Evaluating Criticality of Nodes in Consensus Network Under False Data Injection Attack

Sawant, Vishal; Wisniewski, Rafal

Published in:
IEEE Control Systems Letters

DOI (link to publication from Publisher):
10.1109/LCSYS.2023.3257265

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

Evaluating Criticality Of Nodes In Consensus Network Under False Data Injection Attack

Vishal Sawant and Rafal Wisniewski

***Abstract*— In this paper, a finite-duration, magnitude-bounded *false data injection (FDI)* attack on consensus network is considered. The aim of the attacker is to induce disagreement between nodes and consequently, influence the convergence of consensus algorithm. In order to measure the induced disagreement, a metric, namely *induced terminal disagreement (ITD)*, is defined. The objective of this study is to determine the criticality of individual nodes in terms of the worst-case ITD resulting from attack on them. To achieve that, for every node, the closed-form expressions for the optimal attack input which results in the maximum ITD and the corresponding value of ITD, are obtained. Based on that, *criticality ranks* are assigned to all nodes. These ranks are beneficial in allocating security resources and designing resilient architecture. Further, the effect of varying attack duration on the worst-case ITDs and criticality ranks, is analyzed. Finally, it is shown that the criticality ranks of nodes have strong negative correlation with their degrees. A numerical example and simulations are presented to illustrate the proposed results.**

***Index Terms*— Consensus, Cyberattack, Node criticality**

I. Introduction

OVER the last few years, cyber-physical security of networked systems has received tremendous research interest (see survey [1] and references therein) due to their application in safety-critical infrastructures such as power grid, water distribution, transportation, etc., and recent cyber-attacks [2] - [3] on such infrastructures. In networked systems, consensus based algorithms are widely used for various purposes such as formation control [4], distributed optimization [5], sensor data fusion [6], etc. If the cyber attack adversely influences the convergence of consensus algorithm, it could compromise the critical control objectives which rely on the consensus value. To avoid that, various attack detection and mitigation techniques [7], [8] have been proposed. However, these techniques require significant time to neutralize the attack, during which consensus algorithm is vulnerable to performance degradation. Such degradation is likely to vary with nodes/links under attack. For a system designer, it is important to know the critical nodes/links, which when attacked, result in the maximum performance degradation, and the corresponding degradation level. This information is vital in allocating security resources and designing resilient architecture [9].

The problem of analyzing the impact of cyber attacks on consensus algorithm has received limited attention. The existing work considers two types of attacks. The first attack is known as the *jamming* attack [10] - [12] in which the attacker disables one or more communication links in the network. The second attack is known as the *false data injection (FDI)* attack [10] in which the attacker injects malicious input, referred to as the *attack input*, into the nodes in the consensus network. In [10], both jamming and FDI attacks on consensus network are considered. For each of these attacks, the authors formulated a finite horizon optimization problem to maximize the distance of nodes' states from the consensus point. Subsequently, authors proposed techniques for computing optimal jamming and FDI strategies. Later, [11] corrected erroneous claim in [10] about the computational efficiency of obtaining optimal jamming strategy, and proved that it is a strongly NP-hard problem. Recently, the optimization framework from [10] was used in [12] to obtain optimal strategy for jamming attack on bipartite consensus algorithm. In [13], a distributed algorithm is proposed to find the most critical node in the consensus network whose removal results in the maximum reduction in the convergence speed of the consensus algorithm. The effect of cyber attacks on consensus algorithm, in presence of defensive action by system designer, is analyzed in [14] - [17] as a zero-sum game between attacker and designer. In [14]-[15], jamming attack is considered whereas in [16] - [17], FDI attack is considered. In [14] - [17], the saddle-point strategies are obtained for both attacker and designer.

Another related problem is of determining the criticality of individual nodes with respect to given performance metrics, which are known as the *network centrality measures* [18]. Some examples of such metrics used for dynamical networks are: the number of sensors required to be manipulated for stealthy FDI attack [19], controllability metrics [20], leader selection metrics [21], etc.

Each of the works in [10] - [21] involves numerical optimization to obtain the optimal attack strategy. If the system designer uses similar approach to determine the criticality of individual nodes/links in terms of performance degradation resulting from attack on them, she would need to execute the above-mentioned numerical optimization once for every node/link. This scheme is inefficient for practical networks with thousands of nodes [22]. We overcome this limitation by obtaining the closed-form expressions for optimal attack strategy and the corresponding performance degradation. We also obtain the relation between the criticality of nodes and network topology.

In this paper, we consider a finite duration FDI attack on a consensus network, that is following the *nearest neighbour rule* [23]. In practice, the attacker is often resource constrained in terms of actuator saturation, power consumption [10], [15], etc. To model these limitations, we assume that the attack input is magnitude bounded. We also assume that the attacker knows the topology of the consensus network, but it does not know the states of un-attacked nodes.

In order to degrade consensus performance, the obvious goal of the attacker is to increase the disagreement between nodes.

The authors are with the Department of Electronic Systems, Aalborg University, Denmark - 9220. E-mail: {vishalss, raf}@es.aau.dk. Corresponding author: Vishal Sawant. This work was supported by CRUCIAL project.

To measure disagreement induced by the attacker, we define a metric, which we refer to as the, *induced terminal disagreement (ITD)*. Our basic objective is to determine the criticality of individual nodes in terms of the worst-case ITD resulting from FDI attack on them. Thus, we consider FDI attack on one node at a time. For every node, we obtain the closed-form expressions for the optimal attack input and the corresponding worst-case ITD (Theorems 6 and 7). Interestingly, for every attack duration, the optimal attack input is a constant signal (Theorem 6). Subsequently, we assign ranks, namely *criticality ranks*, to nodes, in decreasing order of worst-case ITDs.

Next, we analyze the effect of varying attack duration on the worst-case ITDs and the criticality ranks of nodes (Problem 4.3). We show that the worst-case ITDs corresponding to all nodes, increase monotonically with the attack duration and eventually converge to same value (Theorem 9). We also show with an example that the order of criticality ranks of nodes can change with the attack duration.

Finally, we investigate the relation between the criticality ranks of nodes and their degrees (Section V). We first prove that for small attack durations, the criticality rank of a node is inversely proportional to its degree (Theorem 11). Then, we demonstrate through simulations that even for larger attack durations, the criticality ranks of nodes have strong negative correlation with their degrees (Section V-C).

Our contributions in this paper are summarized as follows:
1) We consider a finite-duration, magnitude-bounded, single-node FDI attack on a consensus network.
2) For every node, we obtain the closed-form expressions for the optimal attack input which maximizes ITD and the corresponding value of ITD (Theorems 6 and 7).
3) We analyze the effect of varying attack duration on the worst-case ITDs and criticality ranks of nodes (Theorem 9).
4) We show that the criticality ranks of nodes have strong negative correlation with their degrees (Theorem 11).

The remaining paper is organized as follows. In Section II, the problem of computing worst-case ITDs and criticality ranks, is formulated. In Section III, the closed-form expressions for the optimal attack input and the corresponding worst-case ITD, are obtained. The effect of varying attack duration on the worst-case ITDs and criticality ranks of nodes, is analyzed in Section IV. In Section V, the relation between the criticality ranks and degrees of nodes is investigated. The paper is concluded in Section VI with future direction.

II. Preliminaries and Problem Formulation

A. Graphs

A graph $G = (V, E)$ is a set of nodes V connected by a set of edges $E \subseteq (V \times V)$. A graph G is said to be *undirected* if $(i, j) \in E$ implies $(j, i) \in E$. In an undirected graph, if $(i, j) \in E$, then the nodes i and j are called as the *neighbours* of each other. A graph is said to be *simple* if $(i, i) \notin E,\ \forall i \in V$. A *path* between nodes i and j is a sequence of edges connecting them. An undirected graph is called as *connected* if there exists a path between any two nodes in it.

Let n be number of nodes in V and d_i be the number of neighbours, i.e., the *degree*, of node i. Then, the *Laplacian* matrix $L \in \mathbb{R}^{n\times n}$ of a simple undirected graph $G = (V, E)$ is defined as

$$L_{i,j} := \begin{cases} d_i, & \text{if} \quad i = j \\ -1, & \text{if} \quad i \neq j \quad \text{and} \quad (i,j) \in E \\ 0, & \text{if} \quad i \neq j \quad \text{and} \quad (i,j) \notin E \end{cases} \tag{1}$$

B. Consensus network

Consider a network of n nodes with underlying undirected simple graph $G = (V, E)$. Let S_i be the set of neighbours of node i, with cardinality d_i. Let $L \in \mathbb{R}^{n\times n}$ be the Laplacian matrix of G. We make the following assumption about G.

Assumption 1: G is *time-invariant* and *connected*.

Define the index set $P = \{1, 2, \ldots, n\}$. Let $x_i(t) \in \mathbb{R}$ denote the state of node i at instant t. In order to achieve consensus, the nodes employ the classical *nearest neighbour rule* [23] which is given as

$$\dot{x}_i(t) = -\sum_{j\in S_i} \Big(x_i(t) - x_j(t)\Big), \qquad \forall i \in P \tag{2}$$

It is well known [23] that dynamics (2) achieves asymptotic consensus of the network, i.e., $\lim_{t\to\infty} x_i(t) = \lim_{t\to\infty} x_j(t),\ \forall i, j \in P$. Let $X(t) := [x_1(t), \ldots, x_n(t)]^T \in \mathbb{R}^n$ be the augmented state of all nodes. Then, it follows from the definition of L in (1) that the augmented dynamics of the network is

$$\dot{X}(t) = -LX(t) \tag{3}$$

C. Attack model

We consider an FDI attack in which the attacker injects malicious input, referred to as the *attack input*, into the nodes in the network. As our basic objective is to determine the criticality of individual nodes in terms of impact of FDI attack on them, we consider attack on one node at a time. Let i be the attacked node and u_i be the attack input. Then, the dynamics of node i becomes

$$\dot{x}_i(t) = -\sum_{j\in S_i} \Big(x_i(t) - x_j(t)\Big) + u_i(t) \tag{4}$$

Let $e_i \in \mathbb{R}^n$ be the ith canonical unit vector. Then, it follows from (2)-(4) that in presence of attack on node i, the augmented dynamics of the network becomes

$$\dot{X}(t) = -LX(t) + e_i u_i(t) \tag{5}$$

We consider attack over a finite time interval, which is referred to as the *attack interval*. Let $T_a \in [0, \infty)$ be the attack duration. Then, without loss of generality, we take the attack interval to be $\big[0, T_a\big]$.

It is well known that the attacker often has resource constraints such as actuator saturation, power consumption [10], [15], etc. To take them into account, we assume that -

Assumption 2: The attack input u_i is magnitude bounded and belongs to the set $\mathcal{U} := \big\{u \in \mathcal{C} \ \big|\ |u(t)| \leq \beta, \forall t\big\}$ where $\mathcal{C}$ denotes the set of piecewise continuous functions from $[0, T_a]$ to $\mathbb{R}$.

Before beginning the attack, the attacker does not have access to the state of the network. However, the attacker can still have prior knowledge of the network topology. Hence, we make the following assumption.

Assumption 3: Before initiating the attack, the attacker does not know the state of the network. But it knows the Laplacian L and the consensus dynamics (3).

D. Performance metric

The basic goal of the attacker is to induce disagreement between nodes which would influence the convergence of consensus dynamics. To measure the induced disagreement, a natural metric, as used in [10, Section 3], is the sum of distances of nodes' states from the consensus point. However,

for this metric, normally, numerical methods are required to compute the optimal attack input and the corresponding worst-case induced disagreement. As a result, it is not easy to characterize the induced disagreement in terms of network topology. To overcome this issue, we choose a metric that is expressible in terms of the Laplacian L of the network.

Define $Z(t) = [z_1(t), \ldots, z_n(t)]^T := LX(t)$. Then, it follows from the definition of Laplacian in (1) that $z_i(t) = \sum_{j \in S_i} \big(x_i(t) - x_j(t)\big)$, which implies that $z_i(t)$ is the disagreement between node i and its neighbours. Hence, the vector $Z(t)$ captures the aggregate disagreement in the network. Further, it is well known [23] that for the Laplacian of a connected graph,

$$Z(t) = LX(t) = 0 \quad \iff \quad x_i(t) = x_j(t), \ \forall i, j \in P$$

As a result, the norm $||Z(t)||_2$, i.e., the distance of $Z(t)$ from the origin, is a good measure of the disagreement in the network at instant t. Then, it is natural for the attacker to aim to maximize the disagreement $||Z(T_a)||_2$ at the end of the attack interval.

The state trajectory of the augmented dynamics (5) at instant $t = T_a$ is

$$X\big(T_a\big) = e^{-LT_a} X(0) + e^{-LT_a} \int_0^{T_a} e^{L\tau}\, e_i\, u_i(\tau)\, d\tau$$

Then, it follows from $Z(t) = LX(t)$ that

$$Z\big(T_a\big) = Le^{-LT_a} X(0) + Le^{-LT_a} \int_0^{T_a} e^{L\tau}\, e_i\, u_i(\tau)\, d\tau \quad (6)$$

Recall from Assumption 3 that the attacker does not know $X(0)$ before the attack. Thus, it can not affect the term $Le^{-LT_a}X(0)$ in (6). However, the attacker knows the Laplacian L and can control the second term on the right hand side of (6), which we denote by

$$F\big(i, u_i, T_a\big) := Le^{-LT_a} \int_0^{T_a} e^{L\tau}\, e_i\, u_i(\tau)\, d\tau \quad (7)$$

Further, by applying the triangle inequality in (6), we get $\big|\big|Z\big(T_a\big)\big|\big|_2 \leq \big|\big|Le^{-LT_a}X(0)\big|\big|_2 + \big|\big|F\big(i, u_i, T_a\big)\big|\big|_2$. It implies that the maximization of $\big|\big|F\big(i, u_i, T_a\big)\big|\big|_2$ results in the maximization of the upper bound on $\big|\big|Z\big(T_a\big)\big|\big|_2$. Then, to maximize $\big|\big|Z\big(T_a\big)\big|\big|_2$, it is rational for the attacker to maximize $\big|\big|F\big(i, u_i, T_a\big)\big|\big|_2$. Thus, we take $\big|\big|F\big(i, u_i, T_a\big)\big|\big|_2$ as the performance metric and call it the *induced terminal disagreement (ITD)* corresponding to attack on node i.

E. Problem statement

Recall that our basic objective is to determine the criticality of individual nodes in terms of the worst-case ITD resulting from FDI attack on them. To achieve that, we first need to compute the worst-case ITDs corresponding to all nodes. Based on that, we could assign a rank to every node, which we refer to as the *criticality rank*. Let $C\big(i, T_a\big)$ denote the criticality rank of node i corresponding to attack duration T_a. Our next objective is to analyze the effect of varying T_a on $C\big(i, T_a\big)$'s. We formalize our objectives as follows.

Problem 4: Consider dynamics (5) corresponding to attack on node i. Let $F\big(i, u_i, T_a\big)$ be as defined in (7). Then,
1) For every node $i \in P$, obtain the attack input $u_i^* \in \mathcal{U}$, which is solution of the following optimization

$$u_i^* = \arg \max_{u_i \in \mathcal{U}} \ \big|\big|F\big(i, u_i, T_a\big)\big|\big|_2 \quad (8)$$

and the corresponding (worst-case) ITD $\big|\big|F\big(i, u_i^*, T_a\big)\big|\big|_2$.
2) Assign criticality ranks $C\big(i, T_a\big)$'s to all nodes.
3) Analyze the effect of varying T_a in $\big[0, \infty\big)$ on $\big|\big|F\big(i, u_i^*, T_a\big)\big|\big|_2$'s and $C\big(i, T_a\big)$'s.

III. Computation of $\big|\big|F\big(i, u_i^*, T_a\big)\big|\big|_2$'s and $C\big(i, T_a\big)$'s

In this section, we first present the eigenstructure of Laplacian L and then express $||F\big(i, u_i, T_a\big)||_2$ in terms of that eigenstructure. Later, we will use this expression for obtaining the solution of optimization (8).

A. Eigenstructure of Laplacian L

Recall that the underlying graph G of our consensus network is undirected. As a result, its Laplacian matrix L is symmetric. Consequently, L has real eigenvalues and it is diagonalizable with real, orthogonal eigenvectors [24, Chapter 8]. Further, the network graph G is assumed to be connected (Assumption 1). As a result, L possesses certain special properties. We present these properties next.

Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ be the eigenvalues of L in an increasing order, with corresponding normalized eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$. Define the diagonal matrix $D := diag\big\{\lambda_1, \ldots, \lambda_n\big\} \in \mathbb{R}^{n \times n}$ and matrix $Q := [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n}$. Let $I_n \in \mathbb{R}^{n \times n}$ be the identity matrix of order n. Then, the following proposition enlists the properties of the eigenstructure of L.

Proposition 5: [24, Chapter 8] [25, Chapter 4] Let L be the Laplacian matrix of a connected undirected graph G. Let the eigenpairs $(\lambda_j, v_j), \ \forall j \in P$, of L and the matrices D, Q and I_n be as defined above. Then, the following holds.

1) $\lambda_1 = 0$ and $v_1 = \big[\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\big]^T \in \mathbb{R}^n$.
2) $\lambda_j > 0$ for $j = 2, \ldots, n$.
3) L is diagonalizable as $L = QDQ^{-1}$.
4) Q is an orthogonal matrix, i.e., $Q^TQ = I_n$.

B. Expressing ||F(i, u_i, T_a)|| in terms of eigenstructure of L

Consider the expression of $F\big(i, u_i, T_a\big)$ in (7). Recall from Proposition 5.3 that $L = QDQ^{-1}$. Then, it follows from the definition of matrix exponential [26, Chapter I] that $e^L = Qe^DQ^{-1}$. By using expressions for L and e^L in (7), we get

$$F\big(i, u_i, T_a\big) = QDQ^{-1}Qe^{-DT_a}Q^{-1}\int_0^{T_a} Qe^{D\tau}Q^{-1}e_iu_i(\tau)d\tau$$

Subsequently, by interchanging the order of Q^{-1} and integration, and canceling product terms $Q^{-1}Q = I_n$, we get

$$F\big(i, u_i, T_a\big) = QDe^{-DT_a} \int_0^{T_a} e^{D\tau}Q^{-1}e_iu_i(\tau)d\tau \quad (9)$$

We know from Proposition 5.4 that Q preserves the 2-norm. Then, (9) leads to

$$\big|\big|F\big(i, u_i, T_a\big)\big|\big|_2 = \Big|\Big|De^{-DT_a}\int_0^{T_a} e^{D\tau}Q^{-1}e_iu_i(\tau)d\tau\Big|\Big|_2 \quad (10)$$

Define $b_i = [b_{i_1}, \ldots, b_{i_n}]^T := Q^{-1}e_i$. Recall that D and $e^{D\tau}$ are diagonal matrices. Then, $De^{-DT_a} = diag\Big\{\lambda_1 e^{-\lambda_1 T_a}, \ldots, \lambda_n e^{-\lambda_n T_a}\Big\} \in \mathbb{R}^{n \times n}$ and $e^{D\tau}b_i =$

$\left[b_{i_1}e^{\lambda_1\tau},\ldots,b_{i_n}e^{\lambda_n\tau}\right]^T \in \mathbb{R}^n$. Then, the term on the right hand side of (10) becomes

$$De^{-DT_a}\int_0^{T_a} e^{D\tau}b_i u_i(\tau)d\tau = \begin{bmatrix} b_{i_1}\lambda_1 e^{-\lambda_1 T_a}\int_0^{T_a} e^{\lambda_1\tau}u_i(\tau)d\tau \\ b_{i_2}\lambda_2 e^{-\lambda_2 T_a}\int_0^{T_a} e^{\lambda_2\tau}u_i(\tau)d\tau \\ \vdots \\ b_{i_n}\lambda_n e^{-\lambda_n T_a}\int_0^{T_a} e^{\lambda_n\tau}u_i(\tau)d\tau \end{bmatrix}$$

For every $j \in P$, define $c_{i_j} := b_{i_j}\lambda_j e^{-\lambda_j T_a} \in \mathbb{R}$. Then, (10) can be written as

$$\left|\left|F\big(i,u_i,T_a\big)\right|\right|_2 = \left(\sum_{j=1}^{n} c_{i_j}^2\left(\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau\right)^2\right)^{1/2} \quad (11)$$

This expression will be instrumental in obtaining the solution of optimization (8) in the next section.

C. Computation of u_i^* and $\left|\left|F(i,u_i^*,T_a)\right|\right|_2$

In this section, we first obtain the solution u_i^* of optimization (8).

Theorem 6: Let β be the magnitude bound of the attack input u_i and $T_a \in \mathbb{R}^+$ be any attack duration. Then, the solutions of optimization (8) are -

$u_i^*(t) = \beta,\ \forall t \in \big[0,T_a\big]$ or $u_i^*(t) = -\beta,\ \forall t \in \big[0,T_a\big]$

Proof: Consider the integral term $\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau$ in the expression of $||F\big(i,u_i,T_a\big)||_2$ in (11). It represents the area under the function $e^{\lambda_j\tau}u_i(\tau)$ in the interval $[0,T_a]$. Note that $e^{\lambda_j\tau} \geq 0,\ \forall\tau \geq 0,\ \forall j \in P$. Thus, if $u_i(\tau)$ changes sign in the interval $[0,T_a]$, then the above-mentioned area has positive and negative components. These components partially cancel out each other and reduce the absolute value of the overall area, i.e., $\left|\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau\right|$, which in turn reduces the squared term $\left(\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau\right)^2$ in (11). Then, it is easy to see that $u_i \in \mathcal{U}$ which maximizes $\left(\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau\right)^2$ has same sign and maximum possible magnitude (i.e., β) in the interval $\big[0,T_a\big]$. Thus, for each $j \in P$, the controls $u_i^*(\tau) = \beta,\ \forall\tau \in \big[0,T_a\big]$ and $u_i^*(\tau) = -\beta,\ \forall\tau \in \big[0,T_a\big]$ are maximizers of $\left(\int_0^{T_a} e^{\lambda_j\tau}u_i(\tau)d\tau\right)^2$. Then, it follows from (11) that these u_i^*'s are solutions of optimization (8). This completes the proof. ■

It is interesting that for every attack duration, the optimal attack input is a constant signal with maximum magnitude. Next, we compute the worst-case ITD corresponding to u_i^*, i.e., $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$. Let $e_i \in \mathbb{R}^n$ be the ith canonical unit vector and $I_n \in \mathbb{R}^{n\times n}$ be the identity matrix.

Theorem 7: The worst-case ITD corresponding to attack on node i with optimal attack input u_i^* is equal to the 2-norm of the ith column of the matrix $\beta\big(I_n - e^{-LT_a}\big) \in \mathbb{R}^{n\times n}$, i.e.,

$$\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2 = \left|\left|\beta\big(I_n - e^{-LT_a}\big)e_i\right|\right|_2 \quad (12)$$

Proof: Consider the expression of $\left|\left|F\big(i,u_i,T_a\big)\right|\right|_2$ in (11) and the expression $c_{i_j} = b_{i_j}\lambda_j e^{-\lambda_j T_a}$ given before (11). Recall that $\lambda_1 = 0$. Then, clearly, $c_{i_1} = 0$. Next, consider λ_j for $j \in \{2,\ldots,n\}$. Under attack input $u_i = u_i^* = \beta$ (similar arguments follow for $u_i^* = -\beta$), the integral term in (11) corresponding to λ_j becomes

$$\int_0^{T_a} e^{\lambda_j\tau}u_i^*(\tau)d\tau = \beta\int_0^{T_a} e^{\lambda_j\tau}d\tau = \beta\left(\frac{e^{\lambda_j T_a}-1}{\lambda_j}\right)$$

Then, it follows from the expression $c_{i_j} = b_{i_j}\lambda_j e^{-\lambda_j T_a}$ that $c_{i_j}^2\left(\int_0^{T_a} e^{\lambda_j\tau}u_i^*(\tau)d\tau\right)^2 = \beta^2 b_{i_j}^2\left(1-e^{-\lambda_j T_a}\right)^2$, for every $j \in P$. Then, (11) leads to

$$\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2 = \beta\left(\sum_{j=2}^{n} b_{i_j}^2\left(1-e^{-\lambda_j T_a}\right)^2\right)^{1/2} \quad (13)$$

Now, recall that $D = diag\{\lambda_1,\ldots,\lambda_n\}$ and $b_i = [b_{i_1},\ldots,b_{i_n}]^T$. Then, it is easy to show that,

$$\big(I_n - e^{-DT_a}\big)b_i = \left[\big(1-e^{-\lambda_1 T_a}\big)b_{i_1},\ldots,\big(1-e^{-\lambda_n T_a}\big)b_{i_n}\right]^T$$

Recall again that $\lambda_1 = 0$ and hence, $(1-e^{-\lambda_1 T_a}) = 0$. Then, (13) leads to

$$\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2 = \left|\left|\beta\big(I_n - e^{-DT_a}\big)b_i\right|\right|_2 \quad (14)$$

We know from Proposition 5.4 that matrix Q preserves the 2-norm. Recall also that $b_i = Q^{-1}e_i$. Then, (14) leads to

$$\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2 = \left|\left|\beta Q\big(I_n - e^{-DT_a}\big)Q^{-1}e_i\right|\right|_2 \quad (15)$$

We also know that $e^{-LT_a} = Qe^{-DT_a}Q^{-1}$ and $QQ^{-1} = I_n$. Then, (15) leads to claim (12) and completes the proof. ■

D. Assigning criticality ranks $C(i,T_a)$'s to nodes

In order to assign criticality ranks to nodes, we first compute the worst-case ITDs $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$'s corresponding to all nodes, by using expression (12). Then, we arrange these $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$'s in decreasing order without repetition. Subsequently, we assign the criticality rank $C\big(i,T_a\big)$ to each node as the position of the corresponding $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$ in the above-mentioned ordered list. We illustrate this procedure with the following example.

Example 8: Consider a consensus network of $n = 7$ nodes with graph Laplacian

$$L = \begin{bmatrix} 3 & 0 & -1 & 0 & 0 & -1 & -1 \\ 0 & 2 & 0 & 0 & 0 & -1 & -1 \\ -1 & 0 & 2 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 3 & -1 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & 0 & -1 \\ -1 & -1 & 0 & -1 & 0 & 4 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 6 \end{bmatrix}$$

Let the magnitude bound of attack input be $\beta = 10$ and the attack duration be $T_a = 0.3$ sec. Then, the worst-case ITDs $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$'s and the criticality ranks $C\big(i,T_a\big)$'s for $T_a = 0.3$ sec and $T_a = 0.6$ sec are given in Table I.

IV. Effect of T_a on $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$'s and $C\big(i,T_a\big)$'s

In this section, we analyze the effect of varying T_a on $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$'s and $C\big(i,T_a\big)$'s.

A. Effect of varying T_a on $\left|\left|F(i,u_i^*,T_a)\right|\right|_2$'s

Consider the expression of $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$ in (13). At $T_a = 0$, we get $\left|\left|F\big(i,u_i^*,0\big)\right|\right|_2 = 0$. Intuitively, with increase in T_a, each $\left|\left|F\big(i,u_i^*,T_a\big)\right|\right|_2$ increases. We formalize this intuition in the next theorem.

Theorem 9: Consider a network of n nodes with dynamics (5), which corresponds to attack on node i. Then, for every $i \in P$, the following holds.

TABLE I
EXAMPLE 8 : WORST-CASE ITDS $||F(i,u_i^*,T_a)||_2$'S AND CRITICALITY RANKS $C(i,T_a)$'S OF ALL NODES

Node i	1	2	3	4	5	6	7
Degree d_i	3	2	2	3	2	4	6
$\|\|F(i,u_i^*,0.3)\|\|_2$	5.921	4.580	4.610	5.921	4.610	6.878	8.124
$C(i,0.3)$	3	5	4	3	4	2	1
$\|\|F(i,u_i^*,0.6)\|\|_2$	7.830	6.632	6.616	7.830	6.616	8.504	9.119
$C(i,0.6)$	3	4	5	3	5	2	1

1) $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$ increases monotonically with T_a.

2) $\lim_{T_a\to\infty}\ \left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2 = \beta\left(\dfrac{n-1}{n}\right)^{1/2}$

Proof: Consider the expression of $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$ in (13). Recall that $\lambda_j > 0$ for $j = 2,\dots,n$. Then, clearly, each term $\left(1-e^{-\lambda_j T_a}\right)$ in (13) increases monotonically with increase in T_a. As a result, $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$ increases monotonically with T_a. This proves the first claim.

Next, it is easy to see that $\lim_{T_a\to\infty}\left(1-e^{-\lambda_j T_a}\right)=1$ for $j=2,\dots,n$. Then, it follows from (13) that

$$\lim_{T_a\to\infty}\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2 = \beta\left(\sum_{j=2}^{n} b_{i_j}^2\right)^{1/2} \tag{16}$$

Recall that $b_i = [b_{i_1},\dots,b_{i_n}]^T = Q^{-1}e_i$ and $Q = [v_1,\dots,v_n]$. Then, we can express e_i as

$$e_i = Qb_i = \sum_{j=1}^{n} b_{i_j} v_j \tag{17}$$

We know from Proposition 5.4 that the vectors v_j's are orthonormal, i.e., $v_j^T v_j = 1$ and $v_j^T v_l = 0$ if $j \neq l$. Then, it follows from (17) that $e_i^T e_i = ||e_i||_2^2 = 1 = \sum_{j=1}^{n} b_{i_j}^2$. Further, recall from Proposition 5.1 that $v_1 = \left[\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}}\right]^T$. Then, (17) leads to $v_1^T e_i = \frac{1}{\sqrt{n}} = b_{i_1}$. Hence, $\sum_{j=2}^{n} b_{i_j}^2 = \left(\sum_{j=1}^{n} b_{i_j}^2\right) - b_{i_1}^2 = 1 - \frac{1}{n} = \frac{n-1}{n}$. Then, (16) proves claim 2 and completes the proof. ■

B. Effect of varying T_a on $C(i,T_a)$'s of nodes

The implication of Theorem 9.2 is that for $T_a = \infty$, $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$'s and consequently, $C(i,T_a)$'s of all nodes are equal. However, as visible in Table I, for $T_a \in (0,\infty)$, $C(i,T_a)$'s of two different agents can be different. Furthermore, the order of $C(i,T_a)$'s of nodes may change with T_a. To prove this claim, consider Example 8 and the corresponding $C(i,T_a)$'s given in Table I. Observe that $C(2,0.3) > C(5,0.3)$ but $C(2,0.6) < C(5,0.6)$.

Next, observe that degree d_i of node 7 is the highest but it has the lowest $C(i,T_a)$ for both $T_a = 0.3$ and $T_a = 0.6$. It is also visible in Table I that $C(i,T_a)$ increases with decrease in d_i. These observations hint at inverse proportionality between $C(i,T_a)$ and d_i, which we investigate next.

V. RELATION BETWEEN $C(i,T_a)$'S AND NODAL DEGREES

Recall from Theorem 9.2 that for $T_a \to \infty$, $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$'s and consequently, $C(i,T_a)$'s of all nodes converge to the same value. It implies that for $T_a \to \infty$, there can not be a relation between $C(i,T_a)$'s and the degrees of nodes. Thus, in this section, we investigate the above-mentioned relation for T_a's prior to the convergence of $C(i,T_a)$'s.

A. For small T_a's

We first obtain the approximation of $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$ for small attack durations.

Lemma 10: Let β be the magnitude bound on attack input and d_i be the degree of node i. Then, for small T_a's,

$$\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2 \approx \beta T_a\sqrt{d_i^2+d_i}, \qquad \forall i \in P \tag{18}$$

Proof: For small T_a's, the first-order approximation of the matrix e^{-LT_a} is $e^{-LT_a} \approx \left(I_n - LT_a\right)$, where I_n is the identity matrix. Then, $\left(I_n - e^{-LT_a}\right) \approx LT_a$. Further, it follows from the definition of L in (1) that the 2-norm of its ith column is equal to $\sqrt{d_i^2+d_i}$. Then, (12) leads to (18). ■

Now, by using Lemma 10, the following theorem shows that the criticality rank $C(i,T_a)$ of node i is inversely proportional to its degree d_i.

Theorem 11: Let d_i be the degree of node i. Then, for small T_a's: $\quad d_i < d_j \implies C(i,T_a) > C(j,T_a)$.

Proof: The claim follows from Lemma 10 and the relation between $C(i,T_a)$'s and $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$'s. ■

Note that nodes with same d_i's can have different $C(i,T_a)$'s. For example, observe in Table I that $d_2 = d_3$, but $C(2,0.3) \neq C(3,0.3)$.

B. For larger T_a's: Simulations

In order to study, the relation between $C(i,T_a)$'s and degrees of nodes for larger T_a's (prior to convergence of $C(i,T_a)$'s), we performed the following simulations. We randomly generated 1000 undirected connected graphs with $n = 75$ nodes. Let these graphs be denoted by $G_1,\dots,G_{1000}$. Let $\beta = 10$ be magnitude bound on the attack input. We began with $T_a = 0.05$ sec and did the following.

1) Consider any graph G_j. Let d_{i_j} be the degree of node i in G_j. Define $\mathcal{D}_j := [d_{1_j},\dots,d_{n_j}] \in \mathbb{R}^n$. We computed $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$'s of all nodes by using (12) and augmented them in a vector $F_j(T_a) \in \mathbb{R}^n$. Then, we computed the *Pearson correlation coefficient* between $F_j(T_a)$ and $\mathcal{D}_j$, and denoted it by $CR_j(T_a)$. Subsequently, we computed the average of all $CR_j(T_a)$'s corresponding to all 1000 graphs, and denoted it by $CR_{avg}(T_a)$.

2) Let $F_j(T_a)$ be as computed above for graph G_j. Define $f_s := \beta\left(\dfrac{n-1}{n}\right)^{1/2}$. Then, we know from Theorem 9.2 that all elements of $F_j(T_a)$, i.e, $\left|\left|F\left(i,u_i^*,T_a\right)\right|\right|_2$'s corresponding

to graph G_j, converge to f_s as $T_a \to \infty$. Thus, every element of $F_j(T_a)$ lies between 0 and f_s, for all $T_a \in [0, \infty)$. In order to measure how far the elements of $F_j(T_a)$ are from f_s, we computed the ratio of the minimum element of $F_j(T_a)$ (which is farthest from f_s) with f_s. Let this ratio be $RF_j(T_a)$. Note that $RF_j(T_a) \in [0, 1]$. Then, we computed the average of all $RF_j(T_a)$'s corresponding to all 1000 graphs, and denoted it by $RF_{avg}(T_a)$.

3) We repeated above steps for $T_a = 0.05k,\ k = 1, \ldots, 10$ and then plotted $CR_{avg}(T_a)$'s and $RF_{avg}(T_a)$'s against T_a in Fig. 1.

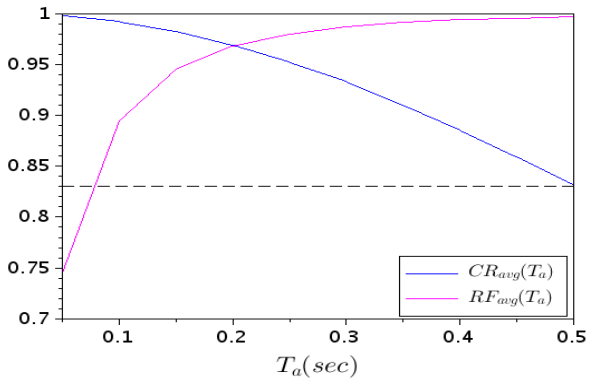

Fig. 1. Simulation: $\boldsymbol{CR_{avg}(T_a)}$ and $\boldsymbol{RF_{avg}(T_a)}$ versus $\boldsymbol{T_a}$

C. Discussion

Observe in Fig. 1 that $RF_{avg}(T_a)$ increases with T_a and eventually (almost) converges to 1 at $T_a = 0.5$ sec. Then, it follows from the definition of $RF_{avg}(T_a)$ that every $RF_j(T_a)$ (corresponding to graph G_j) converges to 1. This implies that for every graph G_j, all elements of $F_j(T_a)$'s, i.e., $\left|\left|F\left(i, u_i^*, T_a\right)\right|\right|_2$'s, converge to f_s (as claimed in Theorem 9.2). Next, observe that $CR_{avg}(T_a)$ is close to 1 for small T_a's. This implies that all $CR_j(T_a)$'s are close to 1 for small T_a's (as expected from Lemma 10). With increase in T_a, $CR_{avg}(T_a)$ decreases. However, even till $T_a = 0.5$ sec, by when all $\left|\left|F\left(i, u_i^*, T_a\right)\right|\right|_2$'s have converged to f_s, $CR_{avg}(T_a)$ is higher than 0.83. This shows strong positive correlation between $\left|\left|F\left(i, u_i^*, T_a\right)\right|\right|_2$'s and degrees of nodes. Equivalently, it shows strong negative correlation between $C(i, T_a)$'s and degrees of nodes.

VI. Conclusion

In this paper, we considered a finite-duration, magnitude-bounded, single-node FDI attack on a consensus network. To measure the disagreement induced by the attacker, we defined a metric abbreviated as ITD. For every node, we obtained the closed-form expressions for the optimal attack input which results in the maximum ITD and the corresponding value of ITD (Theorems 6 and 7). Then, we assigned criticality ranks to nodes in decreasing order of worst-case ITDs (Section III-D). Next, we analyzed the effect of varying attack duration T_a on the worst-case ITDs and criticality ranks of nodes. We showed that with increase in T_a, the worst-case ITDs corresponding to all nodes increase monotonically and eventually, as T_a approaches infinity, they converge to the same value (Theorem 9). We also showed with an example that the order of criticality ranks of nodes may change with T_a. Finally, we demonstrated with simulations that the criticality ranks have strong negative correlation with node degrees. The extension of presented work to networks with complex node dynamics, is in progress.

References

[1] H. Sandberg, V. Gupta and K. Johansson, Secure Networked Control Systems, *Annual Review of Control, Robotics and Autonomous Systems*, vol. 5, pp. 445-464, 2022.

[2] Cybersecurity and Infrastructure Security Agency, ICS Alert (IR-ALERT-H-16-056-01): Cyber-attack Against Ukrainian Critical Infrastructure, 2016.

[3] D. Kushner, The real story of Stuxnet, *IEEE Spectrum*, vol. 50, no. 3, pp. 48-53, 2013.

[4] K. Oh, M. Park, H. Ahn, A survey of multi-agent formation control, *Automatica*, vol. 53, pp 424-440, 2015.

[5] J. Wang and N. Elia, Control approach to distributed optimization, *Annual Allerton Conference on Communication, Control and Computing*, Monticello, USA, 2010.

[6] R. Olfati-Saber and J. Shamma, Consensus filters for sensor networks and distributed sensor fusion, *IEEE Conference on Decision and Control*, Seville, Spain 2005.

[7] F. Pasqualetti, A. Bicchi and F. Bullo, Consensus computation in unreliable networks: A system theoretic approach, *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90-104, 2011.

[8] I. Shames, A. Teixeira, H. Sandberg, and K. Johansson, Distributed fault detection for interconnected second-order systems, *Automatica*, vol. 47, no. 12, pp. 2757-2764, 2011.

[9] J. Milosevic, Security metrics and allocation of security resources for control systems, PhD dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2020.

[10] A. Khanafer, B. Touri and T. Başar, Consensus in the presence of an adversary, *IFAC Proceedings Volumes*, vol. 45, no. 26, pp. 276-281, 2012.

[11] S. Etesami, Consensus under network interruption and effective resistance interdiction, *American Control Conference*, New Orleans, USA, 2021.

[12] R. Gopika, A. Sharma and R. Warier, Bipartite consensus in the presence of denial of service adversary, *IFAC-PapersOnLine*, vol. 55, no. 1, pp. 771-776, 2022.

[13] H. Liu, X. Cao, J. He, P. Cheng, C. Li, J. Chen and Y. Sun, Distributed identification of the most critical node for average consensus, *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4315-4328, 2015.

[14] A. Khanafer and T. Başar, Robust distributed averaging: When are potential-theoretic strategies optimal ?, *IEEE Transactions on Automatic Control*, vol. 61, no. 7, pp. 1767-1779, 2015.

[15] Y. Nugraha, A. Cetinkaya, T. Hayakawa, H. Ishii and Q. Zhu, Dynamic resilient network games with applications to multi-agent consensus, *IEEE Transactions on Control of Network Systems*, vol. 8, no. 1, pp. 246-259, 2020.

[16] M. Pirani, E. Nekouei, H. Sandberg and K. Johansson, A graph-theoretic equilibrium analysis of attacker-defender game on consensus dynamics under $\mathcal{H}_2$ performance metric, *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 1991-2000, 2020.

[17] K. Vamvoudakis and J. Hespanha, Game-theory-based consensus learning of double-integrator agents in the presence of worst-case adversaries, *Journal of Optimization Theory and Applications*, vol. 177, no. 1, pp. 222-253, 2018.

[18] Z. Wan, Y. Mahajan, B. Kang, T. Moore and J. Cho, A survey on centrality metrics and their network resilience analysis, *IEEE Access*, vol. 9, pp. 104773-104819, 2021.

[19] H. Sandberg, A. Teixeira, and K. Johansson, On security indices for state estimators in power networks, *First Workshop On Secure Control Systems*, Stockholm, Sweden, 2010.

[20] F. Pasqualetti, S. Zampieri and F. Bullo, Controllability metrics, limitations and algorithms for complex networks, *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 40-52, 2014.

[21] A. Clark, L. Bushnell, and R. Poovendran, A supermodular optimization framework for leader selection under link noise in linear multi-agent systems, *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 283-296, 2013.

[22] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D. Hwang, Complex networks: Structure and dynamics, *Physics Reports*, vol. 424, no. 4-5, pp. 175-308, 2006.

[23] A. Jadbabaie, J. Lin and A. Morse, Coordination of groups of mobile autonomous agents using nearest neighbor rules, *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988-1001, 2003.

[24] K. Hoffman and R. Kunze, Linear Algebra, *Prentice-Hall*, 1971.

[25] R. Bapat, Graphs and Matrices, *Springer*, 2014.

[26] R. Bhatia, Matrix Analysis, *Springer*, 1997.