Design of Attack-Resilient Consensus Dynamics: A Game-Theoretic Approach

Mohammad Pirani and Henrik Sandberg

***Abstract*— We propose a game-theoretic framework for improving the resilience of multi-agent consensus dynamics in the presence of a strategic attacker. In this game, the attacker selects a set of network nodes to inject the attack signals. The attacker's objective is to minimize the required energy for steering the consensus towards its desired direction. This energy is captured by the trace of controllability Gramian of the system when the input is the attack signal. The defender improves the resilience of dynamics by adding self-feedback loops to certain nodes of the system and its objective is to maximize the trace of controllability Gramian. The Stackelberg equilibrium of the game is studied with the defender as the game leader. When the underlying network topology is a tree and the defender can select only one node, we show that the optimal strategy of the defender is determined by a specific distance-based network centrality measure, called network's f-center. In addition, we show that the degree-based centralities solutions may lead to undesirable payoffs for the defender. At the end, we discuss the case of multiple attack and defense nodes on general graphs.**

I. Introduction

A. Motivation

As the scale of network control systems becomes larger and interactions between different parts become more sophisticated, vulnerability of the system to unexpected faults and attacks comes as a side effect. There exists a vast literature on various methods to encounter, in both forms of mitigation or bypassing, such unexpected and undesirable events which are mostly referred to as fault-tolerant or resilient distributed control [1]–[4]. However, the subtle difference between faults and attacks is that in the latter case, the attacker uses a knowledge she has about the vulnerabilities of the victim system and targets its attack in such a way to maximize its effect and/or minimize its visibility or effort to attack [5], [6]. When the attacker uses such a level of intelligence to steer the system to its desired direction, the defending mechanism has to adopt an intelligent strategy to encounter the attacker. One of the approaches to model such battles between intelligent attacker and defender is via game theory. In this direction, a large body of research is dedicated to discuss game-theoretic methods to the security of cyber-physical systems . A branch of such research pertains to study the effect of the network structure on the game value and the equilibrium strategies [7], [8].

The role of network structure on the robustness, resilience, and fault tolerance of network control systems has been investigated in the past decade [9], [10]. Most of such works belong to characterizing the effect of network connectivity in damping the disturbances or faults, which are sometimes referred to as *network coherence*. However, some researches have been dedicated to identify the most effective nodes (or a set of nodes) for placing controllers to maximize the robustness of the network control system. The counterpart of this problem is to look at these effective nodes as the most vulnerable nodes to be attacked. In both cases, the key approach is to relate such nodes to specific well-known network centrality metrics [11], [12]. In this direction, our approach in this paper is to relate the best actions of the defender (and consequently the attacker) to some network centrality metrics.

M. Pirani and H. Sandberg are with the Department of Automatic Control, KTH Royal Institute of Technology. E-mail: {pirani,hsan}@kth.se.

B. Related Work

There is a vast literature on resilient distributed algorithms in the presence of adversarial agents [13]. Among them, resilient consensus has attracted attentions in recent years. One approach to resilient consensus problems was to obtain (or recover) the initial values of all agents in the network (despite the actions of malicious agents) and then compute the function of initial values [10], [14]. The other approach is to bypass the effects of malicious agents while doing the averaging [15], [16]. The former method reaches to the exact averaging of the initial conditions; however, demands large computational cost. The latter requires much less computational cost; however, it only guarantees that the final value will be in a convex hull of initial conditions (not necessarily the average). Both approaches, however, require a large level of connectivity for the underlying interaction network. In many of the real-world applications of multi-agent systems, e.g., networks of power generators, the underlying topology is given and can not be changed. Hence, if the network connectivity does not satisfy the requirements of the above-mentioned resilient consensus algorithms, some other methods to overcome the actions of the malicious actions have to be proposed.

C. Contributions

In this paper, we discuss a specific resilient distributed consensus algorithm based on a game between the attacker (which tries to steer the consensus to its desired direction with minimum energy) and a defender (which tries to maximize this energy). More specifically, the contributions of the paper are:

- We introduce an attacker-defender zero-sum game in consensus dynamics of multi-agent systems where the game payoff is the trace of the controllability Gramian (which captures the average energy needed to steer the

system over all controllable subspace). Moreover, we discuss the Stackelberg game between the two players when the defender is chosen as the leader.
- For the cases of single defender and f attackers, $f \geqslant 1$, when the underlying network is a tree, we show that the solution of the Stackelberg game for the defender implies a specific network centrality, called f-center of the graph. Moreover, we show that degree-based centralities can exhibit bad performance if they are chosen by the defender.
- We discuss these results to the general case of multiple defense and attacked nodes on general weighted undirected graphs.

II. System Model and Preliminaries

A. Notations and Definitions

We use $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ to denote a weighted undirected graph where $\mathcal{V}$ is the set of vertices (or nodes) and $\mathcal{E}$ is the set of undirected edges where $(v_i, v_j) \in \mathcal{E}$ if an only if there exists a weighted undirected edge between v_i and v_j. Let $|\mathcal{V}| = n$ and define the adjacency matrix for $\mathcal{G}$, denoted by $\mathcal{A}_{n\times n}$, to be a matrix where $\mathcal{A}_{ij} = w_{ij}$ if and only if there is an edge with weight w_{ij} between v_j and v_i in $\mathcal{G}$ (the adjacency matrix will be a symmetric matrix when the graph is undirected). The *neighbors* of vertex $v_i \in \mathcal{V}$ in the graph $\mathcal{G}$ are denoted by the set $\mathcal{N}_i = \{v_j \in \mathcal{V} \mid (v_j, v_i) \in \mathcal{E}\}$. We define the degree for node v_i as $d_i = \sum_{v_j \in \mathcal{N}_i} \mathcal{A}_{ij}$. The Laplacian matrix of an undirected graph is denoted by $L = D - \mathcal{A}$, where $D = \mathrm{diag}(d_1, d_2, ..., d_n)$. We use $\mathbf{e}_i$ to indicate the i-th vector of the canonical basis. The eccentricity $\epsilon(v)$ of a vertex v in a connected weighted graph $\mathcal{G}$ is the maximum graph distance (or weighted distance) between v and any other vertex u of $\mathcal{G}$. The center of a graph is a set of vertices with minimum (weighted) eccentricity. The f-eccentricity, $\epsilon_f(v)$, of a vertex v in a connected weighted graph $\mathcal{G}$ is the maximum sum of graph distances (or weighted distances) between v and any combination of f vertices $u_1, u_2, ..., u_f$ in $\mathcal{G}$. The f-center of a graph is a set of vertices with minimum (weighted) f-eccentricity. The *effective resistance*, R_{ij}, between two vertices v_i and v_j in a graph is the equivalent resistance between these two vertices when we treat the resistance of each edge e as $\frac{1}{w_e}$, where w_e is its weight.

B. Consensus Model

Consider a connected undirected network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. The state of each agent $v_j \in \mathcal{V}$ evolves based on the interactions with its neighbors as

$$\dot{x}_i(t) = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_j(t) - x_i(t)\right), \tag{1}$$

where $w_{ij} > 0$ is a positive weight representing the communication strength. There are some agents who add a feedback from their own state and initial condition and evolve as[1]

$$\dot{x}_i(t) = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_j(t) - x_i(t)\right) - k\left(x_i(t) - x_i(0)\right), \tag{2}$$

where k is the gain of the self-feedback.[2] The reason for adding such pure state feedbacks to some agents' dynamics is to make the consensus resilient to attack signals, as will be discussed in the next subsection. Aggregating the dynamics of agents into a matrix form yields the following linear time-invariant networked dynamical system

$$\dot{\boldsymbol{x}}(t) = -\bar{L}\boldsymbol{x}(t) + K\boldsymbol{x}(0), \tag{3}$$

where $\bar{L} = L + K$, $K = kD_y$ and $D_y = \mathrm{diag}(\boldsymbol{y})$ in which $\boldsymbol{y}$ is a binary vector. The i-th element of $\boldsymbol{y}$, *i.e.,* $y_i \in \{0, 1\}$, indicates that node i has a self feedback if $y_i = 1$ and it does not have a self-feedback loop if $y_i = 0$. It is well known that in the absence of self-feedback loops, dynamics (3) converges to the average of the initial conditions [18]. If there exists at least one defense node, then matrix $\bar{L}$ is invertible [19]. The following proposition indicates that in the absence of attack (and when the self feedback loops exist), the above dynamics converges to a convex combination of agents' initial conditions (but not necessarily to the average of initial conditions).

Proposition 1: Each component of the steady-state solution of (3) lies in the convex hull of the agents' initial conditions.

Proof: We have $\bar{L} = L + K$ and by multiplying both sides with vector $\mathbf{1}$, we get $L\mathbf{1} = (\bar{L} - K)\mathbf{1} = \mathbf{0}$ which results in $\bar{L}^{-1}K\mathbf{1} = \mathbf{1}$. This implies that $\bar{L}^{-1}K$ is a row stochastic matrix. For the steady-state solution, $\boldsymbol{x}_{ss}$, of (3) we have $-\bar{L}\boldsymbol{x}_{ss} + K\boldsymbol{x}(0) = \mathbf{0}$ which yields $\boldsymbol{x}_{ss} = \bar{L}^{-1}K\boldsymbol{x}(0)$ and based on the fact the $\bar{L}^{-1}K$ is row stochastic, we conclude that $\boldsymbol{x}_{ss}$ will be some convex combination of the elements of $\boldsymbol{x}(0)$. ■

C. Attack Model

To examine the resilience of the dynamic (2) against cyber-physical attacks, we assume that an attacker injects attack signals to a set of nodes in the network. Let $\mathcal{B}$ denote the set of nodes under attack. Thus, the dynamic of the node $v_i \in \mathcal{B}$ will evolve according to

$$\dot{x}_i(t) = \sum_{j\in\mathcal{N}_i} w_{ij}\left(x_j(t) - x_i(t)\right) - k\left(x_i(t) - x_i(0)\right) + \zeta_i(t), \tag{4}$$

where $\zeta_i(t)$ represents the attack signal on node v_i. In this paper we assume that $\zeta_i(t) = \zeta(t)$, $\forall i \in \mathcal{B}$. Aggregating the dynamics of agents into a matrix form yields the following linear time-invariant network dynamical system

$$\dot{\boldsymbol{x}}(t) = -\bar{L}\boldsymbol{x}(t) + K\boldsymbol{x}(0) + B\boldsymbol{\zeta}(t), \tag{5}$$

where matrix $B = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_f]$ specifies the selected nodes by attacker. Note that the attacker selects the columns of the

[1]This type of state evolution where agents used their initial states as feedbacks is used in opinion dynamics literature [17].

[2]Here, we assume that the gain values k are fixed and same for all agents which have self-feedback.

matrix B and the value of the attack signal $\boldsymbol{\zeta}(t)$ is not a decision variable. We assume that the value of the feedback parameter k is private and is unknown to the attacker.[3]

In the presence of an attack, dynamics (5) can be potentially deviated from the convex hull of the initial conditions, unless the underlying network is sufficiently connected [15], [16]. However, having a highly connected network is a demanding condition for many applications. In this case, we have to adopt an alternative strategy to prohibit or mitigate the effect of the attack even when the underlying network topology is sparse as it will be discussed in the following section.

III. Attacker-Defender Game

In this section, we propose a game-theoretic framework for designing secure consensus algorithms. More formally, we pose the problem as a zero-sum game between an attacker and a defender. The objective of the game is the energy that the attacker requires to steer the consensus to its desired direction. This energy is captured via the spectrum of the controllability Gramian of the system. What we use is the trace of the controllability Gramian which is inversely related to the average energy and can be interpreted as the average controllability in all directions in the state space [20]. The attacker selects a set of network nodes, to inject the attack signals, such that the average energy is minimized (the trace of controllability Gramian is maximized). The defender selects a set of nodes to place self-feedback loops such that this energy is maximized (the trace of controllability Gramian is minimized). Mathematically speaking, the decision variable of the attacker is matrix B and the decision variable of the defender is matrix D_y. The notion of average energy here is captured by the trace of the controllability Gramian matrix, associated with the dynamics in (5), which can be written as

$$\mathcal{W}_c = \int_0^\infty \mathbf{e}^{\bar{L}\tau} B B^T \mathbf{e}^{\bar{L}^T\tau} d\tau. \tag{6}$$

Based on the following calculations, the trace of the controllability Gramian of (5) has a closed form

$$\begin{aligned}\operatorname{tr}(\mathcal{W}_c) &= \operatorname{tr}\left(\int_0^\infty \mathbf{e}^{\bar{L}\tau} B B^T \mathbf{e}^{\bar{L}^T\tau} d\tau\right) \\ &= \int_0^\infty \operatorname{tr}\left(B^T \mathbf{e}^{2\bar{L}\tau} B\right) d\tau = \operatorname{tr}\left(B^T \int_0^\infty \mathbf{e}^{2\bar{L}\tau} d\tau B\right) \\ &= \frac{1}{2}\operatorname{tr}\left(B^T \bar{L}^{-1} B\right) = \sum_{i\in\mathcal{B}} [\bar{L}^{-1}]_{ii}, \end{aligned} \tag{7}$$

where $[\bar{L}^{-1}]_{ii}$ is the i-th diagonal element of $\bar{L}^{-1}$. Based on the above closed form, the following game is defined.

[3]Otherwise the attacker can choose $\zeta_i(t) = k\left(x_i(t) - x_i(0)\right)$ and makes the defender as an ordinary agent.

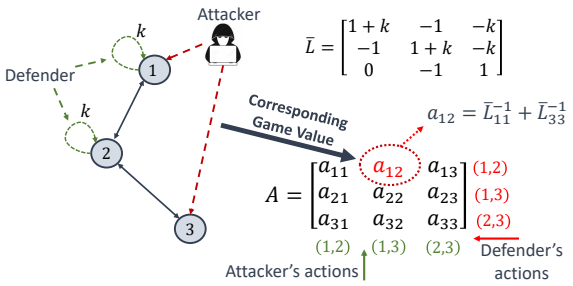

Fig. 1: An example of the attacker-defender game.

Attacker and Defender Game: The attacker tries to inject attack signal $\zeta_i(t)$ to f nodes in the network to maximize the average controllability $\operatorname{tr}(\mathcal{W}_c)$, while the defender places the self feedbakcs to $\bar{f}$ nodes to minimize the average controllability of the system. Thus the game payoff for this attacker-defender zero-sum game is:

$$J(B, D_y) = \operatorname{tr}\left(B^T \underbrace{(L + kD_y)}_{\bar{L}}{}^{-1} B\right), \tag{8}$$

where the attacker's decision determines matrix B to maximize $J(B, D_y)$ and the defender's decision affects matrix D_y to minimize $J(B, D_y)$.

Based on the actions of the attacker and the defender, when f nodes are under attack and $\bar{f}$ nodes are defended, matrix game $A_{\binom{n}{\bar{f}}\times\binom{n}{f}}$ is formed where $A_{ij} = J(B_j, D_{yi})$ in which B_j corresponds to the set chosen by the attacker and D_{yi} corresponds to the set chosen by the defender. The attacker, which acts as the maximizer, chooses columns of matrix A and the defender, the minimizer, chooses the rows. Fig. 1 shows an example of the attacker defender game. Based on this figure, the attacker chooses nodes 1 and 3, while the defender chooses nodes 1 and 2. For this example, the game value is $a_{12} = \bar{L}_{11}^{-1} + \bar{L}_{33}^{-1}$.

Through the rest of the paper, we investigate the equilibrium of the above game for the case of single (and multiple) attacker and defender. Note that the equilibrium strategy of the game determines the optimal location of self-feedback loops.

Remark 1: It is known that optimizing the trace of controllability Gramian does not preserve the controllability of the system, e.g., having $\mathcal{W}_c \succ 0$ [20]. If minimizing $\operatorname{tr}(\mathcal{W}_c)$ (from defender's perspective) yields to an uncontrollable system, then the action of the defender can be interpreted as maximizing the energy of steering the system over the *controllable subspace*, which is the range space of $\mathcal{W}_c$. □

IV. Equilibrium Analysis: Single Attack Single Defense Nodes On Trees

In this section, we investigate the attacker-defender game when the attacker and the defender choose single nodes and the underlying network is a tree. The reason of choosing trees

is that it provides insights about the role of network topology on the game strategy. We will discuss more general graphs in Section VI.

For the single attack-single defense nodes case, the game payoff will simply become

$$J(B, D_y) = [(L + k\mathbf{e}_j\mathbf{e}_j^T)^{-1}]_{ii}, \tag{9}$$

where v_i, and v_j are the node under attack and the node with the self-feedback, respectively. We need the following lemma to further study the game value and the equilibrium strategies. The proof of this lemma is postponed to Section VI.

Lemma 1: Suppose that $\mathcal{G}$ is an undirected tree and there exists a single defense node v_j (which has a self loop with weight k). Then we have

$$[\bar{L}^{-1}]_{ii} = \frac{1}{k} + \sum_{h\in\mathcal{P}_{ij}} \frac{1}{w_h}, \tag{10}$$

where $\mathcal{P}_{ij}$ is the set of edges in the (unique) path from v_i to v_j and w_h is the weight of edge h.

We next show that the attacker defender game does not admit a Nash equilibrium.

Proposition 2: The single attacker-single defender game with a single attack and single defense node does not admit a Nash equilibrium.

Proof: The proof is based on the fact that the value of all diagonal elements of the game matrix A is $A_{ii} = \frac{1}{k}$ and each off-diagonal element is $A_{ij} = \frac{1}{k} + \sum_{h\in\mathcal{P}_{ij}} \frac{1}{w_h}$. Thus, each diagonal element is strictly less than the elements of its corresponding row and column. Now assume that a NE exists and let (i^*, j^*) denote the equilibrium strategies of the attacker and defender. Thus, we should have

$$[A]_{i^*j} \leqslant [A]_{i^*j^*} \leqslant [A]_{ij^*} \tag{11}$$

for all $i \neq i^*$ and $j \neq j^*$. If element $[A]_{i^*j^*}$ is in one of the diagonal elements then the left inequality will be violated and if it is in one of the non-diagonal elements, the right inequality will be violated. ■

Although the attacker-defender game does not admit a Nash equilibrium, the optimal defense strategy can be determined by finding the solution of the Stackelberg game between the attacker and defender. In the Stackelberg game formulation, the defender acts as the game leader, *i.e.,* the leader solves the following optimization problem

$$J^*(D_y) = \min_{D_y} \operatorname{tr}\left(B^{*^T}(D_y)\bar{L}^{-1}B^*(D_y)\right). \tag{12}$$

where D_y is chosen over all $\bar{f}$ defense nodes (here $\bar{f} = 1$) in $\mathcal{V}$ and $B^*(D_y)$ is the best response of the attacker when the strategy of the defender is D_y, i.e., $B^*(D_y)$ is the solution of the following optimization problem

$$B^*(D_y) = \arg\max_{B} \operatorname{tr}\left(B^T\bar{L}^{-1}B\right), \tag{13}$$

where B is chosen over all f attacked nodes (here $f = 1$) in $\mathcal{V}$. In particular, for a given strategy of the defender, i.e., D_y, the attacker finds its best response strategy to the defender's decision, which is given by $\arg\max_B \operatorname{tr}\left(B^T\bar{L}^{-1}B\right)$. Then, the defender optimizes its decision based on all possible best response strategies of the attacker. Unlike Nash equilibrium, a Stackelberg game always admits an equilibrium strategy.

Based on the above discussion and the definition of graph center presented in Section II-A, we have the following theorem.

Theorem 1: Consider the Stackelberg attacker-defender game, with the defender as the game leader, over the connected undirected tree $\mathcal{G}$. Then, a solution of the game happens when the defender chooses the weighted center of the graph and the attacker chooses the node with longest (weighted) distance from the center. □

Proof: We know that for the game matrix A we have $A_{ij} = \frac{1}{k} + \sum_{h\in\mathcal{P}_{ij}} \frac{1}{w_h}$. As the defender is the leader of the Stackelberg game, it minimizes (over all rows) the maximum element of each row of A. Since the term $\frac{1}{k}$ is shared over all elements of A, then the optimal place for the defender is $v^* = \arg\min_i \max_j \sum_{h\in\mathcal{P}_{ij}} \frac{1}{w_h}$ and this is the center of the graph, whose (weighted) eccentricity is minimized. Note that this solution (strategies of the defender and attacker) may not be unique since the center of the network may not be a single node. However, the value of the game is unique. ■

V. Equilibrium Analysis: Multiple Attack Nodes, Single defense Node On Trees

In this section, we analyze the attacker-defender game with multiple attack nodes and a single defense node. One can interpret this as the lack of knowledge about the number of attacks to the network. In this case, if the defender knows an upper bound on the number of attacks f, then it can choose a strategy which corresponds to the worst case scenario, i.e., f nodes are under attack. The following remark discusses this worst case more formally.

Remark 2: **(Effect of Adding Attackers)**: By increasing the number of attackers, the game payoff will increase (the energy required to attack the system will decrease). More formally, since $\bar{L}^{-1}$ is a positive matrix, for two attack sets $\mathcal{B}$ and $\bar{\mathcal{B}}$ where $\mathcal{B} \subseteq \bar{\mathcal{B}}$, we have $\sum_{i\in\mathcal{B}}[\bar{L}^{-1}]_{ii} \leqslant \sum_{i\in\bar{\mathcal{B}}}[\bar{L}^{-1}]_{ii}$. □

For this scenario, the optimal strategy of the single defender depends on the number of nodes under attack. The following theorem characterizes the equilibrium solution of the game with multiple attack nodes and a single defense node. The proof of this theorem follows a similar logic to that of Theorem 1. The notion of graph's f-center was defined in Section II-A.

Theorem 2: Consider the Stackelberg attacker-defender game with single defender and f attacks, $f > 1$, with the defender as the game leader, over the connected undirected tree $\mathcal{G}$. Then, a solution of the game is when the defender chooses the weighted f-center of the graph and the attackers choose the farthest f nodes from the f-center. □

It should be noted that the graph's closeness central node (the node whose summation of its distances to all other nodes is minimized) is equivalent to $(n-1)$-center node. In general,

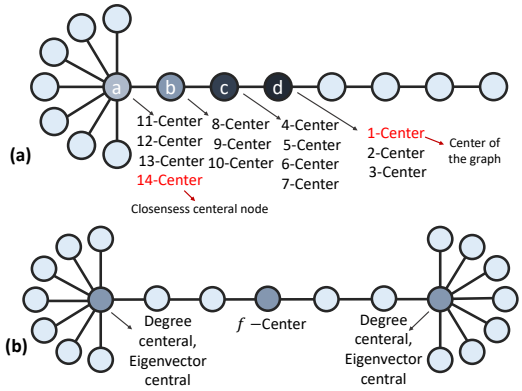

Fig. 2: An example of a graph showing that the f-center of a graph can change with f.

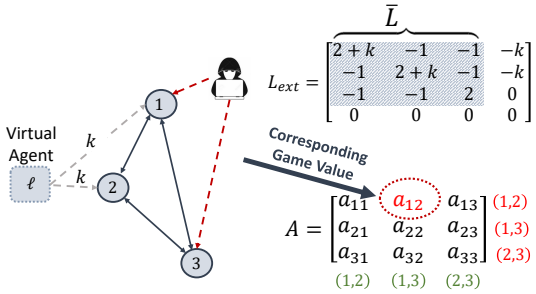

Fig. 3: Extended graph and the virtual leader.

the location of graph's f-center in the network changes by changing f. This is discussed in the following example.

Example 1: A broom tree is a graph which is formed by a star, comprised of m nodes connected to the center, followed by a tail of length m. In the broom tree shown in Fig. 2 (a), the closeness central node of the graph always is in the center of the star; however, the center (and in general its f-center) will go left through the tail as m increases. □

Theorem 2 together with Example 1 show that in the absence of knowledge of the number of attacks on the system, choosing one of f-centers of the graph results in a sub-optimal solution. However, it gives a message that centralities which are distance-based (like f-centrlities mentioned above) are among appropriate choices for the defender. In other words, degree-based centralities, such as degree central node and eigenvector central node [21], can reach to inappropriate decisions, as discussed in the following example.

Example 2: In Fig. 2 (b), the degree and eigenvector central nodes are the two nodes in the end of the line; however, the f-central node, for any $f \leqslant n-1$, is located in the middle of the line. By increasing the length of the line, the f-central node becomes arbitrarily far from the degree (and eigenvector) central nodes. □

VI. Multiple Attack, Multiple defense Nodes on General Graphs

In this section, we discuss the case where there are f attack and $\bar{f}$ defense (self-feedback loops) in the general network where $f, \bar{f} \geqslant 1$. In order to tackle this problem, we need to reinterpret the self-feedback loops in terms of connections to some virtual agent as shown in Fig. 3. We call the graph including such a virtual agent the *extended graph* and its Laplacian is denoted by L_{ext}. Any defense node is connected to that virtual agent with an edge of weight k (the self-feedback). This virtual agent is interpreted as a grounded node if we interpret the network as a circuit. Thus, it does not have any effect on the dynamics and exists only to facilitate the graph-theoretic interpretation of the attacker-defender game.

Here we relax the assumption of acyclic networks and solve the game on general graphs. In this case, as there may be multiple paths between each couple of nodes, each element of the game matrix a_{ij} will be the sum of effective resistances (instead of the physical distance) between nodes in the i-th attack set (consists of f attacks) when the node set j (consists of $\bar{f}$ defenders) is chosen to be the defender set.

If we treat $\bar{L}$ as a grounded Laplacian matrix (grounded at the virtual node ℓ), it is known that the i-th diagonal element of $\bar{L}^{-1}$ is the effective resistance from node v_i and the virtual node ℓ [22]. [4] With this in mind, consider nodes 1 and 2 in Fig. 3 which are chosen as defenders and nodes 1 and 3 which are under attack. In this case, element a_{12} in matrix game A determines the game value which is equal to $a_{12} = \sum_{i\in\mathcal{B}}[\bar{L}^{-1}]_{ii} = R_{1\ell} + R_{3\ell}$. Based on this fact, the following theorem discusses the equilibrium of Stackelberg game for general case of f attack nodes and $\bar{f}$ defense nodes on connected undirected networks.

Theorem 3: Consider the Stackelberg attacker-defender game with $\bar{f}$ defense nodes and f attack nodes, $f, \bar{f} \geqslant 1$, with the defender as the game leader, over the connected undirected graph $\mathcal{G}$. Let's denote the virtual agent corresponding to a set of $\bar{f}$ defense nodes $\mathcal{D}$ by $\ell(\mathcal{D})$. Then, a solution of the game is when the defender chooses $\bar{f}$ nodes $\mathcal{D}$ in which the maximum sum of effective resistances between $\ell(\mathcal{D})$ and all combinations of f nodes in the network is minimized, i.e., $\mathcal{D}^* = \arg\min_{\mathcal{D}\subseteq\mathcal{V}} \max_{,\mathcal{B}\subseteq\mathcal{V}} \sum_{j\in\mathcal{B}} R_{\ell(\mathcal{D})j}$. Moreover, the attacker chooses the set of f attack nodes as $\mathcal{B}^*(\mathcal{D}) = \arg\max_{\mathcal{B}\subseteq\mathcal{V}} \sum_{j\in\mathcal{B}} R_{\ell(\mathcal{D}^*)j}$. □

Remark 3: **(The Effect of Increasing Connectivity)**: Since the effective resistance between two nodes in the graph is an increasing function of edge weights (or decreasing function of edge conductance, as mentioned in [22]), adding extra edges to the network (or increasing the weight of edges) will decrease the diagonal elements of $\bar{L}^{-1}$ and consequently decreases the trace of the controllability Gramian which results in increasing the attack energy. Hence, it would be beneficial from the defender's perspective. With the same reasoning, removing edges (or weights) from the network results in a less secure system. □

Remark 4: **(The Effect of Adding defense nodes)**: In the extended graph, adding a self-feedback loop to node v_i is equivalent to adding an edge from v_i to the virtual agent with

[4] When the graph is a tree, the effective resistance and physical distance become the same and this proves the statement of Lemma 1.

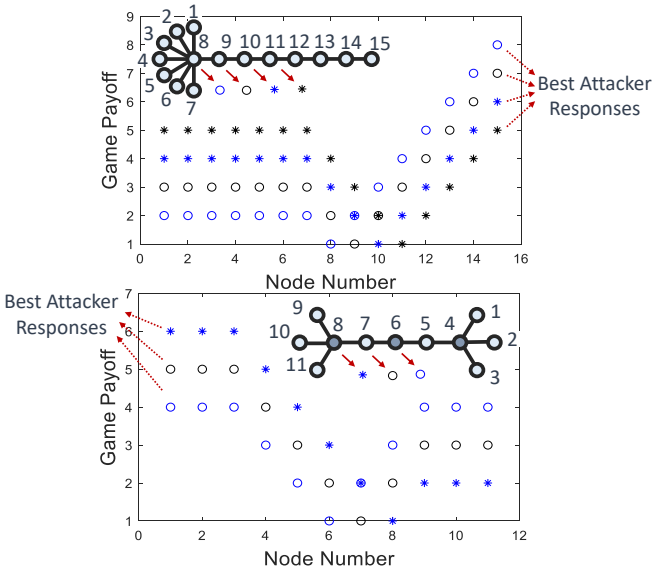

Fig. 4: single attacker-single defender, (Top) Effect of choosing the center of the graph as the defender, (b) comparing degree-based centralities with the center of the graph.

weight k. Hence, similar to what was discussed in Remark 3, adding extra defense node results in decreasing the game value; thus, it is beneficial for the security of the system. □

VII. Simulations

In this section, we present simulation results for an attacker-defender game over a graph with single attack-single defense nodes. Fig. 4 (top) shows the game payoffs for four sample defense nodes (nodes 8, 9, 10, and 11) each of which for all possible choices of the attack nodes (the horizontal axis belongs to the attacker's label). According to this figure, the best response of the attacker to each strategy of the defender is to select the end of the line (node 15) as the attack node. Thus, the optimal value from the defender's perspective (leader of the Stackelberg game) is to choose node 11 whose attacker's best response (node 15) is less than the other four (black star in node 15). As another example, we look at the structure which was discussed in Fig. 3 as well where there are two stars linked with a line. In this case, similar to the previous example, the game payoff for three sample nodes (6, 7, and 8) as defenders is depicted for all possible choices of the attack nodes (horizontal axis) and the best response of the attacker is node 1.[5] Hence, the minimum of these values belongs to the case where the defender is located in node 6.

VIII. Conclusion

A game-theoretic approach to the resilience of consensus problems in the presence of attacks was proposed. The motivation behind this work was to come up with a method to increase the cost of attack for the case where the underlying network is not highly connected (large connectivity is necessity for previous methods in the literature to mitigate the effect of attacks). It was shown that the optimal solution of the Stackelberg attacker-defender game on trees (from a single defender's perspective) in the presence of f attackers is to choose a specific centrality metric, called f-center of the graph. The results of Stackelberg game were extended to general connected graphs and multiple attackers and defenders.

[5]The best attacker's response it not unique since nodes 2 and 3 are also best responses.

References

[1] F. Pasqualetti, F. Dorfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 110–127, 2015.

[2] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conf.*, 2009, pp. 91–918.

[3] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *IEEE Conf. on Decision and Control*. IEEE, 2010, pp. 5991–5998.

[4] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Proc. IEEE Int. Conf. Smart Grid Commun*, 2010, pp. 214–219.

[5] Y. Mo, J. P. Hespanha, and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 31–43, 2014.

[6] S. M. Dibaji, M. Pirani, D. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of CPS security," *Submitted for Journal Publication.*, 2019.

[7] Q. Zhu and T. Basar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems," in *IEEE control systems*, vol. 35, no. 1, 2015, pp. 45–65.

[8] M. Manshaei, Q. Zhu, T. Alpcan, T. Basar, and J. P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys*, vol. 45, pp. 53–73, 2013.

[9] M. Pirani, E. M. Shahrivar, B. Fidan, and S. Sundaram, "Robustness of leader - follower networked dynamical systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 4, pp. 1752–1763, 2018.

[10] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[11] K. E. Fitch and N. E. Leonard, "Information centrality and optimal leader selection in noisy networks," *Proceedings of IEEE Conference on Decision and Control*, pp. 7510–7515, 2013.

[12] N. Bof, G. Baggio, and S. Zampieri, "On the role of network centrality in the controllability of complex networks," *IEEE Transactions on Control of Network Systems*, vol. 4, pp. 643–653, 2017.

[13] N. A. Lynch, *Distributed Algorithms*. Elsevier, 1996.

[14] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, pp. 1454–1467, 2014.

[15] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 766–781, 2013.

[16] S. M. Dibaji and H. Ishii, "Resilient consensus of second-order agent networks: Asynchronous update rules with delays," *Automatica*, vol. 81, pp. 123–132, 2017.

[17] J. Ghaderi and R. Srikant, "Opinion dynamics in social networks with stubborn agents: Equilibrium and convergence rate," *Automatica*, vol. 50, pp. 3209–3215, 2014.

[18] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 95, pp. 215–233, 2007.

[19] M. Pirani and S. Sundaram, "On the smallest eigenvalue of grounded Laplacian matrices," *IEEE Transactions on Automatic Control*, vol. 61, no. 2, pp. 509–514, 2016.

[20] F. Pasqualetti, S. Zampieri, and F. Bullo, "Controllability metrics, limitations and algorithms for complex networks," *Control of Network Systems, IEEE Transactions on*, vol. 1, pp. 40–52., 2014.

[21] M. E. J. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[22] A. Ghosh, S. Boyd, and A. Saberi, "Minimizing effective resistance of a graph," *SIAM Review*, vol. 50, no. 1, pp. 37–66, 2008.