A Risk-Security Tradeoff in Graphical Coordination Games

Keith Paarporn, Mahnoosh Alizadeh, and Jason R. Marden

***Abstract*—A system relying on the collective behavior of decision makers can be vulnerable to a variety of attacks. How well can a system operator protect performance in the face of these risks? We frame this question in the context of graphical coordination games, where the agents in a network choose between two conventions and derive benefits from coordinating neighbors, and system performance is measured in terms of the agents' welfare. In this article, we assess an operator's ability to mitigate two types of attacks: 1) broad attacks, where the adversary distributes targeted incentives to all agents in the network; and 2) focused attacks, where the adversary can force a selected subset of the agents to commit to a prescribed convention. The system operator can among a class of distributed algorithms that define the local interactions. Our main contribution characterizes the operator's fundamental tradeoff between security against worst-case broad attacks and vulnerability from focused attacks. We show that this tradeoff significantly improves when the operator selects a decision-making process at random. This article highlights the design challenges a system operator faces in maintaining resilience of networked distributed systems.**

***Index Terms*—Game theory, networked control systems, network security.**

I. Introduction

NETWORKED distributed systems typically operate without centralized planning or control, and instead rely on local interactions and communication among comprising agents. These systems arise in a variety of engineering applications such as teams of mobile robots and sensor networks [2]–[4]. They are also prevalent in social dynamics [5], [6] and biological populations [7].

The distributed nature of these systems may leave the individual agents vulnerable to adversarial manipulations. Such perturbations can potentially lead to unwanted outcomes. For example in social networks, individual opinions can be shaped from external information sources, resulting in a polarized society [8], [9]. When feasible, a system operator takes measures to mitigate adversarial influences. The literature on cyber-physical system security studies many aspects of this interplay. For instance, optimal controllers are designed to mitigate denial-of-service, estimation, and deception attacks [10]–[14].

This article investigates measures a system operator can take to mitigate adversarial influences when the underlying system is a graphical coordination game [6], [15]. Agents in a network decide between two choices x or y. One may think of these choices as two competing products, e.g., iPhone versus Android, two conflicting social norms, or two opposing political parties. Each agent derives a positive benefit from interactions with coordinating neighbors, and zero benefits from miscoordinating ones. A measure of system efficiency is the ratio of total benefits of all agents to the maximal attainable benefits over all configurations of choices.

In the absence of adversaries, the goal of a system operator is to design local decision-making algorithms for the agents so that the emergent collective behavior optimizes system efficiency. One algorithm that achieves this goal in graphical coordination games is log-linear learning [16]–[18]. More formally, the agents follow a "perturbed" best reply dynamics where the agents' local objectives are precisely equal to their local welfare. We seek to address the question of whether this particular algorithm is robust to adversarial influences. That is, can an adversary manipulate their decision-making processes to render the system suboptimal? If so, how can the operator modify the base algorithms to mitigate such attacks?

We consider two adversarial attack models—*broad* and *focused* attacks. In broad attacks, the adversary attaches impostor nodes to every agent in the network (hence, broad), influencing their decision-making process. This could depict distributing political ads with the intention of polarizing voters. In focused attacks, the adversary targets a specific set of agents in the network, forcing them to commit to x or y. These targeted fixed agents consequently do not update their choices over time but still influence the decisions of others. They could portray loyal consumers of a brand or product, or staunch supporters of a political party. Fixed agents and their effects on system performance have been extensively studied in the context of opinion dynamics and optimization algorithms [14], [19], [20].

This article concerns the consequences a system operator faces in making modifications to the local algorithms. In particular, we answer the question "if the operator succeeds in

Manuscript received January 10, 2020; revised January 17, 2020; accepted June 7, 2020. Date of publication June 15, 2020; date of current version April 26, 2021. This work was supported in part by the University of California under Grant LFR-18-548175, in part by the Office of Naval Research under Grant #N00014-17-1-2060, and in part by the National Science Foundation under Grant #ECCS-1638214. This paper was presented in part at the 58th IEEE Conference on Decision and Control, Nice, France, Dec. 1919. Recommended by Associate Editor R. M. Jungers. *(Corresponding author: Keith Paarporn.)*

The authors are with the Department of Electrical and Computer Engineering, the University of California, Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: kpaarporn@gmail.com; alizadeh@ece.ucsb.edu; jrmarden@ece.ucsb.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2020.3002499

protecting the system from one type of attack, how vulnerable does it leave the system to the other?" We identify the fundamental tradeoffs between performance guarantees against broad attacks and focused attacks.

Specifically, we provide a characterization of worst-case risk metrics from both types of adversarial attacks as a function of the operator's algorithm design parameter (see Section III). We define risk in this article as the system's distance to optimal efficiency. By worst-case, here, we mean the maximum risk among all possible connected network topologies subject to any admissible adversarial attack (see Section V). These results appeared in an earlier version of this article [1] for deterministic design parameters. The present manuscript extends the analysis to randomized operator designs (see Sections IV and VI). In particular, randomized operator strategies significantly improve the set of attainable risk levels and their associated optimal tradeoffs (see Section IV).

This article offers a system operator's perspective of ensuring security against adversarial influences. This is in contrast to a body of previous works that studied system vulnerabilities from an adversarial perspective [21], [22]. These works analyze the extent to which system efficiency can be degraded as a function of the adversary's strength, intelligence, and sophistication. However, they do not consider any mitigation strategies a system operator could implement. Furthermore, these results restricted attention to specific graph structures, namely ring graphs, whereas the analysis in this article considers adversarial influence on any graph topology.

II. Preliminaries

A. Graphical Coordination Games

A graphical coordination game is played between a set of agents $\mathcal{N} = \{1, \ldots, N\}$ over a connected undirected network $G = (\mathcal{N}, \mathcal{E})$ with node set $\mathcal{N}$ and edge set $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$. Agent i's set of neighbors is written as $\mathcal{N}_i = \{j : (i,j) \in \mathcal{E}\}$. Each agent i selects a choice a_i from its action set $\mathcal{A}_i = \{x, y\}$. The choices of all the agents constitutes an action profile $a = (a_1, \ldots, a_N)$, and we denote the set of all action profiles as $\mathcal{A} = \Pi_{i=1}^N \mathcal{A}_i$. The local interaction between two agents $(i,j) \in \mathcal{E}$ is based on a 2×2 matrix game, described by the payoff matrix $V : \{x,y\}^2 \to \mathbb{R}$

$$\begin{array}{ccc|c|c|} & & \multicolumn{2}{c}{\text{Player } j} \\ & & \multicolumn{1}{c}{x} & \multicolumn{1}{c}{y} \\ \cline{3-4} \text{Player } i & x & 1+\alpha_{\text{sys}}, 1+\alpha_{\text{sys}} & 0,0 \\ \cline{3-4} & y & 0,0 & 1,1 \\ \cline{3-4} \end{array} \tag{1}$$

where $\alpha_{\text{sys}} > 0$ is the system *payoff gain*. It indicates that x is an inherently superior product over y when users coordinate. Here, agents would rather coordinate than not, but prefer to coordinate on x. Agent i's *benefit* is the sum of payoffs derived from playing the game (1) with each of its network neighbors

$$W_i(a_i, a_{-i}) := \sum_{j \in \mathcal{N}_i} V(a_i, a_j). \tag{2}$$

A measure of system *welfare* defined over $\mathcal{A}$ is

$$W(a) := \sum_{i=1}^{N} W_i(a_i, a_{-i}) \tag{3}$$

which is simply the sum of all agent benefits. The *system efficiency* for action profile $a \in \mathcal{A}$ is defined as

$$\frac{W(a)}{\max_{a' \in \mathcal{A}} W(a')}. \tag{4}$$

For $\mathcal{A} = \{x,y\}^N$, the all-x profile $\vec{x}$ maximizes welfare. This does not necessarily hold for arbitrary action spaces.

B. Log-Linear Learning Algorithm

Log-linear learning is a distributed stochastic algorithm governing how players' decisions evolve over time [15]–[17]. It may be applied to any instance of a game with each player having a well-defined local utility function $U_i : \mathcal{A} \to \mathbb{R}$ over a set of action profiles $\mathcal{A}$ with an underlying interaction graph G. That is, agent i's local utility is a function of its action a_i and actions of its neighbors in G.

Agents update their decisions $a(t) \in \mathcal{A}$ over discrete time steps $t = 0, 1, \ldots$. Assume $a(0)$ is arbitrarily determined. For step $t \geq 1$, one agent i is selected uniformly at random from the population. It updates its action to $a_i(t) = z \in \mathcal{A}_i$ with probability

$$\frac{\exp(\beta U_i(z, a_{-i}(t-1)))}{\sum_{z' \in \mathcal{A}_i} \exp(\beta U_i(z', a_{-i}(t-1)))} \tag{5}$$

where $\beta > 0$ is the rationality parameter. All other agents repeat their previous actions: $a_{-i}(t) = a_{-i}(t-1)$. For large values of β, i selects a best response to the previous actions of others with high probability, and for values of β near zero, i randomizes among its actions $\mathcal{A}_i$ uniformly at random. This induces an irreducible Markov chain over the action space $\mathcal{A}$, with a unique stationary distribution $\pi_\beta \in \Delta(\mathcal{A})$. The *stochastically stable states* (SSS) $a \in \mathcal{A}$ are the action profiles contained in the support of the stationary distribution in the high rationality limit: they satisfy $\pi(a) = \lim_{\beta \to \infty} \pi_\beta(a) > 0$. Such a limiting distribution exists and is unique [16], [23], [24]. We write the set of stochastically stable states as

$$\text{LLL}(\mathcal{A}, \{U_i\}_{i \in \mathcal{N}}; G). \tag{6}$$

For graphical coordination games, the log-linear learning algorithm specified by the action set $\mathcal{A} = \{x,y\}^N$ and utilities $\{W_i\}_{i \in \mathcal{N}}$ selects the welfare-maximizing profile $\vec{x}$ as the stochastically stable state irrespective of the graph topology G. This can be shown using potential game arguments [17]. That is, $\vec{x} = \text{LLL}(\mathcal{A}, \{W_i\}_{i \in \mathcal{N}}; G)$ for all $G \in \mathcal{G}_N$, where $\mathcal{G}_N$ is the set of all connected undirected graphs on N nodes.

C. System Operator

If the agents' local decision-making rules can be manipulated, optimality of the log-linear learning algorithm may no longer hold true. We consider a system operator able to alter the agents' local utility functions with the goal of mitigating the loss of system efficiency from adversarial manipulations. Such modifications to distributed algorithms could represent updates to software features, for example.

In particular, we consider the class of local utility functions $\{U_i^\alpha\}_{i\in\mathcal{N}}$ parameterized by a payoff gain $\alpha > 0$. They take the same form as the benefit function (2), except the system operator chooses a *perceived gain* α to replace α_{sys}. These local utilities change the underlying game interactions between agents, but do not change the *experienced* benefits of coordination, i.e., these are still given by (1). As such, the system's welfare and efficiency (3), (4) remain defined in terms of the system payoff gain α_{sys}.

III. Models of Adversarial Influence

In this section, we outline two models of attacks in graphical coordination games—*broad* and *focused* attacks. The system operator specifies the local utility functions $\{U_i^\alpha\}$ that govern the log-linear learning algorithm by selecting the perceived payoff gain $\alpha > 0$. Our goal is to assess the performance of this range of algorithms on two corresponding worst-case risk metrics, which we will define. We, then, identify fundamental tradeoff relations between mitigating these two risk metrics.

A. Broad Attacks and Worst-Case Risk Metric

We consider a scenario where the system is subject to broad attacks. For each agent in the network, the adversary attaches a single impostor node that acts as a neighbor that always plays x or y. These nodes are not members of the network but affect the decision making of agents that are. Let $S_x \subseteq \mathcal{N}$ (S_y) be the set of agents targeted with an impostor x (y) node. We call the *target set* $S = (S_x, S_y)$. Any target set satisfies $S_x \cap S_y = \varnothing$ and $S_x \cup S_y = \mathcal{N}$. We call $\mathcal{T}(G)$ the set of all possible target sets S on the graph G. Given $\alpha > 0$, the agents' *perceived* utilities are

$$\tilde{U}_i^\alpha(a_i, a_{-i}) := \begin{cases} U_i^\alpha(a_i, a_{-i}) + \mathbb{1}(a_i = y) & i \in S_y \\ U_i^\alpha(a_i, a_{-i}) + (1+\alpha)\mathbb{1}(a_i = x) & i \in S_x \end{cases}. \tag{7}$$

In the notation of (6), the set of stochastically stable states is written $\text{LLL}(\mathcal{A}, \{\tilde{U}_i^\alpha\}_{i\in\mathcal{N}}; G)$. However, in the context of broad attacks, we will refer to it as $\text{LLL}(\mathcal{A}, \alpha, S; G)$. The induced network *efficiency* is defined as

$$\begin{aligned} J_b(\alpha, S; G) &:= \frac{\min_{a\in\text{LLL}(\mathcal{A},\alpha,S;G)} W(a)}{\max_{a'\in\mathcal{A}} W(a')} \\ &= \frac{\min_{a\in\text{LLL}(\mathcal{A},\alpha,S;G)} W(a)}{(1+\alpha_{\text{sys}})|\mathcal{E}|} \end{aligned} \tag{8}$$

which is the ratio of the welfare induced by the welfare-minimizing SSS to the optimal welfare. The second equality abovementioned is due to the fact that optimal welfare is attained at $\vec{x}$ (all play x). We reiterate that the impostor nodes do not contribute to the system efficiency. They serve only to possibly modify the stochastically stable state. The *risk* from broad attacks faced by the system operator in choosing gain α is defined as

$$R_b(\alpha, S; G) := 1 - J_b(\alpha, S; G). \tag{9}$$

Risk measures the distance from optimal efficiency under operating gain α. Fig. 1(a) illustrates an example of a three-node network subject to a broad attack. The extent to which systems are susceptible to broad attacks is captured by the following definition of worst-case risk.

Definition 1: The *worst-case risk* to broad attacks is given by

$$R_{\text{b}}^*(\alpha) := \max_{N\geq 3} \max_{G\in\mathcal{G}_N} \max_{S\in\mathcal{T}(G)} R_{\text{b}}(\alpha, S; G). \tag{10}$$

The quantity $R_b^*(\alpha)$ is the cost metric that the system operator wishes to reduce given uncertainty of the network structure and target set. In other words, the adversary observes the design choice of the system operator, and chooses an optimal attack strategy.

Theorem 1: Let $\alpha > 0$. The worst-case broad risk is

$$R_{\text{b}}^*(\alpha) = \begin{cases} 1 - \left(\frac{k}{k+1}\right)\left(\frac{1}{1+\alpha_{\text{sys}}}\right), & \text{if } \alpha \in I_k, \text{ for } k = 1, 2, \ldots \\ 1 - \frac{1}{1+\alpha_{\text{sys}}}, & \text{if } \alpha \in \left[1, \frac{3}{2}\right] \\ 0, & \text{if } \alpha > \frac{3}{2} \end{cases} \tag{11}$$

where

$$I_k := \left(\frac{k-1}{k}, \frac{k}{k+1}\right]. \tag{12}$$

An illustration of R_{b}^* is given in Fig. 2(a), along with the graphs and target sets that achieve the worst-case risks. For sufficiently high gains $\alpha > \frac{3}{2}$, the system is safeguarded from any broad attack, i.e., the worst-case risk is zero. The technical results needed for the proof are given in Section V.

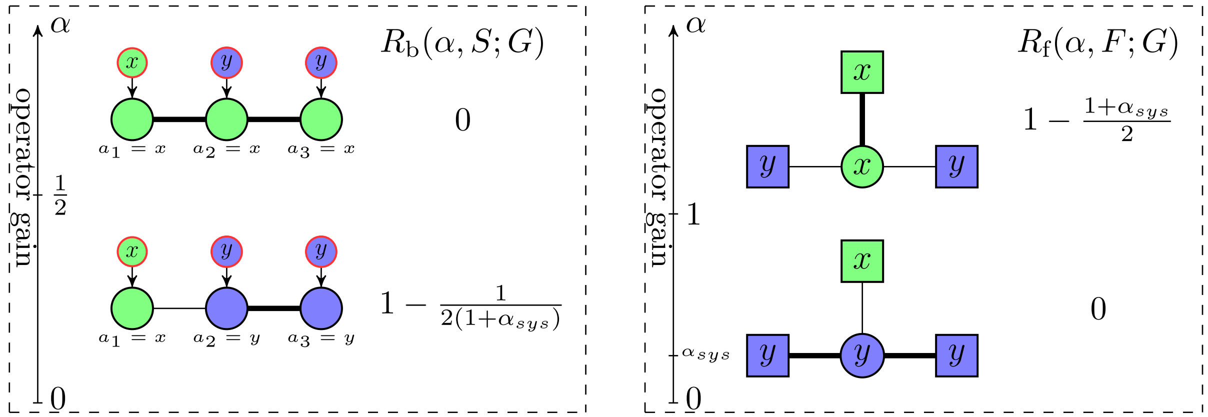
B. Focused Attacks and Worst-Case Risk Metric

An adversary is able to choose a strict subset of agents and force them to commit to prescribed choices. This causes them to act as *fixed agents*, or agents that do not update their choices over time. One could consider this as allowing the adversary an unlimited number of impostor nodes (instead of one) at its dispatch to attach to each agent in the subset, thereby solidifying their choices. This focused influence on a single agent is stronger than the influence a broad attack has on a single agent in the sense that the latter type does not require the agent to commit to a choice.

Let $F_x \subset \mathcal{N}$ (F_y) be the set of fixed x (y) agents. We call the *fixed set* $F = (F_x, F_y)$, which satisfies $F_x \cap F_y = \varnothing$ and $F_x \cup F_y \subset \mathcal{N}$. We call $\mathcal{F}(G)$ the set of all feasible fixed sets on a graph $G \in \mathcal{G}_N$. A fixed set $F \in \mathcal{F}(G)$ restricts the action space to $\mathcal{A}_F$, where $(\mathcal{A}_F)_i = \{x\}$ ($\{y\}$) $\forall i \in F_x$ (F_y) and $(\mathcal{A}_F)_i = \{x, y\}$ $\forall i \notin F$. We assume the adversary selects at least one fixed agent. The strict subset assumption avoids pathological cases (e.g., alternating x and y fixed nodes for an entire line network yields an efficiency of zero).

The set of stochastically stable states given a fixed set F is written as $\text{LLL}(\mathcal{A}_F, \{U_i^\alpha\}_{i\in\mathcal{N}}; G)$. However for brevity, we will refer to it as $\text{LLL}(\mathcal{A}_F, \alpha; G)$. The induced efficiency is

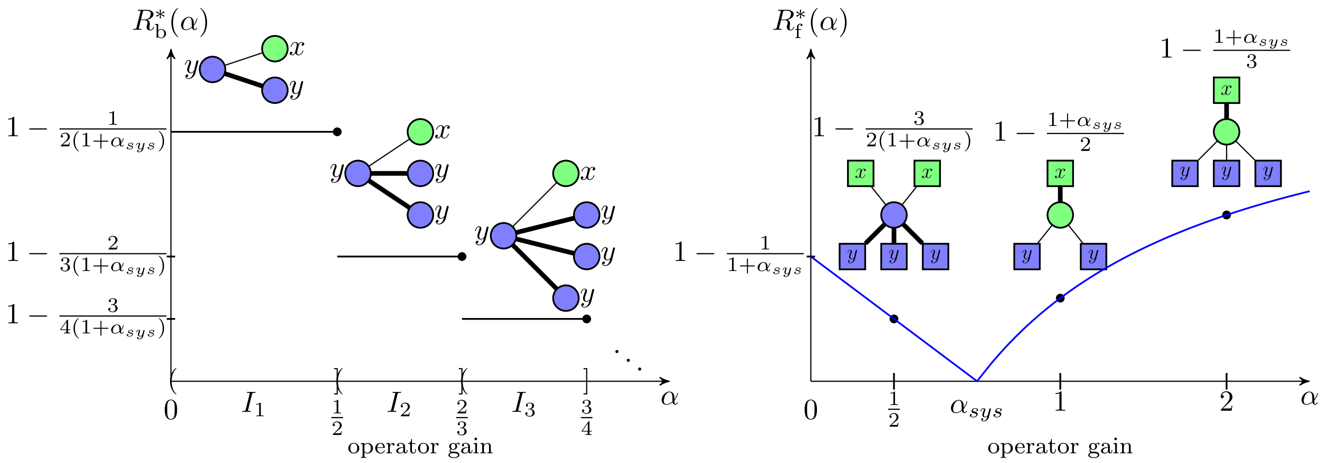
$$J_f(\alpha, F; G) := \frac{\min_{a\in\text{LLL}(\mathcal{A}_F,\alpha;G)} W(a)}{\max_{a\in\mathcal{A}_F} W(a)} \tag{13}$$



(a)

(b)

Fig. 1. (Left) An example three-node line network under a broad attack. The impostor nodes that are part of the target set S are depicted as the labeled smaller circles and agents in the network are the bigger circles. The color of each circle indicates the node's action—green for x, blue for y. For operator gains $\alpha \leq \frac{1}{2}$, the SSS is $(a_1, a_2, a_3) = (x, y, y)$. From (9), this gives a risk of $R_b(\alpha, S; G) = 1 - \frac{1}{2(1+\alpha_\mathrm{sys})}$. In other words, it induces suboptimal efficiency. For $\alpha > \frac{1}{2}$, the network is resilient to this attack because the welfare-minimizing SSS is (x, x, x). This gives optimal efficiency, i.e., a risk of 0. (Right) An example of a four node star network under a focused attack where a subset F of three nodes are targeted to be fixed (squares). Only the center node is unfixed. In this example, the maximum welfare is $\max_{a \in \mathcal{A}_F} W(a) = 4$, achieved when the center plays y. This is because the alternative action (when center plays x) gives the suboptimal welfare $2(1 + \alpha_\mathrm{sys}) < 4$ due to $\alpha_\mathrm{sys} < 1$. For operator gains $\alpha < 1$, the center node plays y in the SSS. This yields optimal efficiency, i.e., the risk is $R_f(\alpha, F; G) = 0$. For $\alpha \geq 1$, the center node plays x, giving a risk of $R_f(\alpha, F; G) = 1 - \frac{1+\alpha_\mathrm{sys}}{2}$. The methods to calculate stochastically stable states under both types of attacks follow standard potential game arguments and are detailed in Section V.



(a)

(b)

Fig. 2. (a) Worst-case risk from broad attacks $R_b^*(\alpha)$ (11) is a piecewise constant function defined over countably infinite half-open intervals. The graphs and their corresponding target set, which attain each level of worst-case broad risk are illustrated for $\alpha < 1$. Here, the x, y labels indicate the type of impostor influence on the agents (circles) in the network, and the color of the circles depict the action played in the welfare-minimizing SSS (green $= x$, blue $= y$). If $\alpha \in I_k$, $k = 1, 2, \ldots$[recall (12)], the worst-case risk occurs on a star graph of $k + 2$ nodes where all nodes but one are targeted with a y impostor. The one leaf node has an x impostor attached, giving a single miscoordinating link in the network. (b) Worst-case risk from focused attacks $R_f^*(\alpha)$ (16). The graphs and their corresponding fixed sets which attain the worst-case focused risks are illustrated for $\alpha = \frac{1}{2}, 1$, and 2. The nodes' color represents the worst-case SSS at α (blue $= y$, green $= x$). The targeted fixed agents are represented as squares and the unfixed agents as circles. Here, $\frac{1}{2} < \alpha_\mathrm{sys} < 1$. The proofs establishing all worst-case graphs are detailed in Section V.

which is the ratio of the welfare induced by the worst-case stable state to the optimal welfare given the fixed set F. The risk faced by the system operator in choosing α is defined as

$$R_f(\alpha, F; G) := 1 - J_f(\alpha, F; G). \tag{14}$$

Again, risk measures the distance from optimal efficiency when choosing α. The fixed nodes here differ from the impostor nodes in that they contribute to the true measured welfare (3) in addition to modifying the SSS by restricting the action profile space and influencing the decisions of their nonfixed neighbors. Fig. 1(b) provides an illustrative example of a network with three fixed agents and one unfixed agent. The extent to which the system is susceptible to focused attacks is defined by the following worst-case risk metric.

Definition 2: The *worst-case risk from focused attacks* is given by

$$R_{\mathrm{f}}^*(\alpha) := \max_{N \geq 3} \max_{G \in \mathcal{G}_N} \max_{F \in \mathcal{F}(G)} R_{\mathrm{f}}(\alpha, F; G). \tag{15}$$

The quantity $R_f^*(\alpha)$ is the cost metric that a system operator wishes to reduce given uncertainty on the graph structure and composition of fixed agents in the network.

Theorem 2: The worst-case risk from focused attacks is

$$R_{\mathrm{f}}^*(\alpha) = \begin{cases} 1 - \frac{1+\alpha}{1+\alpha_{\mathrm{sys}}}, & \text{if } \alpha < \alpha_{\mathrm{sys}} \\ 0, & \text{if } \alpha = \alpha_{\mathrm{sys}} \,. \\ 1 - \frac{1+\alpha_{\mathrm{sys}}}{1+\alpha}, & \text{if } \alpha > \alpha_{\mathrm{sys}} \end{cases} \tag{16}$$

The technical results needed for the proof are given in Section V. An illustration of this quantity as well as the graphs that induce worst-case risk are portrayed in Fig. 2(b). We observe the choice $\alpha = \alpha_{\mathrm{sys}}$ recovers optimal efficiency for any $G \in \mathcal{G}_N$ and $F \in \mathcal{F}(G)$. In other words, by operating at the system gain α_{sys}, the system operator safeguards efficiency from any focused attack. Furthermore, $R_f^*(\alpha)$ monotonically increases for $\alpha > \alpha_{\mathrm{sys}}$, approaching 1 in the limit $\alpha \to \infty$. Intuitively, the risk in this regime comes from inflating the benefit of the x convention, which can be harmful to system efficiency when there are predominantly fixed y nodes in the network. For $\alpha < \alpha_{\mathrm{sys}}$, $R_f^*(\alpha)$ monotonically decreases. The risk here stems from devaluing the x convention, which hurts efficiency when coordinating with fixed x nodes is valuable.

C. Fundamental Tradeoffs Between Risk and Security

We describe the operator's tradeoffs between the two worst-case risk metrics. That is, given a level of security $\gamma \in [0, 1]$ is ensured on one worst-case risk metric through a particular design choice α, what is the minimum achievable risk level of the other? These relations are direct consequences of Theorems 1 and 2.

Remark 1: Before presenting the tradeoff relations, we first observe that since $R_{\mathrm{f}}^*(\alpha)$ is decreasing on $\alpha < \alpha_{\mathrm{sys}}$ and $R_{\mathrm{b}}^*(\alpha)$ is decreasing in α, the operator should not select any gain $\alpha < \alpha_{\mathrm{sys}}$, as it worsens both risk levels. Hence, for the rest of this article, we only consider gains greater than α_{sys}.

Corollary 1: Fix $\gamma_{\mathrm{f}} \in [0, 1)$. Suppose $R_{\mathrm{f}}^*(\alpha) \leq \gamma_{\mathrm{f}}$ for some α. Then,

$$R_{\mathrm{b}}^*(\alpha) \geq R_{\mathrm{b}}^*\left(\frac{1+\alpha_{\mathrm{sys}}}{1-\gamma_{\mathrm{f}}} - 1\right). \tag{17}$$

Proof: From (16), $R_f^*(\alpha) \leq \gamma_f$ implies $\alpha \leq \frac{1+\alpha_{\mathrm{sys}}}{1-\gamma_f} - 1$. Since $R_b^*(\alpha)$ is a decreasing function in α, we obtain the result. ■

In words, as the security from worst-case focused attacks improves (γ_f lowered), the risk from worst-case broad attacks increases. A tradeoff relation also holds in the opposite direction.

Corollary 2: Fix $\gamma_{\mathrm{b}} \in [0, 1]$. Suppose $R_{\mathrm{b}}^*(\alpha) \leq \gamma_{\mathrm{b}}$ for some α. Suppose $\alpha_{\mathrm{sys}} \in I_{k_{\mathrm{sys}}}$ for some $k_{\mathrm{sys}} \in \{1, 2, \ldots\}$. Then,

$$R_{\mathrm{f}}^*(\alpha) \begin{cases} \geq 0, & \text{if } \gamma_{\mathrm{b}} \in \left[1 - \frac{k_{\mathrm{sys}}}{k_{\mathrm{sys}}+1}\frac{1}{1+\alpha_{\mathrm{sys}}}, 1\right] \\ > R_{\mathrm{f}}^*\left(\frac{k}{k+1}\right), & \text{if } \gamma_{\mathrm{b}} \in \left[1 - \frac{k}{k+1}\frac{1}{1+\alpha_{\mathrm{sys}}}, 1 - \frac{k-1}{k}\frac{1}{1+\alpha_{\mathrm{sys}}}\right) \\ & \quad \text{for } k = k_{\mathrm{sys}}, k_{\mathrm{sys}}+1, \ldots \\ \geq R_{\mathrm{f}}^*(1), & \text{if } \gamma_{\mathrm{b}} = 1 - \frac{1}{1+\alpha_{\mathrm{sys}}} \\ > R_{\mathrm{f}}^*\left(\frac{3}{2}\right), & \text{if } \gamma_{\mathrm{b}} \in \left[0, 1 - \frac{1}{1+\alpha_{\mathrm{sys}}}\right). \end{cases} \tag{18}$$

If $\alpha_{\mathrm{sys}} \in [1, 3/2]$

$$R_{\mathrm{f}}^*(\alpha) \begin{cases} \geq 0, & \text{if } \gamma_{\mathrm{b}} \in \left[1 - \frac{1}{1+\alpha_{\mathrm{sys}}}, 1\right] \\ > R_{\mathrm{f}}^*\left(\frac{3}{2}\right), & \text{if } \gamma_{\mathrm{b}} \in \left[0, 1 - \frac{1}{1+\alpha_{\mathrm{sys}}}\right). \end{cases} \tag{19}$$

If $\alpha_{\mathrm{sys}} > \frac{3}{2}$, then $R_{\mathrm{f}}^*(\alpha) \geq 0$ for any γ_{b}.

Proof: All bounds are computed by finding $\inf_\alpha R_f^*(\alpha)$ s.t. $R_b^*(\alpha) \leq \gamma_b$. The relations $\geq$ and $>$ follow from the fact that $R_f^*(\alpha)$ is increasing in $\alpha > \alpha_{\mathrm{sys}}$, and depending on whether R_f^* can attain the resulting value. ■

Here, as the security from worst-case broad attacks improves (γ_b lowered), the risk from worst-case focused attacks increases. Each of the broad risk levels can be attained for a range of focused risks. An illustration of the attainable worst-case risk levels is given in Fig. 3 (blue).

IV. Randomized Operator Strategies

In this section, we consider the scenario where the operator randomizes over multiple gains. We present a definition and a characterization of worst-case expected risks. We then identify the risk-security tradeoffs available in the randomized gain setting. We observe they significantly improve upon the deterministic gain setting (see Fig. 3). We then identify ways to further improve these tradeoffs through different randomizations.

A. Worst-Case Expected Risks

Suppose the operator selects a gain from the M distinct values $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^M$ satisfying $\alpha_1 < \alpha_2 < \cdots < \alpha_M$ with the probability distribution $\boldsymbol{p} = [p_1, \ldots, p_M]^\top \in \Delta_M$. Here, we denote $\Delta_M = \{\boldsymbol{p} \in \mathbb{R}_+^M : \sum_{j=1}^M p_j = 1\}$ as the set of all M-dimensional probability vectors. In other words, the operator employs the payoff gain α_j with probability p_j.

We consider the following natural definitions of expected risks. Given a graph $G \in \mathcal{G}_N$ and target set $S \in \mathcal{T}(G)$, let $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b|S, G] := \sum_{j=1}^M p_j R_b(\alpha_j, S; G)$ be the expected adversarial risk of the operator's strategy $\boldsymbol{\alpha}, \boldsymbol{p}$. The worst-case expected risk from broad attacks is defined as

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_b] := \max_{N \geq 3} \max_{G \in \mathcal{G}_N} \max_{S \in \mathcal{T}(G)} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b|S, G]. \tag{20}$$

Similarly, given a fixed set $F \in \mathcal{F}(G)$, let $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_f|F, G] := \sum_{j=1}^M p_i R_f(\alpha_j, F; G)$ be the expected risk from focused attacks. The worst-case expected risk from focused attacks is

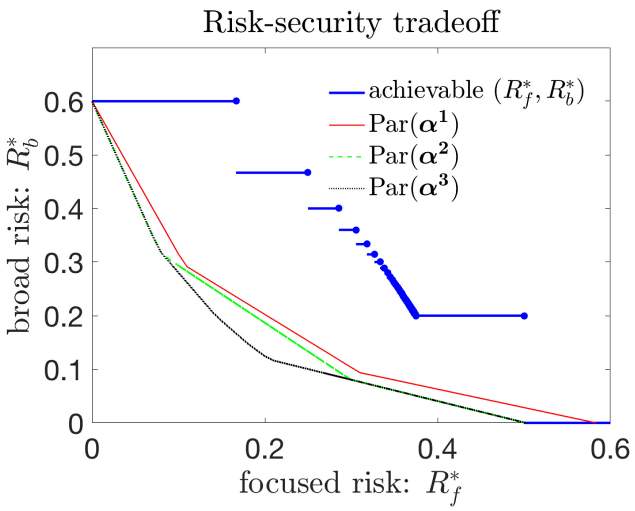


Fig. 3. Security-risk tradeoffs are depicted by the achievable worst-case risk levels from deterministic gains (blue) and randomized gains (red, green, black). The Pareto frontiers for three different randomized strategies $\boldsymbol{\alpha}^1, \boldsymbol{\alpha}^2 \in \mathbb{R}_+^5$, and $\boldsymbol{\alpha}^3 \in \mathbb{R}_+^{300}$, are shown in increasing order of improvement. The strategies $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ randomize over the highest three broad risk levels in addition to the lowest two. The strategy $\boldsymbol{\alpha}^3$ randomizes over the highest 298 broad risk levels and the lowest two. We chose the values as follows. For $k = 1, 2$, we set $\alpha_1^k = \alpha_{\text{sys}}$, $\alpha_j^k = (1 - \epsilon_k)\frac{j-1}{j} + \epsilon_k \frac{j}{j+1} \in I_j$ for $j = 2, 3$, $\alpha_4^k = 1 + \epsilon_k$, and $\alpha_5^k = \frac{3}{2} + \epsilon_k$. We have set $\epsilon_1 = 0.5$ and $\epsilon_2 = .01$. Hence, Par($\boldsymbol{\alpha}^2$) improves upon Par($\boldsymbol{\alpha}^1$) via Claim 1. For $k = 3$, we set $\alpha_1^3 = \alpha_{\text{sys}}$, $\alpha_j^3 = (1 - \epsilon_3)\frac{j-1}{j} + \epsilon_3 \frac{j}{j+1} \in I_j$, $j = 2, 3, \ldots, 298$, $\alpha_{299}^3 = 1 + \epsilon_3$, and $\alpha_5^3 = \frac{3}{2} + \epsilon_3$. Claim 2 ensures Par($\boldsymbol{\alpha}^3$) improves upon Par($\boldsymbol{\alpha}^2$). We chose $\epsilon_3 = .01$ and $\alpha_{\text{sys}} = 1/4$.

defined as

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_f] := \max_{N \geq 3} \max_{G \in \mathcal{G}_N} \max_{F \in \mathcal{F}(G)} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_f | F, G]. \tag{21}$$

Theorem 3: Suppose the operator randomizes with gains $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^M$ according to $\boldsymbol{p} \in \Delta_M$. Then, the worst-case expected broad risk is

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{b}}] = \max_{k=1,\ldots,M} \left\{ \left(\sum_{j=1}^{k} p_j \right) R_{\text{b}}^*(\alpha_k) \right\}. \tag{22}$$

The worst-case expected focused risk is

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{f}}] = \max_{k=1,\ldots,M} \left\{ \left(\sum_{j=k}^{M} p_j \right) R_{\text{f}}^*(\alpha_k) \right\}. \tag{23}$$

The proofs are given in Section VI. The characterization of worst-case expected risk is a discounted weighting of a deterministic worst-case risk level. This suggests that the risk levels achievable by randomization can improve upon the risks induced from a deterministic gain.

B. Risk Tradeoffs Under Randomized Operator Strategies

Given a level of security $\gamma \in [0, 1]$ is ensured on one expected worst-case metric, what is the minimum achievable risk level on the other? We find this can be calculated through a linear program. We formalize these tradeoffs in the following two statements, which are analogous to Corollaries 1 and 2.

Corollary 3: Fix $\gamma_{\text{f}} \in [R_{\text{f}}^*(\alpha_1), 1]$ and a set of gains $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^M$. Suppose $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{f}}] \leq \gamma_{\text{f}}$ for some $\boldsymbol{p} \in \Delta_M$. Then,

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{b}}] \geq v_{\text{b}}(\gamma_{\text{f}}, \boldsymbol{\alpha}) \tag{24}$$

where $v_b(\gamma, \boldsymbol{\alpha})$ is the value of the following linear program:

$$\begin{aligned} v_b(\gamma_{\text{f}}, \boldsymbol{\alpha}) &= \min_{\boldsymbol{p}', v} v \\ &\text{s.t.} \sum_{i=1}^{M} p_i' = 1,\ p_i \geq 0\ \forall i = 1, \ldots, M \\ &v \in [0, 1] \\ &A_{\text{LP}} \begin{bmatrix} \boldsymbol{p}' \\ v \end{bmatrix} \preceq \begin{bmatrix} 0_M \\ \gamma_{\text{f}} \mathbb{1}_M \end{bmatrix} \end{aligned} \tag{25}$$

where $\preceq$ denotes elementwise $\leq$, 0_M and $\mathbb{1}_M$ are column M-vectors of zeros and ones, respectively, and A_{LP} is the $2M \times (M+1)$ matrix

$$A_{\text{LP}} = \left[\begin{array}{cccc:c} R_{\text{b}}^*(\alpha_1) & 0 & \cdots & 0 & -1 \\ R_{\text{b}}^*(\alpha_2) & R_{\text{b}}^*(\alpha_2) & \cdots & \vdots & \vdots \\ \vdots & & \ddots & 0 & \\ R_{\text{b}}^*(\alpha_M) & \cdots & \cdots & R_{\text{b}}^*(\alpha_M) & -1 \\ \hdashline R_{\text{f}}^*(\alpha_1) & \cdots & \cdots & R_{\text{f}}^*(\alpha_1) & 0 \\ 0 & R_{\text{f}}^*(\alpha_2) & \cdots & R_{\text{f}}^*(\alpha_2) & \vdots \\ \vdots & & \ddots & \vdots & \\ 0 & \cdots & 0 & R_{\text{f}}^*(\alpha_M) & 0 \end{array} \right]. \tag{26}$$

Moreover, $v_b(\gamma_{\text{f}}, \boldsymbol{\alpha})$ is decreasing in γ_{f}.

Proof: We need to show equivalence between the linear program (25) and the optimization problem

$$\min_{\boldsymbol{p}' \in \Delta_M} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_b] \text{ subject to } \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_f] \leq \gamma_f. \tag{27}$$

Let $A_b(\boldsymbol{\alpha}) \in \mathbb{R}^{M \times M}$ be the matrix defined by the upper left block of (26) and $A_f(\boldsymbol{\alpha})$ by the bottom left block. From Theorem 3, we can express $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_b]$ as the maximum element of the M-vector $A_b(\boldsymbol{\alpha})\boldsymbol{p}'$, and similarly $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_f]$ as the maximum element of $A_f(\boldsymbol{\alpha})\boldsymbol{p}'$. Hence, $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_f] \leq \gamma_f$ is the linear constraint $[A_f(\boldsymbol{\alpha})\boldsymbol{p}']_i \leq \gamma$ for all $i = 1, \ldots, M$. The objective $\min_{\boldsymbol{p}' \in \Delta_M} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}^*[R_b]$ itself can be cast as a linear objective with linear constraints, i.e., $\min_{\boldsymbol{p}' \in \Delta_M, v \in [0,1]} v$ s.t. $[A_b(\boldsymbol{\alpha})\boldsymbol{p}']_i \leq v$. Combining these two, we obtain (25). The claim $v_b(\gamma, \boldsymbol{\alpha})$ is decreasing in γ follows as a consequence of the linear program (25). ■

We note that a worst-case expected focused risk $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_f] < R_f^*(\alpha_1)$ is not attainable because α_1 is the smallest gain it mixes with. Hence, the linear program (25) is infeasible for $\gamma_f < R_f^*(\alpha_1)$. The following tradeoff relation holds in the opposite direction.

Corollary 4: Fix $\gamma_{\text{b}} \in [R_{\text{b}}^*(\alpha_M), 1]$ and a set of gains $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^M$. Suppose $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{b}}] \leq \gamma_{\text{b}}$ for some $\boldsymbol{p} \in \Delta_M$. Then,

$$\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}^*[R_{\text{f}}] \geq v_{\text{f}}(\gamma_{\text{b}}, \boldsymbol{\alpha}) \tag{28}$$

where $v_{\mathrm{f}}(\gamma_{\mathrm{b}}, \boldsymbol{\alpha})$ is the value of the following linear program:

$$\begin{aligned} v_{\mathrm{f}}(\gamma_{\mathrm{b}}, \boldsymbol{\alpha}) = \min_{\boldsymbol{p}, v} &\; v \\ \text{s.t.} &\sum_{i=1}^{M} p_i = 1,\ p_i \geq 0\ \forall i = 1, \ldots, M \\ & v \in [0, 1] \\ & \left[\begin{array}{c:c} A_{\mathrm{f}}(\boldsymbol{\alpha}) & -\mathbb{1}_M \\ \hdashline A_{\mathrm{b}}(\boldsymbol{\alpha}) & 0_M \end{array}\right] \begin{bmatrix} \boldsymbol{p} \\ v \end{bmatrix} \preceq \begin{bmatrix} 0_M \\ \gamma_{\mathrm{b}} \mathbb{1}_M \end{bmatrix} \end{aligned} \tag{29}$$

where $A_{\mathrm{f}}(\boldsymbol{\alpha})$ and $A_{\mathrm{b}}(\boldsymbol{\alpha})$ are defined as the bottom and top-left blocks of (26), respectively. Furthermore, $v_{\mathrm{f}}(\gamma_{\mathrm{b}}, \boldsymbol{\alpha})$ is decreasing in γ_{b}.

We omit the proof as it is similar to that of Corollary 3. Note a worst-case expected broad risk $\mathbb{E}^*_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b] < R_b^*(\alpha_M)$ is not attainable since α_M is the highest gain it mixes with—(29) is infeasible for $\gamma_b < R_b^*(\alpha_M)$. Fig. 3 plots the best achievable risk levels of three randomized operator strategies (red, green, and black).

C. Improvement of Risk Tradeoffs

The tradeoff relations describe the best achievable level on one risk metric given the other is subject to a security constraint when the gains $\boldsymbol{\alpha}$ are fixed. One way to improve the achievable risks is to decrease the available gains.

Claim 1: Let $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathbb{R}^M$. Suppose $\alpha_j \in I_{k_j}$ [recall (12)], $j = 1, \ldots, M$ for some nondecreasing subsequence $k_j \geq 1$. Let $\boldsymbol{\alpha}'$ satisfy $\alpha_j' \in I_{k_j}$ with $\alpha_j' < \alpha_j$. Then, for all $\gamma_{\mathrm{b}} \in [R^*_{\mathrm{b}}(\alpha_M), 1]$, $v_{\mathrm{f}}(\gamma_{\mathrm{b}}, \boldsymbol{\alpha}') \leq v_{\mathrm{f}}(\gamma_{\mathrm{b}}, \boldsymbol{\alpha})$. Similarly, for all $\gamma_{\mathrm{f}} \in [R^*_{\mathrm{f}}(\alpha_1), 1]$, $v_{\mathrm{b}}(\gamma_{\mathrm{f}}, \boldsymbol{\alpha}') \leq v_{\mathrm{b}}(\gamma_{\mathrm{f}}, \boldsymbol{\alpha})$.

Randomizing over additional gains can also improve the achievable risks.

Claim 2: Suppose $\boldsymbol{\alpha} \in \mathbb{R}^M$ and $\boldsymbol{\alpha}' \in \mathbb{R}^{M'}$ with $M < M'$, and assume $\boldsymbol{\alpha}'$ contains the elements of $\boldsymbol{\alpha}$. Then, the assertion of Claim 1 holds.

The proofs of the abovementioned two Claims follow directly from the formulation of the LPs (25), (29), and hence we omit them.

Fig. 3 depicts the best achievable risk levels of three randomized operator strategies of increasing improvement due to Claims 1 and 2 (red, green, and black curves). In particular, these plots constitute the *Pareto frontier* of all attainable expected risks among distributions $\boldsymbol{p}$ given a fixed set of gains. That is, for any $\boldsymbol{\alpha}$, we say a risk level $\begin{bmatrix} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_f] \\ \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b] \end{bmatrix} \in \mathbb{R}^2$ belongs to the frontier $\text{Par}(\boldsymbol{\alpha})$ if there does not exist a $\boldsymbol{p}' \neq \boldsymbol{p}$ such that $\begin{bmatrix} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}[R_f] \\ \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}'}[R_b] \end{bmatrix} \preceq \begin{bmatrix} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_f] \\ \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b] \end{bmatrix}$. Within $\text{Par}(\boldsymbol{\alpha})$, the operator can only improve upon one worst-case risk metric by sacrificing performance on the other.

From Corollary 4, the frontier given gains $\boldsymbol{\alpha}$ is the set of points

$$\text{Par}(\boldsymbol{\alpha}) = \left\{ \begin{bmatrix} v_f(\gamma_b, \boldsymbol{\alpha}) \\ \gamma_b \end{bmatrix} \in \mathbb{R}^2 : \gamma_b \in [R_b^*(\alpha_M), R_b^*(\alpha_1)] \right\}. \tag{30}$$

The parameter γ_b is upper bounded here by $R_b^*(\alpha_1)$ since any risk level with $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b] > R_b^*(\alpha_1)$ is unattainable under $\boldsymbol{\alpha}$. Hence, the values $v_f(\gamma_b, \boldsymbol{\alpha})$ and $v_f(R_b^*(\alpha_1), \boldsymbol{\alpha})$ are equivalent for $\gamma_b > R_b^*(\alpha_1)$. The frontiers in Fig. 3 are generated by numerically solving the linear program (29) for a finite grid of points $\gamma_b \in [R_b^*(\alpha_M), R_b^*(\alpha_1)]$.

As we have seen, the transition from deterministic to randomized gains ensures a reduction of risk levels. Randomizing over only a few different gains substantially improves upon the attainable deterministic worst-case risks. However, a detailed quantification of such improvements remains a challenge due to the high dimensionality of the model. In particular, we have yet identified a "limit" frontier that could be obtained by selecting gains detailed by Claims 1 and 2.

V. Proof of Theorems 1 and 2: Deterministic Worst-Case Risks

In this section, we develop the technical results that characterize the worst-case risk metrics $R_b^*(\alpha)$ and $R_f^*(\alpha)$ (see Theorems 1 and 2). Before presenting the proofs, we first present some preliminaries on potential games [25], which are essential to calculating stochastically stable states. We, then, define relevant notations for the forthcoming analysis.

A. Potential Games

Graphical coordination games fall under the class of potential games–games where individual utilities $\{U_i\}_{i \in \mathcal{N}}$ are aligned with a global objective, or potential function. A game is a potential game if there exists a potential function $\phi : \mathcal{A} \to \mathbb{R}$, which satisfies

$$\phi(a_i, a_{-i}) - \phi(a_i', a_{-i}) = U_i(a_i, a_{-i}) - U_i(a_i', a_{-i}) \tag{31}$$

for all $i \in \mathcal{N}$, $a \in \mathcal{A}$, and $a_i' \neq a_i$ [25]. In potential games, the set of stochastically stable states (6) are precisely the action profiles that maximize the potential function [16], [17]. Specifically, $\text{LLL}(\mathcal{A}, \{U_i\}_{i \in \mathcal{N}}; G) = \arg\max_{a \in \mathcal{A}} \phi(a)$. Our analysis relies on characterizing a potential function for the graphical coordination game in the presence of adversarial influences. This allows us to compute stochastically stable states in a straightforward manner.

B. Relevant Notations for Analysis

Any action profile a on a graph $G = (\mathcal{N}, \mathcal{E}) \in \mathcal{G}_N$ decomposes $\mathcal{N}$ into x and y-partitions. A node that belongs to a y-partition (x-partition) has $a_i = y$ (x). The partitions are enumerated $\{\mathcal{P}_y^1, \ldots, \mathcal{P}_y^{k_y}\}$ and $\{\mathcal{P}_x^1, \ldots, \mathcal{P}_x^{k_x}\}$, are mutually disjoint, and cover the graph. Each partition is a connected subgraph of G. It is possible that $k_x = 0$ with $k_y = 1$ (when $a = \vec{y}$), $k_x = 1$ with $k_y = 0$ (when $a = \vec{x}$), or $k_y, k_x \geq 1$.

For any subset of nodes $A, B \subseteq \mathcal{N}$, let us denote

$$e(A, B) := \{(i, j) \in \mathcal{E} : i \in A, j \in B\} \tag{32}$$

as the set of edges between A and B. We write A^c as the complement of A. We extensively use the notation

$$W^\alpha(E, a) := \sum_{(i,j)\in E} V^\alpha(a_i, a_j) \tag{33}$$

as the welfare due to edge set $E \subseteq \mathcal{E}$ in action profile a, where V^α is of the form (1) with α_{sys} replaced by α. For compactness, we will denote $W(E, a)$ as $W^{\alpha_{\text{sys}}}(E, a)$ for the local system welfare generated by the edges E. Our analysis will also rely on the following mediant inequality.

Fact 1: Suppose $n_i \geq 0$ and $d_i > 0$ for each $i = 1, \ldots, m \in \mathbb{N}$. Then,

$$\frac{\sum_{i=1}^m n_i}{\sum_{i=1}^n d_i} \geq \min_i \frac{n_i}{d_i}. \tag{34}$$

We refer to the LHS abovementioned as the *mediant sum* of the $\frac{n_i}{d_i}$.

C. Characterization of R_b^*: Worst-Case Broad Risk

To prove Theorem 1, we seek a pair (S, G) with $G \in \mathcal{G}_N$ of any size $N \geq 3$ and $S \in \mathcal{T}(G)$, that minimizes efficiency $J_b(\alpha, S; G)$ (maximizes risk $R_b(\alpha, S; G)$). Our method to find the minimizer is to show any (S, G) can be transformed into a star network with a particular target set that has lower efficiency, when $\alpha < 1$. Thus, in this regime the search for the worst-case graph reduces to the class of star networks of arbitrary size. For $\alpha \geq 1$, structural properties allow us to deduce the minimal efficiency.

The graphical coordination game defined by $\mathcal{A} = \{x, y\}^N$, perceived utilities $\{\tilde{U}_i^\alpha\}_{i\in\mathcal{N}}$ (7), target set S, and graph G falls under the class of potential games [25]. A potential function is given by

$$\frac{1}{2} W^\alpha(a) + (1 + \alpha) \sum_{i\in S_x} \mathbb{1}(a_i = x) + \sum_{i\in S_y} \mathbb{1}(a_i = y) \tag{35}$$

where

$$W^\alpha(a) := \sum_{i\in\mathcal{N}} U_i^\alpha(a). \tag{36}$$

Hence, the stochastically stable states $\text{LLL}(\mathcal{A}, \alpha, S; G)$ are maximizers of (35). Suppose $\hat{a} = \arg\min_{a\in\text{LLL}(\mathcal{A},\alpha,S;G)} W(a)$ is the welfare-minimizing SSS inducing the partitions $\{\mathcal{P}_z^k\}_{k=1}^{k_z}$, $z = x, y$. We can express its efficiency from (8) as

$$\frac{\sum_{k=1}^{k_y} |e(\mathcal{P}_y^k, \mathcal{P}_y^k)| + (1 + \alpha_{\text{sys}}) \sum_{k=1}^{k_x} |e(\mathcal{P}_x^k, \mathcal{P}_x^k)|}{(1 + \alpha_{\text{sys}})(\sum_{k=1}^{k_y} |e(\mathcal{P}_y^k, \mathcal{N})| + \sum_{k=1}^{k_x} |e(\mathcal{P}_x^k, \mathcal{P}_x^k)|)}. \tag{37}$$

Note the denominator is simply the number of edges in G multiplied by $1+\alpha_{\text{sys}}$. From (35), each y-partition $\mathcal{P}_y^k$ in $\hat{a}$ satisfies[1]

$$|\mathcal{P}_y^k| + |e(\mathcal{P}_y^k, \mathcal{P}_y^k)| \geq \max_{a_{\mathcal{P}_y^k} \neq \vec{y}_{\mathcal{P}_y^k}} W^\alpha(e(\mathcal{P}_y^k, \mathcal{N}), (a_{\mathcal{P}_y^k}, \hat{a}_{-\mathcal{P}_y^k})) + \sum_{i\in\mathcal{P}_y^k} \mathbb{1}(a_i = y). \tag{CY}$$

In words, no subset of agents in $\mathcal{P}_y^k$ can deviate from y to improve the collective perceived welfare of $\mathcal{P}_y^k$. A similar stability condition holds for each x-partition $\mathcal{P}_x^k$

$$\begin{aligned}(1 + \alpha)\big(|\mathcal{P}_x^k| + |e(\mathcal{P}_x^k, \mathcal{P}_x^k)|\big) &\geq \max_{a_{\mathcal{P}_x^k} \neq \vec{x}_{\mathcal{P}_x^k}} W^\alpha(e(P_x^k, \mathcal{N}), (a_{\mathcal{P}_x^k}, \hat{a}_{-\mathcal{P}_x^k})) \\ &\quad + (1 + \alpha) \sum_{i\in\mathcal{P}_x^k} \mathbb{1}(a_i = x).\end{aligned} \tag{CX}$$

The following result characterizes the threshold on α above which any network is safeguarded from any impostor attack.

Lemma 1: Let $N \geq 3$. Then, $\alpha > \frac{N}{N-1}$ if and only if

$$\min_{G\in\mathcal{G}_N} \min_{S\in\mathcal{T}(G)} J_b(\alpha, S; G) = 1. \tag{38}$$

Proof: ($\Rightarrow$) Let $\alpha > \frac{N}{N-1}$. Suppose there is a pair (S, G) with $J_b(\alpha, G, S) < 1$. Then there must exist a y-partition $\mathcal{P}_y \subset \mathcal{N}$. From (CY)

$$|\mathcal{P}_y| + |e(\mathcal{P}_y, \mathcal{P}_y)| \geq (1 + \alpha)|e(\mathcal{P}_y, \mathcal{N})| > 2|e(\mathcal{P}_y, \mathcal{N})|. \tag{39}$$

Since G is connected, $|e(\mathcal{P}_y, \mathcal{P}_y)| \geq |\mathcal{P}_y| - 1$ and there is at least one outgoing link from $\mathcal{P}_y$, i.e., $|e(\mathcal{P}_y, \mathcal{P}_y^c)| \geq 1$. Consequently, $|e(\mathcal{P}_y, \mathcal{N})| \geq |\mathcal{P}_y|$, from which we obtain

$$|e(\mathcal{P}_y, \mathcal{N})| + |e(\mathcal{P}_y, \mathcal{P}_y)| > 2|e(\mathcal{P}_y, \mathcal{N})| \tag{40}$$

which is impossible.

($\Leftarrow$) Assume $\min_{G\in\mathcal{G}_N} \min_{S\in\mathcal{T}(G)} J_b(\alpha, S; G) = 1$. Then, no y-partition can exist for any graph. In particular, (CY) is violated for $\mathcal{P}_y = \mathcal{N}$

$$N + |\mathcal{E}| < (1 + \alpha)|\mathcal{E}| \Rightarrow \alpha > \frac{N}{|\mathcal{E}|}. \tag{41}$$

Since $|\mathcal{E}| \geq N - 1$, we obtain $\alpha > \frac{N}{N-1}$. ■

We also deduce the following minimal efficiencies for any graph when $1 \leq \alpha \leq \frac{N}{N-1}$.

Lemma 2: Suppose $N \geq 3$. Then, $\alpha \in [1, \frac{N}{N-1}]$ if and only if

$$\min_{G\in\mathcal{G}_N} \min_{S\in\mathcal{T}(G)} J_b(\alpha, S; G) = \frac{1}{1 + \alpha_{\text{sys}}}. \tag{42}$$

Proof: The ($\Rightarrow$) direction follows the same argument as Lemma 1.

[1] Since we are seeking worst-case pairs (S, G), we may consider any y-partition as only having y impostors placed among its nodes. This is because any x impostors that were placed in a resulting y-partition can be replaced by y-impostors and retain stability. We reflect this generalization in (CY) and (CX), where influence from only y (x) impostors is considered.

($\Leftarrow$) The assumption implies the only y-partition that is stabilizable is $\mathcal{N}$. Then, for any $\mathcal{P}_y \subset \mathcal{N}$, (CY) is violated, i.e.,

$$|\mathcal{P}_y| + |e(\mathcal{P}_y,\mathcal{P}_y)| < (1+\alpha)|e(\mathcal{P}_y,\mathcal{N})|. \tag{43}$$

Since G is connected and there is at least one outgoing edge from $\mathcal{P}_y$, we obtain

$$\frac{2|\mathcal{P}_y|-1}{|\mathcal{E}|} < 1+\alpha. \tag{44}$$

The abovementioned equation holds for any graph $G=(\mathcal{N},\mathcal{E})$ and subset of nodes $\mathcal{P}_y \subset \mathcal{N}$. From the facts that $|\mathcal{P}_y| \le N-1$ and $|\mathcal{E}| \ge N-1$, we have $\alpha > \frac{N-2}{N-1}$ for any $N \ge 3$. Consequently, $\alpha \ge 1$ and Lemma 1 establishes that $\alpha \le \frac{N}{N-1}$. ■

The class of star graphs is central to the worst-case analysis in the interval $0<\alpha<1$.

Definition 3: Let $\mathbb{S}_N$ be the set of all (S,G) where G is the star graph with N nodes, S_y contains the center node, and $S_x = \mathcal{N}\backslash S_y$.

An immediate consequence of this definition is the leaf nodes S_x satisfy (CX). The efficiency is, then, proportional to the fraction of leaf nodes that are stable to y, if any. Furthermore, the stability condition (CY) of $\mathcal{P}_y = S_y$ for members of $\mathbb{S}_N$ simplifies to

$$2|e(\mathcal{P}_y,\mathcal{P}_y)| + 1 \ge (1+\alpha)(N-1). \tag{45}$$

In other words, stability of the target set S_y as a y-partition hinges on (CY) being satisfied for the selection $a_{\mathcal{P}_y} = \vec{x}$. The following result reduces the search space for efficiency minimizers to $\mathbb{S}_N$ when $\alpha < 1$.

Lemma 3: Suppose $0<\alpha<1$ and $n \ge 3$. Consider any (S,G) with $G \in \mathcal{G}_N$, $S \in \mathcal{T}(G)$. Then, there is a $(S',G') \in \mathbb{S}_{N'}$ such that $J_{\text{b}}(\alpha,S';G') \le J_{\text{b}}(\alpha,S;G)$ for some $N' \ge N$.

The idea of the proof is to construct a member of $\mathbb{S}_{N'}$ by recasting the y and x-partitions of (S,G) as star subgraphs while preserving the same number and type of edges, thus preserving efficiency. Further efficiency reduction can be achieved by converting excess x links into y links in this star configuration. We provide the proof detailing the constructive procedure in the Appendix. We now characterize the minimal efficiency for the star graph of size N, $J_N^*(\alpha) := \min_{(G,S)\in\mathbb{S}_N} J_b(\alpha,S;G)$ for $\alpha < 1$.

Lemma 4: Suppose $\alpha<1$ and fix $N \ge 3$. Then,

$$J_N^*(\alpha) = \frac{1}{(1+\alpha_{\text{sys}})(N-1)}\left\lceil \frac{(1+\alpha)(N-1)-1}{2} \right\rceil. \tag{46}$$

Proof: The goal is to find the smallest y-partition of the n star that is still stabilizable under a gain α. This is written

$$J_N^*(\alpha) = \min_{N_y} \frac{1}{1+\alpha_{\text{sys}}}\frac{N_y}{N-1}$$
$$\text{s.t.} \begin{cases} N_y \le N-1, & \text{(size of } y\text{-partition)} \\ 2N_y + 1 \ge (1+\alpha)(N-1), & \text{(stability).} \end{cases} \tag{47}$$

The smallest integer N_y that satisfies the constraints is $\left\lceil \frac{(1+\alpha)(N-1)-1}{2} \right\rceil$ for $\alpha \in (0,1)$. ■

Proof of Theorem 1: For $\alpha<1$, by Lemma 3, the worst-case efficiency is

$$\min_{N\ge 3}\min_{(G,S)\in\mathbb{S}_N} J_b(\alpha,S;G) = \min_{N\ge 3} J_N^*(\alpha). \tag{48}$$

Using the formula of Lemma 4, we obtain the first entry in (11). Lemma 2 asserts the minimal efficiency is $\frac{1}{1+\alpha_{\text{sys}}}$ for $\alpha \in [1,\frac{3}{2}]$ because the upper bound $\frac{N}{N-1}$ is maximized at $N=3$ (for $N \ge 3$). This gives the second entry in (11). Finally, Lemma 1 asserts the minimal efficiency is 1 for $\alpha > \frac{3}{2}$. ■

D. Characterization of R_f^*: Worst-Case Focused Risk

Our approach for the proof of Theorem 2 differs from that of R_b^*. Instead of reducing the search of worst-case graphs, we simply provide an upper bound on $R_f^*(\alpha,F;G)$ for any G and fixed set $F \in \mathcal{F}(G)$, and show one can construct a graph with fixed nodes that achieves it..

We observe $\frac{1}{2}W^\alpha(a) : \mathcal{A}_F \to \mathbb{R}$ serves as a potential function [recall (36)] for the game with restricted action set $\mathcal{A}_F$ and utilities $\{U_i^\alpha\}_{i\in\mathcal{N}}$. Hence, the stochastically stable states $\text{LLL}(\mathcal{A}_F,\alpha;G)$ are maximizers of $\frac{1}{2}W^\alpha(a)$. Suppose $\hat{a} = \arg\min_{a\in\text{LLL}(\mathcal{A}_F,\alpha;G)} W(a)$ decomposes the graph into the x and y-partitions $\{\mathcal{P}_z^k\}_{k=1}^{k_z}$, $z=x,y$. We express its efficiency (13) as

$$\frac{\sum_{k=1}^{k_y}|e(\mathcal{P}_y^k,\mathcal{P}_y^k)| + (1+\alpha_{\text{sys}})\sum_{k=1}^{k_x}|e(\mathcal{P}_x^k,\mathcal{P}_x^k)|}{\sum_{k=1}^{k_y}W^{\alpha_{\text{sys}}}(e(\mathcal{P}_y^k,\mathcal{N}),a^*) + \sum_{k=1}^{k_x}W^{\alpha_{\text{sys}}}(e(\mathcal{P}_x^k,\mathcal{P}_x^k),a^*)} \tag{49}$$

where $a^* = \arg\max_{a\in\mathcal{A}_F} W(a)$ is the welfare-maximizing action profile. Similar to (CY), each y-partition $\mathcal{P}_y^k$ formed from $\hat{a}$ satisfies the stability condition

$$|e(\mathcal{P}_y^k,\mathcal{P}_y^k)| \ge \max_{a_{\mathcal{P}_y^k}\neq\vec{y}} W^\alpha(e(\mathcal{P}_y^k,\mathcal{N}),(a_{\mathcal{P}_y^k},\hat{a}_{-\mathcal{P}_y^k})). \tag{CYE}$$

To reduce cumbersome notation, it is understood the max is taken over actions of unfixed nodes, $a_{\mathcal{P}_y^k\backslash F}$. Likewise, each x-partition $\mathcal{P}_x^k$ satisfies

$$(1+\alpha)|e(\mathcal{P}_x^k,\mathcal{P}_x^k)| \ge \max_{a_{\mathcal{P}_x^k}\neq\vec{x}} W^\alpha(e(\mathcal{P}_x^k,\mathcal{N}),(a_{\mathcal{P}_x^k},\hat{a}_{-\mathcal{P}_x^k})). \tag{CXE}$$

The following lemma asserts that agents playing y in the SSS under the gain α remain playing y under a lower gain $\alpha' < \alpha$. The result is crucial for establishing a lower bound on efficiency for any graph G with arbitrary fixed set $F \in \mathcal{F}(G)$.

Lemma 5: Suppose $\alpha' < \alpha$. Denote $\hat{a}' = \arg\min_{a\in\text{LLL}(\mathcal{A}_F,\alpha';G)} W(a)$ as the welfare-minimizing SSS under α'. Then, for any y-partition $\mathcal{P}_y$ induced from α, $\hat{a}_i' = y$ for all $i \in \mathcal{P}_y\backslash F$.

Proof: Condition (CYE) asserts for all $a_{\mathcal{P}_y} \neq \vec{y}$ that

$$W^\alpha(e(\mathcal{P}_y,\mathcal{N}),(\vec{y}_{\mathcal{P}_y},\hat{a}_{-\mathcal{P}_y})) \ge W^\alpha(e(\mathcal{P}_y,\mathcal{N}),(a_{\mathcal{P}_y},\hat{a}_{\mathcal{P}_y})). \tag{50}$$

It also holds for all $a_{\mathcal{P}_y} \neq \vec{y}$ and for any $a_{-\mathcal{P}_y} \neq \hat{a}_{-\mathcal{P}_y}$ that

$$W^\alpha(e(\mathcal{P}_y,\mathcal{N}),(\vec{y}_{\mathcal{P}_y},a_{-\mathcal{P}_y})) \ge W^\alpha(e(\mathcal{P}_y,\mathcal{N}),(a_{\mathcal{P}_y},a_{-\mathcal{P}_y})) \tag{51}$$

because any y-links garnered in the RHS abovementioned by changing $\hat{a}_{-\mathcal{P}_y}$ to $a_{-\mathcal{P}_y}$ also contribute to the LHS. In particular, the abovementioned equation holds for $a_{-\mathcal{P}_y} = \hat{a}'_{-\mathcal{P}_y}$. Lowering the gain to α' preserves the abovementioned inequality as well, as it devalues x-links garnered on the RHS. ■

A dual statement holds—agents playing x in the SSS under α remain so under a higher gain $\alpha' > \alpha$.

Lemma 6: Suppose $\alpha' > \alpha$. Then, for any x-partition $\mathcal{P}_x$ induced from α, $\hat{a}'_i = x$ for all $i \in \mathcal{P}_x \backslash F$.

We omit the proof for brevity, as it is analogous to the proof of Lemma 5. We are now ready to prove Theorem 2.

Proof of Theorem 2: Consider any graph $G \in \mathcal{G}_N$ with fixed set F. Recall that efficiency is one for $\alpha = \alpha_{\text{sys}}$. Thus, we first consider $\alpha < \alpha_{\text{sys}}$. Observe that

$$\begin{aligned} |e(\mathcal{P}_y^k, \mathcal{P}_y^k)| &\geq W^\alpha(e(\mathcal{P}_y, \mathcal{N}), (a^*_{\mathcal{P}_y}, \hat{a}_{-\mathcal{P}_y})) \\ &= W^\alpha(e(\mathcal{P}_y, \mathcal{N}), (a^*_{\mathcal{P}_y}, a^*_{-\mathcal{P}_y})) \end{aligned} \tag{52}$$

where the inequality is due to (CYE). The equality results from Lemma 6—the agents ($\notin \mathcal{P}_y$) that neighbor any member of $\mathcal{P}_y$ remain playing x in a^*. We, then, obtain

$$\frac{|e(\mathcal{P}_y^k, \mathcal{P}_y^k)|}{W(e(\mathcal{P}_y^k, \mathcal{N}), a^*)} \geq \frac{1+\alpha}{1+\alpha_{\text{sys}}}. \tag{53}$$

The inequality results since the expressions of the numerator and denominator garner the same edges for welfare. It occurs with equality if and only if $a_i^* = x \ \forall i \in \mathcal{P}_y^k \backslash F$. Applying the mediant inequality (34) to (49), $J_f(\alpha, F; G) \geq \frac{1+\alpha}{1+\alpha_{\text{sys}}}$. The case when $\alpha > \alpha_{\text{sys}}$ follows analogous arguments. From Lemma 5, $|e(\mathcal{P}_y^k, \mathcal{P}_y^k)| = W^\alpha(e(\mathcal{P}_y^k, \mathcal{P}_y^k), a^*)$. For x-partitions

$$\begin{aligned} \frac{(1+\alpha_{\text{sys}})|e(\mathcal{P}_x^k, \mathcal{P}_x^k)|}{W(e(\mathcal{P}_x^k, \mathcal{N}), a^*)} &\geq \frac{1+\alpha_{\text{sys}}}{1+\alpha} \frac{W^\alpha(e(\mathcal{P}_x^k, \mathcal{N}), a^*)}{W(e(\mathcal{P}_x^k, \mathcal{N}), a^*)} \\ &\geq \frac{1+\alpha_{\text{sys}}}{1+\alpha} \end{aligned} \tag{54}$$

where the first inequality is from (CXE) and the second occurs with equality if $a_i^* = y \ \forall i \in \mathcal{P}_x^k \backslash F$. From (34) and (49), $J_f(\alpha, F; G) \geq \frac{1+\alpha_{\text{sys}}}{1+\alpha}$. ■

We have just shown fundamental lower bounds on efficiency for any graph with fixed agents. The bounds are tight as they can be achieved for any gain α by arranging N_x fixed x and N_y fixed y leaf nodes that influence a single unfixed agent in the center of a star graph. If $\alpha < \alpha_{\text{sys}}$, choosing $\frac{N_x}{N_y} = \frac{1}{1+\alpha}$ gives the minimal efficiency $\frac{1+\alpha}{1+\alpha_{\text{sys}}}$. If $\alpha > \alpha_{\text{sys}}$, choosing $\frac{N_x}{N_y} = \frac{1}{1+\alpha}$ gives the minimal efficiency $\frac{1+\alpha_{\text{sys}}}{1+\alpha}$. Note that if α is rational, one could choose finite integers N_y, N_x that achieve such ratios. Recall Fig. 2(b) for illustrative examples. However, if it is irrational, they must be taken arbitrarily large to better approximate the ratio.

VI. Proof of Theorem 3: Worst-Case Risks Under Randomized Operator Designs

Recall a randomized strategy consists of gains $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^M$ with distribution $\boldsymbol{p} \in \Delta_M$. The gains are ordered $\alpha_{\text{sys}} \leq \alpha_1 < \cdots < \alpha_M$. To prove Theorem 3, we outline a few technical Lemmas. The key insight is the expected efficiency of any graph G can be expressed in the form $\sum_{i=1}^M p_i s_i$, where the coefficient s_i is a mediant sum over local efficiencies of partitions in G when gain α_i is used. The following two mathematical facts are the basis of this insight.

Fact 2: Let $\nu_i < \frac{n_i}{d_i} \leq 1$ with $r_i \geq 0$ and $n_i, d_i > 0$ for all $i = 1, \ldots, M$. Then, for all $\boldsymbol{p} \in \Delta_M$

$$\sum_{i=1}^M p_i s_i \geq 1 + \min_{i=1,\ldots,M} \left\{ \left(\sum_{j=1}^i p_j \right) (\nu_i - 1) \right\} \tag{55}$$

where $s_i := \frac{\sum_{j=1}^{i-1} d_j + \sum_{j=i}^M n_j}{\sum_{j=1}^M d_j}, \quad i = 1, \ldots, M.$

We provide a proof in the Appendix. The following dual result follows directly.

Fact 3: For all $\boldsymbol{p} \in \Delta_M$

$$\sum_{i=1}^M p_i s'_i \geq 1 + \min_{i=1,\ldots,M} \left\{ \left(\sum_{j=i}^M p_j \right) (\nu_i - 1) \right\} \tag{56}$$

where $s'_i := \frac{\sum_{j=1}^{i} n_j + \sum_{j=i}^M d_j}{\sum_{j=1}^M d_j}, \quad i = 1, \ldots, M.$

Proof: The proof follows similarly to Fact 2, where the indices of the s_i coefficients are reversed. ■

We will show for any (S, G) that $\mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[J_b|S, G] = 1 - \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[R_b|S, G]$ can be expressed in the form $\sum_{i=1}^M p_i s_i$ from the LHS of (55). The lower bounds establish worst-case expected efficiencies—and, hence risks. The ν_i correspond to the worst-case deterministic efficiencies $J_b^*(\alpha_i) = 1 - R_b^*(\alpha_i)$ of the M gains and $\frac{n_i}{d_i}$ to local efficiencies of selected partitions in the graph. Fact 2 will be used to establish (22), and Fact 3 for (23) (see Theorem 3). We now identify a structural property required of worst-case graphs.

Lemma 7: A worst-case graph, i.e., a member of $\arg\min_{G \in \mathcal{G}_N, S \in \mathcal{T}(G)} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[J_{\text{b}}|S, G]$, has no active x-links in α_1.

Proof: Any active x-links in α_1 remain so for all $\{\alpha_i\}_{i=2}^M$. The efficiency corresponding to each gain can be reduced in the following manner. Delete all such x-links and associated agents. For each miscoordinating link between an x and y agent that existed, replace with a single link to a newly created isolated agent with an x-impostor attached. This preserves the stochastically stable states of all other nodes while reducing efficiency in each gain. ■

Intuitively, a graph that has coordinating x nodes in each gain $\alpha_1, \ldots, \alpha_M$ can be modified by removing these links, resulting in a lower efficiency. We are now ready to prove (22) (see Theorem 3).

Proof of (22) (see Theorem 3): Consider any graph $G = (\mathcal{N}, \mathcal{E}) \in \mathcal{G}_N$ and $S \in \mathcal{T}(G)$. Let us denote the M (worst-case) stochastically stable states that correspond to each gain α_i with $\hat{a}^i$. Define for each $k = 1, \ldots, M$

$$P^k = \{n \in \mathcal{N} : \hat{a}_n^i = y \ \forall i \leq k, \ \hat{a}_n^i = x \ \forall i > k\} \tag{57}$$

as the set of nodes that play y in the SSS in $\alpha_1, \ldots, \alpha_k$ and play x in $\alpha_{k+1}, \ldots, \alpha_M$. Note that P^k is possibly composed of multiple y-partitions. Also note it is possible that $P^k = \varnothing$ for all $k > \bar{m}$ for some $\bar{m} \in \{2, \ldots, M-1\}$, i.e., $\hat{a}^i = \vec{x}$ for all $i > \bar{m}$. We first consider the case when $P^k \neq \varnothing$ for every $k = 1, \ldots, M$.

Let $Q^k := \{n \in \mathcal{N} : \hat{a}_n^k = y\} = \bigcup_{i=k}^M P^i$. Denote $P^x := \{n \in \mathcal{N} : \hat{a}_n^1 = x\} = (Q^1)^c$ as the set of nodes stable to x for all α_i. Consider the gain α_i with $i \leq k$. Then, the local efficiency $\frac{W(e(P^k,\mathcal{N}),\hat{a}^k)}{W(e(P^k,\mathcal{N}),(a^*_{P^k},\hat{a}^k_{-P^k}))}$ of P^k is

$$\frac{|e(P^k,Q^{k+1})| + |e(P^k,P^k)|}{(1+\alpha_{\text{sys}})(|e(P^k,P^k)| + |e(P^k,P^x)| + |e(P^k,(Q^k)^c)|)} > J_A^*(\alpha_i). \qquad (58)$$

The inequality is due to Proposition 1. For gains α_i with $i > k$, the local efficiency of P^k is

$$\frac{(1+\alpha_{\text{sys}})(|e(P^k,P^k)|+|e(P^k,P^x)|+|e(P^k,(Q^k)^c)|)}{(1+\alpha_{\text{sys}})(|e(P^k,P^k)|+|e(P^k,P^x)|+|e(P^k,(Q^k)^c)|)} = 1. \qquad (59)$$

Hence, the overall system efficiency under gain α_i is the mediant sum of the local efficiencies of the P^k. An application of Fact 2 gives the result. The case when $P^k = \varnothing$ for $k > \bar{m} \in \{2, \ldots, M-1\}$ also follows directly from Fact 2. From the notation of Fact 2, $\frac{n_k}{d_k} = 1$ for $k > \bar{m}$. ■

The details for the proof of (22) (see Theorem 3) follow analogous arguments pertaining to focused attacks. Recall for a graph $G \in \mathcal{G}_n$ and restricted action set $\mathcal{A}$, we denote $F = F_x \cup F_y \subset \mathcal{N}$ as its set of fixed nodes. Additionally, we restrict attention to gains $\alpha_i \geq \alpha_{\text{sys}}$, as these are not strictly dominated in the risk curve. The following structural property holds in a worst-case graph for focused risk.

Lemma 8: A worst-case graph, i.e., a member of $\arg\min_{G \in \mathcal{G}_N, F \in \mathcal{F}(G)} \mathbb{E}_{\boldsymbol{\alpha},\boldsymbol{p}}[J_{\text{f}}|F,G]$, has no active y-links in α_M. Additionally, $a^*_{F^c} = \vec{y}$.

Proof: A graph that has active y-links in α_M remain active for all $\alpha_1, \ldots, \alpha_{M-1}$. The efficiency corresponding to each gain can be reduced by removing all such links and keeping the border nodes as fixed y agents. This preserves the stability properties of all other nodes. The claim $a^*_{F^c} = \vec{y}$ follows from Lemma 5. ■

We are now ready to prove (23) in Theorem 3.

Proof of (23) (see Theorem 3): Consider any graph $G = (\mathcal{N}, \mathcal{E}) \in \mathcal{G}_N$ and fixed nodes $F \in \mathcal{F}(G)$. The M stochastically stable states that correspond to each gain α_i are denoted $\hat{a}^i$. Define for each $k = 1, \ldots, M$

$$P^k = \{n \in F^c : \hat{a}_n^i = x \; \forall i \geq k, \; \hat{a}_n^i = y \; \forall i < k\} \qquad (60)$$

as the set of unfixed nodes that play x in the SSS for $\alpha_k, \ldots, \alpha_M$ and play y in $\alpha_1, \ldots, \alpha_{k-1}$. Note that it is possible $P^k = \varnothing$ for all $k < \bar{m}$ for some $\bar{m} \in \{2, \ldots, M-1\}$. That is, $a^k_{F^c} = \vec{y}$ for $k = 1, \ldots, \bar{m}-1$. We first consider the case when $P^k \neq \varnothing$ for every $k = 1, \ldots, M$.

Let $Q^k = \{n \in F^c : \hat{a}_n^k = y\} = \bigcup_{i=k}^M P^i$. Consider the gain α_i with $i \geq k$. Then, the local efficiency $\frac{W(e(P^k,\mathcal{N}),\hat{a}^k)}{W(e(P^k,\mathcal{N}),(a^*_{P^k},\hat{a}^k_{-P^k}))}$ of P^k is

$$\frac{(1+\alpha_{\text{sys}})(|e(P^k,P^k)| + |e(P^k,(Q^{k-1})^c)| + |e(P^k,F_x)|)}{|e(P^k,P^k)| + |e(P^k,Q^k)| + |e(P^k,F_y)|} > J_f^*(\alpha_i). \qquad (61)$$

Here, we use the convention $|e(P^1,(Q^0)^c)| = 0$. For gains α_i with $i < k$, the local efficiency of P^k is

$$\frac{|e(P^k,P^k)| + |e(P^k,Q^k)| + |e(P^k,F_y)|}{|e(P^k,P^k)| + |e(P^k,Q^k)| + |e(P^k,F_y)|} = 1. \qquad (62)$$

Hence, the overall system efficiency under α_i is the mediant sum of the local efficiencies of the P^k. An application of Fact 3 gives the result. The case when $P^k = \varnothing$ for $k > \bar{m} \in \{2, \ldots, M-1\}$ also follows directly from Fact 3. ■

VII. Summary

In this article, we framed graphical coordination games as a distributed system subject to two types of adversarial influences. The focus of our article concerned the performance of a class of distributed algorithms against the associated worst-case risks. We identified fundamental tradeoffs between ensuring security against one type of risk and vulnerability to the other, and vice versa. Furthermore, our analysis shows randomized algorithmic designs significantly improves the available tradeoffs. This article highlights the design challenges a system operator faces in maintaining the efficiency of networked, distributed systems.

Appendix

Proof of Lemma 3: This proof outlines a procedure to transform any (S,G) into a star graph with lower efficiency if $\alpha < 1$. We split into two cases—either (S,G) induces a single y-partition or more than one. First, assume (S,G) induces a single y-partition $\mathcal{P}_y$. An illustration of the constructive process is shown in Fig. 4.

Construct a star subgraph Γ_y that has $1 + |e(\mathcal{P}_y,\mathcal{P}_y)|$ nodes, each having a y impostor attached. Call the center node i_y. Construct similar star configurations Γ_x^k for each x-partition $\mathcal{P}_x^k$. Call their center nodes i_x^k. Connect Γ_y to each Γ_x^k with a link between i_x^k and i_y. If there are multiple edges between $\mathcal{P}_x^k$ and $\mathcal{P}_y$ ($|e(\mathcal{P}_x^k,(\mathcal{P}_x^k)^c)| \geq 2$), create $|e(\mathcal{P}_x^k,(\mathcal{P}_x^k)^c)| - 1$ new isolated nodes with a single x impostor attached, and connect each to i_y with a single link. At this point, Γ_y and Γ_x^k are stable y and x-partitions, and the isolated nodes are stable playing x. We have obtained a graph of $N' \geq N$ nodes with identical efficiency to (S,G) since the number and type of edges are preserved.

We can further reduce efficiency if there are active x links, i.e., if $|e(\Gamma_x^k,\Gamma_x^k)| \geq 1$ for at least one Γ_x^k. If there are none, then the graph belongs to $\mathbb{S}_{N'}$ and we are done. Otherwise for each leaf node $j \in \mathcal{P}_x^k$, redirect the edge (j, i_x^k) to (j, i_y), and replace j's x impostor with a y impostor. Call m_x the total number of such converted nodes. The resulting graph-target pair (S',G') belongs to $\mathbb{S}_{N'}$. We claim the resulting (larger) y-partition Γ_y' is stable. For this claim to hold, (45) requires that

$$2|e(\mathcal{P}_y,\mathcal{P}_y)| + 2m_x + 1 \geq (1+\alpha)(|e(\mathcal{P}_y,\mathcal{N})| + m_x). \qquad (63)$$

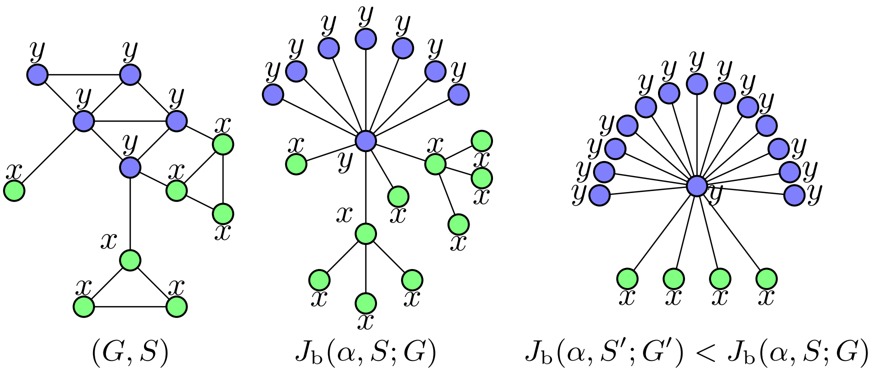


Fig. 4. Illustration of the constructive process (proof of Lemma 3) that generates a member $(S',G') \in \mathbb{S}_m$ from any graph (S,G) with one y-partition, and $\alpha < 1$. Here, the labels on each node indicate the type of impostor influence. Green (blue) nodes play x (y) in the SSS. (Left) Start with an arbitrary graph-adversary pair (S,G). (Center) The partitions of (S,G) are recast as star subgraphs with the same number of edges. When there is more than one edge between the y and an x-partition, new nodes are created for the excess outgoing edges. This recasting preserves the original efficiency $J_b(\alpha,S;G)$. (Right) The active x-links are converted into y-links by redirecting them to the center of the y-partition. This results in a graph $(S',G') \in \mathbb{S}_m$.

From the original $\mathcal{P}_y$, it holds that

$$\begin{aligned}&|e(\mathcal{P}_y,\mathcal{P}_y)| + |\mathcal{P}_y| \ge (1+\alpha)|e(\mathcal{P}_y,\mathcal{N})|\\ &\Rightarrow 2|e(\mathcal{P}_y,\mathcal{P}_y)| + 2m_x + 1 > (1+\alpha)(|e(\mathcal{P}_y,\mathcal{N})| + m_x)\end{aligned} \tag{64}$$

due to $|\mathcal{P}_y| \le 1 + |e(\mathcal{P}_y,\mathcal{P}_y)|$ and $\alpha < 1$. All x-partitions in (S',G'), now just a collection of single nodes connected to i_y with an x-impostor, are stable. The efficiency is less than the original because active x-links increase efficiency more than active y-links do. Hence

$$J_b(\alpha,S;G) > J_b(\alpha,S';G'). \tag{65}$$

Now, we consider the remaining case when (S,G) induces $k_y > 1$ y-partitions $\{\mathcal{P}_y^k\}_{k=1}^{k_y}$ and $k_x \ge 1$ x-partitions $\{\mathcal{P}_x^k\}_{k=1}^{k_x}$. Consider k_y such star subgraphs $\{\Gamma_y^k\}_{k=1}^{k_y}$ with center nodes i_y^k. Recast the x-partitions into similar star subgraphs $\{\Gamma_x^k\}_{k=1}^{k_x}$ with center nodes i_x^k. We first connect each Γ_x^k to some Γ_y^j with a single link (i_x^k, i_y^j) in any manner as long as a link between the original $\mathcal{P}_x^k$ and $\mathcal{P}_y^j$ exists. For each excess outgoing edge, we create an isolated node with an x-impostor attached. Each isolated node is attached to a corresponding i_y^k such that the original number of outgoing edges for each $\mathcal{P}_y^k$ is satisfied. At this point, there are $k' \le k_y$ connected components G_k in the construction, and the efficiency of this construction is identical to the original. Finally, we apply the efficiency reduction procedure from before for each G_k to obtain $(S'_k, G'_k) \in \mathbb{S}_{m_k}$. From (37) and (34) we have

$$\begin{aligned}J_b(\alpha,S;G) &> \frac{\sum_{k=1}^{k'} |e(\Gamma_y^k,\Gamma_y^k)|}{(1+\alpha_{\rm sys})\sum_{k=1}^{k'} |e(\Gamma_y^k,\mathcal{N})|}\\ &\ge \min_{k=1,\dots,k'} J_b(\alpha,S'_k;G'_k).\end{aligned} \tag{66}$$

■

Proof of Lemma 2. Technical result for expected risks: Let us define $f(\boldsymbol{p}) := \sum_{i=1}^M p_i s_i$ and $f_i(\boldsymbol{p}) := (\sum_{j=1}^k p_j)(\nu_k - 1) + 1$. The set of probability vectors such that $k = \arg\ \min_{i=1,\dots,M} f_i(\boldsymbol{p})$ can be written as the set

$$\begin{aligned}V_k &:= \{\boldsymbol{p} \in \Delta_M : f_k(\boldsymbol{p}) \le f_\ell(\boldsymbol{p})\ \forall \ell \ne k\}\\ &= \bigcap_{\ell \ne k} \left\{\boldsymbol{p} \in \Delta_M : \frac{\sum_{j=1}^k p_j}{\sum_{j=1}^\ell p_j} \ge \frac{1-\nu_\ell}{1-\nu_k}\right\}.\end{aligned} \tag{67}$$

Define $\lambda_k = \frac{\sum_{j=1}^k d_j}{\sum_{j=1}^{k+1} d_j}$ for each $k = 1,\dots,M-1$. With some algebra, we can express each s_i as

$$s_i = \left[\sum_{j=1}^{M-i+1} \frac{n_j}{d_j}(1-\lambda_{j-1})\left(\prod_{k=j}^{M-1} \lambda_k\right)\right] + \left(1 - \prod_{j=M-i+1}^{M-1} \lambda_j\right). \tag{68}$$

Using the identities $\sum_{k=1}^M (1-\lambda_{k-1})(\prod_{j=k}^{M-1} \lambda_j) = 1$ and $\sum_{k=1}^\ell (1-\lambda_{k-1})(\prod_{j=k}^{M-1} \lambda_j) = \prod_{j=\ell}^{M-1} \lambda_j$ for $\ell = 1,\dots,M-1$, we obtain (omitting the algebraic steps)

$$f(\boldsymbol{p}) = \sum_{i=1}^M (1-\lambda_{i-1})\left(\prod_{j=i}^{M-1} \lambda_j\right)\left[\left(\frac{n_i}{d_i} - 1\right)\left(\sum_{j=1}^{M-i+1} p_j\right) + 1\right]. \tag{69}$$

Now, suppose $\boldsymbol{p} \in V_k$ for $k \in \{1,\dots,M\}$. Using (67) and $\nu_i \le n_i/d_i \le 1$, we then have

$$f(\boldsymbol{p}) \ge \sum_{i=1}^M \lambda_{i-1}\left(\prod_{j=i}^{M-1} \lambda_j\right) f_k(\boldsymbol{p}) = f_k(\boldsymbol{p}). \tag{70}$$

■

References

[1] K. Paarporn, M. Alizadeh, and J. R. Marden, "Risk and security tradeoffs in graphical coordination games," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 4409–4414.

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

[3] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton, NJ, USA: Princeton Univ. Press, 2010.

[4] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[5] H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ, USA: Princeton Univ. Press, 2001.

[6] A. Montanari and A. Saberi, "The spread of innovations in social networks," *Proc. Nat. Acad. Sci.*, vol. 107, no. 47, pp. 20196–20201, 2010.

[7] S. A. West, S. P. Diggle, A. Buckling, A. Gardner, and A. S. Griffin, "The social lives of microbes," *Annu. Rev. Ecology, Evolution, Systematics*, vol. 38, no. 1, pp. 53–77, 2007.

[8] M. Del Vicario *et al.*, "The spreading of misinformation online," *Proc. Nat. Acad. Sci.*, vol. 113, no. 3, pp. 554–559, 2016.

[9] Y. Mao, S. Bolouki, and E. Akyol, "Spread of information with confirmation bias in cyber-social networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 688–700, Apr.–Jun. 2020.

[10] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*, R. Majumdar and P. Tabuada, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 31–45.

[11] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Stealthy deception attacks on water scada systems," in *Proc. 13th ACM Int. Conf. Hybrid Syst., Comput. Control*, 2010, pp. 161–170.

[12] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.*, 2011, pp. 337–344.

[13] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[14] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Trans. Autom. Control*, vol. 64, no. 3, pp. 1063–1076, Mar. 2019.

[15] V. Auletta, D. Ferraioli, F. Pasquale, and G. Persiano, "Metastability of logit dynamics for coordination games," in *Proc. 23rd Annu. ACM-SIAM Symp.*, 2012, pp. 1006–1024.

[16] L. E. Blume, "The statistical mechanics of best-response strategy revision," *Games Econ. Behav.*, vol. 11, no. 2, pp. 111–145, 1995.

[17] J. R. Marden and J. S. Shamma, "Revisiting log-linear learning: Asynchrony, completeness and payoff-based implementation," *Games Econ. Behav.*, vol. 75, no. 2, pp. 788–808, 2012.

[18] T. Tatarenko, "Proving convergence of log-linear learning in potential games," in *Proc. Amer. Control Conf.*, 2014, pp. 972–977.

[19] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi, "Spread of (mis) information in social networks," *Games Econ. Behav.*, vol. 70, no. 2, pp. 194–227, 2010.

[20] J. Ghaderi and R. Srikant,"Opinion dynamics in social networks with stubborn agents: Equilibrium and convergence rate," *Automatica*, vol. 50, no. 12, pp. 3209–3215, 2014.

[21] P. N. Brown, H. Borowski, and J. R. Marden, "Security against impersonation attacks in distributed systems," *IEEE Trans. Control Netw. Syst.*, vol. 6, no. 1, pp. 440–450, Mar. 2019.

[22] B. Canty, P. N. Brown, M. Alizadeh, and J. R. Marden, "The impact of informed adversarial behavior in graphical coordination games," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 1923–1928.

[23] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. 57–84, 1993.

[24] D. Foster and H. P. Young, "Stochastic evolutionary game dynamics," *Theor. Population Biol.*, vol. 38, no. 2, pp. 219–232, 1990.

[25] D. Monderer and L. S. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, no. 1, pp. 124–143, 1996.

Keith Paarporn received the B.S. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2013, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2016 and 2018, respectively.

He is currently a Postdoctoral Scholar with the Department of Electrical and Computer Engineering, the University of California, Santa Barbara. His research interests include game theory and the control and analysis of multiagent systems.

Mahnoosh Alizadeh received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2009, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of California Davis, Davis, CA, USA, in 2013 and 2014, respectively.

She is an Assistant Professor of electrical and computer engineering with the University of California Santa Barbara. From 2014 to 2016, she was a Postdoctoral Scholar with Stanford University. Her research interests include designing scalable control, learning, and market mechanisms to promote efficiency and resiliency in societal-scale cyber-physical systems.

Dr. Alizadeh is a recipient of the NSF CAREER Award and the Best Paper Award from HICSS-53 power systems track. She was the recipient of the Richard C. Dorf Award for outstanding research accomplishment.

Jason R. Marden received the B.S. and Ph.D. degrees in mechanical engineering from the University of California, Los Angeles, Los Angeles, CA, USA, in 2001 and 2007, respectively.

He is an Associate Professor in the Department of Electrical and Computer Engineering, University of California, Santa Barbara. After graduating from UCLA, he was a junior Fellow with the Social and Information Sciences Laboratory, the California Institute of Technology until 2010 when he joined the University of Colorado. His research interest includes game theoretic methods for the control of distributed multiagent systems.

Dr. Marden was awarded the Outstanding Graduating Ph.D. Student in mechanical engineering with UCLA. He has been recipient of the NSF Career Award in 2014, the ONR Young Investigator Award in 2015, the AFOSR Young Investigator Award in 2012, the American Automatic Control Council Donald P. Eckman Award in 2012, and the SIAG/CST Best SICON Paper Prize in 2015.