \end{bmatrix},
$$

which implies, for any $\lambda \in (0, 1)$,

$$
\frac{dL_{im}}{d\lambda} \leq 0, \quad i \in \mathcal{R}, m \in \mathcal{M}.
$$

In particular, the limit of the derivative of element $L_{im}$ at $\lambda = 0^+$ is nonpositive in virtue of the theorem of sign permanence.

**E. Proof of Proposition \ref{prop:critical_points}**

Computing the partial derivative of $e_R(d_1, \ldots, d_M)$ (C.1) yields

$$
\frac{\partial e_R(d_1, \ldots, d_M)}{\partial d_m} = \text{Tr} \left( \begin{bmatrix} 0 & 0 \\ 0 & S_m \end{bmatrix} E^T E \right),
$$

where

$$
\frac{\partial \Sigma(d_1, \ldots, d_M)}{\partial d_m} = \begin{bmatrix} 0 & 0 \\ 0 & S_m \end{bmatrix}
$$

and $S_m \in \mathbb{R}^{M \times M}$ has all zero elements except for the $m$th diagonal element equal to 1. Hence, the argument of the trace in (E.1) has all zero rows except for the $(R + m)$th row, which equals the $(R + m)$th row of

$$
\begin{bmatrix}
(L_1 - C_R)^2 & (L_1 - C_R) L_2 \\
L_2 (L_1 - C_R) & L_2 L_2
\end{bmatrix}.
$$

The trace then selects the $m$th diagonal element of $L_2^T L_2$, which has all positive elements (see \ref{lem:positive} and discussion in Section IV-A). Hence, it follows that the partial derivative in (E.1) is strictly positive for any $m \in \mathcal{M}$.

**F. Proof of Proposition \ref{prop:critical_points}**

We start by computing the partial derivative of the error, first with respect to $\lambda$ and then with respect to $d_m$:

$$
\frac{\partial^2 e_R(\lambda, d_1, \ldots, d_M)}{\partial d_m \partial \lambda} = \frac{1}{\lambda} \text{Tr} \left( L \frac{\partial \Sigma(d_1, \ldots, d_M)}{\partial d_m} S^{\top} R S_R \right).
$$

It holds

$$
L \frac{\partial \Sigma(d_1, \ldots, d_M)}{\partial d_m} = \begin{bmatrix} 0 & L_2 S_m \\ 0 & S_m \end{bmatrix}
$$

and the argument of the trace in (F.1) is

$$
\begin{bmatrix}
-L_2 S_m L_2^T W_1^T L_1 - L_2 S_m W_2^T L_1^T \\
* & 0
\end{bmatrix},
$$

whose upper-left block is a negative matrix for all $\lambda \in (0, 1)$, and is the zero matrix for $\lambda = 1$. Hence, the error derivative with respect to $\lambda$ (C.2) is strictly decreasing with $d_m$ for any $\lambda \in (0, 1)$, and does not depend on $d_m$ at $\lambda = 1$. In virtue of continuity of (C.2) in $\lambda$, we conclude that the critical points of $e_R$ are strictly increasing with $d_m$.

**G. Proof of Proposition \ref{prop:critical_points}**

We expand (C.2) to highlight dependence on $d_m$. First, we note that

$$
\frac{de_R}{d\lambda} = \frac{1}{\lambda} \text{Tr} \left( V_{\mathcal{M}} N + k(\lambda, W, \Sigma) \right),
$$

where $N$ is a nonpositive matrix defined as

$$
N \doteq -(L_2^T W_1^T + W_2^T) L_1^T L_2,
$$

and $k(\lambda, W, \Sigma) > 0$ does not depend on any $d_m$, $m \in \mathcal{M}$. Then, we have

$$
\frac{de_R}{d\lambda} = \frac{1}{\lambda} \sum_{m \in \mathcal{M}} N_{mm} d_m + k(\lambda, W, \Sigma).
$$

Note that $N_{mm} \neq 0$, because the opposite implies that the $m$th malicious agent has no (even indirect) interactions with regular agents. It follows that, for any $m \in \mathcal{M}$, there always exists $d_m > 0$ such that the error derivative (G.1) is negative, for any $\lambda < 1$. In fact, given $\lambda$, the minimal such value of $d_m$
can be obtained from the following inequality,

\[ d_m > -\frac{\lambda}{N_{mm}} k(\lambda, W, \Sigma) - \sum_{m' \in \mathcal{M}} \frac{N_{m'm'}}{N_{mm}} d_{m'} > 0. \quad \text{(G.4)} \]

The claim follows by combining (G.4) with Proposition 3.