Securing Mobile Robotic Networks Against Replacement Attack

Yushan Li†, *Member, IEEE*, Jianping He†, *Senior Member, IEEE*, Cailian Chen†, *Senior Member, IEEE*, Xinping Guan†, *Fellow, IEEE*, and Lin Cai‡, *Fellow, IEEE*

***Abstract*—The security of mobile robotic networks (MRNs) has been an active research topic in recent years. This paper aims to secure the ubiquitous formation control of MRNs against the replacement attack, where an external robot can replace a formation robot by compromising the communication and physically interfering with the victim simultaneously. To counter this advanced attack, the novel idea of this work is to leverage the physical proximity of the formation shape and the interaction topology among robots for defense design. First, from the physical proximity perspective, we propose the convex neighbor polygon (CNP) to capture the geometric characteristic of the formation robots, and design a CNP-based security mechanism for the robots during the replacement attack. Then, from the interaction perspective, we introduce the indirect controllability to characterize the possibility that the attacker leverages the interaction between robots to deactivate the CNP mechanism, and establish the conditions regarding the topology structure and the attack input to counter the replacement. Finally, based on the obtained conditions, we demonstrate how to design the initiatory topology among the formation robots to enhance the defense performance. Comparative simulations verify the effectiveness of the proposed method.**

***Index Terms*—Mobile robotic network, formation control, replacement attack, cyber physical system security.**

I. Introduction

The latest advances in robotics, control, and communication technologies have raised a surge of interest in mobile robotic networks (MRNs). Compared with a single robot, MRNs can tackle complex tasks in a coordinated and parallel fashion, where formation control serves as a fundamental technique to maintain a geometric shape [1]. They are deployed in numerous practical applications, e.g., environment exploitation and distributed surveillance [2].

As a typical type of cyber-physical systems (CPSs), MRNs have not only the networked characteristic as CPSs but also the special physical moving ability. Nevertheless, their interconnected characteristic in cyberspace and openness in the environment introduce severe vulnerabilities in adversarial scenarios [3]–[5]. For example, a malicious attacker can block the communication channels (Denial-of-Service attacks) or manipulate the process data of nodes (false-data injection attacks) [6]–[9], severely degrading the system performance. In this paper, we aim to secure the fundamental formation control of MRN from a replacement attack, where a malicious robot aims to physically replace a victim robot of the formation. Once the attack is achieved, the malicious robot can update its states arbitrarily and endanger the whole system by spreading abnormal information.

A. Motivations

The considered problem is motivated by two aspects. First, most existing CPS security works focus on detecting and withstanding attacks from cyberspace. Indeed, the physical characteristics of MRNs are also promising to be leveraged to prevent cyber-attacks [10], e.g., a robot can use onboard sensors to verify the actual states of its neighbors even if the received information from communication channels are false. However, attack threats from physical space are rarely studied. It is more stealthy for an attacker to replace a robot in a cyber-physical hybrid way than merely manipulating its transmitted information in cyberspace. Second, the considered replacement attack in MRNs can be regarded as the primary stage to make a misbehaving (or Byzantine) node [11], [12], a typical attack scenario in CPS security. With the physical defense design that exploits the mobility of the robots to constrain the attack capability, the security threats brought by a misbehaving node can be further eliminated along with conventional cyber security methods.

B. Related Works

The security issues of MRNs have been extensively investigated in the literature [13], [14]. For example, [15], [16] focused on the secure rendezvous problem for multiple robots considering abnormal nodes in the network. The connectivity maintenance problem of the topology of MRNs was investigated in [17]–[19]. The works [20], [21] investigated how to ensure the cooperation performance of MRNs when some communication links failed. A few works can also be found that are dedicated to countering attacks from physical space, e.g., spoofing/disrupting onboard sensors by interfering with the physical sensing mechanism [22]–[24]. Some recent works have revealed that the attacker can also infer the internal topology of the MRN [25], or learn the interaction mechanism of avoiding obstacles [26] from physical observations. These developments will better support the implementation of the considered replacement attack.

This work was supported by the National Natural Science Foundation of China under Grant 62373247, 62025305, 62432009, 61933009, and 92167205. *(Corresponding author: Jianping He.)*

†: The authors are with the Dept. of Automation, Shanghai Jiao Tong University, the Key Laboratory of System Control and Information Processing, Ministry of Education of China, and Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai, 200240, China. E-mail address: {yushan_li, jphe, cailianchen, xpguan}@sjtu.edu.cn.

‡: The authors are with the Dept. of Electrical and Computer Engineering, University of Victoria, BC, Canada. Email address: cai@ece.uvic.ca.

It is worth noting that the considered attack has a close relationship with the conventional resilient consensus problem, where some nodes are hijacked by external attackers and their states can be maliciously manipulated. The resilient defense methods aim to make the normal nodes reach consensus, where each node ignores suspicious values sent by its neighbors during the interaction process [27], [28]. The key of these methods is generally built upon certain topology conditions of the network to tolerate the misbehaving nodes, e.g., r-connectivity [29], r-robustness [30] and graph isolability [31]. Despite the fruitful achievements in this direction, the resilient methods mainly focus on *how to overcome the malicious influences after the nodes are compromised*, and they may fail to work when the above topology conditions do not hold [32]. By contrast, this paper focuses on the previous stage before the nodes are hijacked and aims to answer *how to prevent a robot from being replaced and becoming malicious*. Therefore, the proposed method is not an alternative to the resilient consensus methods, and can comprehensively enhance the security of MRNs together with the latter ones.

C. Contributions

In this paper, we propose a physical proximity based defense method to secure a robot in the MRN from being replaced by an external attack robot. Specifically, the attacker is not only consistently spoofing the identity of the victim robot from cyberspace, but also physically approaching it to confuse the authentication of normal robots, as shown in Fig. 1. To address this challenging problem, our key insight is to exploit the physical proximity characteristic of the formation shape and leverage the interaction among robots for real-time defense. The contributions of this paper are summarized as follows.

- We investigate the secure formation control problem that protects a robot in the MRN from being replaced by an external robot. Specifically, we exploit the characteristics of the formation robots' geometric shape and interaction structure to constrain the attacker's capability. The notion of the convex neighbor polygon (CNP) is proposed to explicitly characterize the physical proximity of MRNs under formation control.
- Based on the proposed CNP, we design a security mechanism for the formation robots when they cannot distinguish the attacker and the victim robot. Then, we introduce the indirect controllability to characterize how the robots will be affected by the attacker. The sufficient and necessary conditions of indirect controllability are derived from the formation topology and the attack input.
- By considering the effects of indirect controllability on forming a CNP, we define the dominantly feasible replacement to characterize the risk for a robot to be replaced. We also provide the topology design guidance that constrains the attack capability from the defense perspective. Extensions to distributed communication cases are discussed. Simulations demonstrate the effectiveness of the proposed security mechanism.

The remainder of this paper is organized as follows. In Section II, some preliminaries of the MRN are presented.

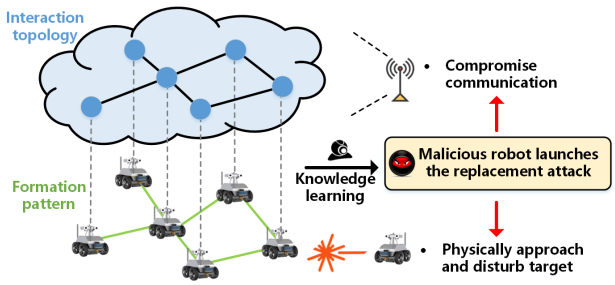


Fig. 1. Illustration of the replacement attack against an MRN.

TABLE I
NOTATION DEFINITIONS

Symbol	Definition
r_a, r_v, r_i	the abbreviation of the attacker, victim, robot i
z_a, z_v, z_i	the state of r_a, r_v, r_i
$\tilde{z}_i$	the state of r_i when the MRN is under attack
$\mathbf{z}_a$, $\mathbf{z}_v$, $\mathbf{z}_i$	the two-dimensional position of r_a, r_v, r_i
$\mathbf{z}_s$	the two-dimensional position of the base station r_s
u_e, u_c	the velocity output of $g(\cdot)$, and the formation leader
h	the shape configuration vector of formation robots
$d_{i,j}$	the distance between the positions of r_i and r_j
$\mathcal{P}(\mathbf{z}, R)$	the R-disk region center at the position $\mathbf{z}$
$\mathcal{P}_c(\mathcal{C}_v^p)$	the convex neighbor polygon of r_v
A, L	the original adjacency and Laplacian matrices
$\tilde{A}$, $\tilde{L}$	the adjacency and Laplacian matrices under attack
p_ℓ, q_ℓ	the ℓ-th left and right eigenvectors of L
$\tilde{p}_\ell$, $\tilde{q}_\ell$	the ℓ-th left and right eigenvectors of $\tilde{L}$

The CNP-based security mechanism is proposed in Section III, along with its performance analysis and parameter design. Simulation results are shown in Section IV. Finally, Section V concludes the paper.

II. Preliminaries and Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph that models an MRN, where $\mathcal{V} = \{1, \cdots, n\}$ is the finite set of nodes (i.e., robots) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of interaction edges. An edge $(i, j) \in \mathcal{E}$ indicates that i will receive information from j. The weighted adjacency matrix $A = [a_{ij}]_{n\times n}$ of $\mathcal{G}$ is defined such that $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$, and $a_{ij} = 0$ otherwise. Denote $\mathcal{N}_i^{in} = \{j \in \mathcal{V} : a_{ij} > 0\}$ and $\mathcal{N}_i^{out} = \{j \in \mathcal{V} : a_{ji} > 0\}$ as the in-neighbor and out-neighbor sets of i, respectively. A node is called a source node if it has no in-neighbors, and a root node is a source node from which all other nodes can be reached through directed paths.

For ease of notation, robot i (j) in the formation, the malicious attack robot, and the victim robot are denoted as r_i (r_j), r_a, and r_v, respectively. Denote by $\mathbf{1}$ and $\mathbf{0}$ the all-one and all-zero vectors with compatible dimensions, respectively. Let the bold font variable $\mathbf{z}_i = [z_i^x, z_i^y]^\intercal$ be the two-dimensional position vector of r_i. To simplify expressions, we omit the superscripts x and y in notations z_i^x and z_i^y, and directly use the non-bold font z_i to describe the state of r_i in one dimension. If the original variables concerning the node state and structure are denoted with a superscript $\tilde{\cdot}$, they represent the same meanings in attack situations. The three specific state

notations z_a, z_v, z_s along with their position versions will not be denoted by the superscript $\tilde{}$. The distance between r_i and r_j is represented by $d_{i,j}(t) = \|\mathbf{z}_i(t) - \mathbf{z}_j(t)\|_2$. Some commonly used symbols are summarized in Table I.

A. Formation Control

Consider that the MRN implements a go-to-goal task with a preset formation shape. Let $h_i \in \mathbb{R}$ be the desired relative displacement between r_i and a reference point in one dimension. Note that the relative displacements $h_{ij} = h_j - h_i$ for all $i, j \in \mathcal{V}$ characterize the desired stable formation shape, which will not change with the reference point selection. Besides, one robot is usually specified as the leader with an extra velocity input to dynamically guide the formation motion. The following assumption is made throughout this paper.

Assumption 1. *Suppose $\mathcal{G}$ has a unique spanning tree, and the root node, r_n, is taken as the leader and the reference node. The velocity of r_n is constant, i.e., $u_n = u_c$.*

For arbitrary follower robot $i \in \mathcal{V}\backslash\{n\}$, its dynamics in each dimension is described by

$$\dot{z}_i = \sum\nolimits_{j\in\mathcal{N}_i^{in}} a_{ij}(z_j - z_i - h_{ij}) + u_i, \tag{1}$$

where u_i is the tracking input about r_i's neighbors and is used to ensure the robots move in an unified velocity. Numerous designs of u_i with convergence guarantees can be found in the literature. For example, one can set u_i by [33]

$$u_i = \beta \operatorname{sgn}\left(\sum\nolimits_{j\in\mathcal{N}_i^{in}} a_{ij}(z_j - z_i - h_{ij})\right), \tag{2}$$

where β is a positive constant larger than the maximum velocity, and $\operatorname{sgn}(\cdot)$ is the signum function. Note that the detailed design of u_i is not the focus of this paper, and in the following sections, we will simplify it without harming the result analysis. Next, denote by $L = \operatorname{diag}\{A\mathbf{1}\} - A$ the Laplacian matrix of $\mathcal{G}$. Assume the eigenvalues of L are distinct and ordered as $|\lambda_1| \le |\lambda_2| \le \cdots \le |\lambda_n|$. Let $p_i \in \mathbb{R}^n$ and $q_i \in \mathbb{R}^n$ be the left and right eigenvectors of λ_i, satisfying

$$\begin{cases} p_i^\intercal(\lambda_i I - L) = 0,\ (\lambda_i I - L)q_i = 0 \\ p_i^\intercal q_i = 1,\ p_i^\intercal q_j = 0\ (i \neq j) \end{cases}. \tag{3}$$

Then, the global form of the system dynamics is given by

$$\dot{z}(t) = -Lz(t) + Lh + u, \tag{4}$$

where $u = [u_1, \cdots, u_n]^\intercal \in \mathbb{R}^n$. Note that $L\mathbf{1} = 0$ always holds, and thus $\lambda_1 = 0$. For simplicity without losing generality, take $q_1 = \mathbf{1}$ and $p_1 = [p_{11} \cdots p_{1n}]^\intercal$.

B. Interaction Modeling for MRNs

To form the desired shape and run safely in the environment, the robots require some interaction abilities to support their operations. The following assumption is made for the robots' capabilities, whose details will be introduced below.

Assumption 2. *The formation robots adopt bounded disk ranges for communication, and are equipped with onboard sensors to detect and avoid obstacles in the environment.*

First, for the communication aspect, the robots need to exchange information (like position, speed, or ID) via wireless communication (e.g., ZigBee, WiFi, LTE, etc). The well-known disk model [2] is commonly used to describe the bounded communication range between robots. Specially, the R-disk region of a position vector $\mathbf{z}$ is defined as

$$\mathcal{P}(\mathbf{z}, R) = \{\mathbf{z}_0 : \|\mathbf{z}_0 - \mathbf{z}\|_2 < R\}. \tag{5}$$

Besides, the communication architecture in MRNs can be generally classified into two types: i) centralized form with a base station robot r_s supporting the communication of the system, and ii) distributed form where the robots independently communicate with each other [34]. Then, let R_c be the communication range between two robots, and their positions in normal situations satisfy

$$\begin{cases} \mathbf{z}_i, \mathbf{z}_j \in \mathcal{P}(\mathbf{z}_s, R_c), \text{ in centralized case} \\ \mathbf{z}_j \in \mathcal{P}(\mathbf{z}_i, R_c), \text{ in distributed case} \end{cases}, \tag{6}$$

where $\mathbf{z}_s$ is the position vector of r_s.

For the obstacle detection and avoidance aspect, this function is supported by the onboard sensors (such as LiDAR, infrared, and ultrasonic sensors), which can detect the relative distances and angles with obstacles in the environment. For instance, a rotating LiDAR provides 360-degree coverage by measuring the time of flight of laser pulses to detect changes in the angle of arrival. Note that the obstacle-avoidance behavior is mainly determined by the relative motion states between a robot and the obstacle (such as relative position and velocity) [35]. Let $\tilde{z}_j$ be the system state when a robot r_j in the formation is influenced by an obstacle, and the dynamics of the robots are updated by

$$\dot{\tilde{z}}_j = g(z_{\text{ob}} - \tilde{z}_j, z_{j*} - \tilde{z}_j, v_j),\ \mathbf{z}_{\text{ob}} \in \mathcal{P}(\tilde{\mathbf{z}}_j, R_o), \tag{7}$$

where z_{j*} is the desired goal of r_j, z_{ob} is the position state of the obstacle, and R_o is the trigger range to avoid obstacles. The aim of $g(\cdot)$ is to make the robot away from the obstacle while approaching the goal position. It is generally required in practice that $2R_o < h_{ij} < R_c$ ($i, j \in \mathcal{V}$ and $i \neq j$), such that a stable formation shape can be achieved. Regardless of the detailed designs of g, its response magnitude is strictly bounded due to energy constraints in practice, satisfying

$$0 \le |g(\cdot)| \le u_e^{\max}, \tag{8}$$

where $0 < u_e^{\max} < \infty$ is the magnitude bound of g.

C. Attack Modeling

Consider an attacker r_a who has no prior knowledge of the formation, and it aims to physically replace one of the formation robots r_v and gain (partial) control over other formation robots. The whole attack process is conducted in a cyber-physical hybrid manner and mainly contains two phases: the formation knowledge learning and the replacement implementation. Note that an exact mathematical expression for the above process is hard to write, and we present the details of each phase as follows.

Phase 1: Knowledge learning. Since r_a does not have any prior knowledge of the system, it needs to actively learn

them to support the final attack, e.g., communication topology L and the obstacle-avoidance mechanism $g(\cdot)$. Recent works [25], [26] have investigated to fulfill such learning tasks based on external observations. The key insights are i) extracting the communication relationships from the shape-forming process of the MRN, and ii) revealing the obstacle-avoidance mechanism from the triggered position-velocity variations when obstacles occur. Note that if the shape of the MRN is already formed, the attacker can actively interfere with the robots to break the shape and collect the observations for topology inference. Specifically, with the learned topology, r_a can further use popular critical node analysis with specific performance objectives [36] to determine the victim r_v to be replaced, e.g., the robot with the maximum number of out-neighbors. The details of this phase are omitted here.

Phase 2: Replacement implementation. In this phase, r_a implements the attack from both cyber and physical space, as described by the following assumption.

Assumption 3. *In the cyber space, r_a can block the communication of r_v and spoof the identity of r_v. In the physical space, r_a can interfere with r_v closely to trigger the obstacle-avoidance mechanism $g(\cdot)$ of r_v.*

Note that the first point for r_a in Assumption 3 aims to significantly impair r_v's ability to transmit data and fake r_v's identity. This process is reasonable and feasible in practice due to the vulnerabilities of wireless communication architecture, such as using jamming and denial-of-service (DoS) attacks to block the communication, using man-in-the-middle attacks to modify the transmitted messages, and broadcasting faked messages with r_v's ID to other robots. Taking a WiFi network an instance, the attacker can fabricate the address resolution protocol (ARP) cache entries of r_v after it receives the ARP reply from the neighbor robots. We refer to [4], [37]–[39] for detailed implementation examples of compromising the communication network.

The second point in Assumption 3 is motivated by the physical sensing abilities of robots for potential verification. For instance, if r_a is not physically detected by the onboard sensors, the formation will recognize no actual robot at the position reported in the faked communication message and will not use this information for state updating, making the replacement directly fail. Besides, by keeping r_a close to r_v, other formation robots can only detect the two close objects by physical sensing, without knowing which one is the victim or the attacker. Hence, from the attack perspective, it is desired to make r_a physically close to r_v and detectable. Notice that the obstacle detection by the physical sensing can be done by all formation robots. For the non-victim robots, their avoidance behaviors will not be frequently triggered by r_a (i.e., $g(\cdot)=0$), because they are mainly driven to form the desired shape and r_a is only very close to r_v during the attack.

Mathematically, let u_a be the input of r_a's movement and it is obtained by solving the following optimization problem

$$\min_{u_a(t)} \quad F(u_a(t), z_v(t)) \tag{9a}$$

$$\text{s.t.} \quad \mathbf{z}_a(t) \in \mathcal{P}(\mathbf{z}_v(t), R_o), \tag{9b}$$

$$\dot{z}_v = g(z_a(u_a) - z_v, z_{v*} - z_v, v_v), \tag{9c}$$

where v_v is the velocity of r_v, and $F(u_a(t), z_v(t))$ is the objective function that describes the attack costs. Specifically, (9b) requires that r_a needs to closely interfere with r_v such that r_v's obstacle-avoidance behavior is triggered, and (9c) represents r_v's dynamics under attack. Since r_a can comprise r_v's communication for neighbor recognition and the physical sensing of a robot is only for the obstacle detection (not the identity recognition), other formation robots will take r_a that broadcasts faked messages of r_v as the true r_v.

D. Problem of Interest

Based on the previous formulation, this paper aims to design a security mechanism for the formation robots to counter the replacement attack. Specifically, we mainly focus on the following two aspects:

- how to exploit the attack's dependency on the physical characteristics of the formation shape and the topology structure from the attacker's perspective.
- how to utilize the obtained analytical relationships to guide the defense design and explicitly characterize its performance.

To address the above issues, we will introduce the notions of convex neighbor polygon and indirect controllability to facilitate the attack condition and defense performance analysis.

III. CNP-based Security Mechanism: Design and Analysis

In this section, we develop a physical proximity based security mechanism. First, we provide a detailed design for the centralized communication case of MRNs. Then, we analyze the attack capability and derive the condition to disable the replacement attack. Furthermore, we present the initiatory topology design to enhance the defense performance and extend the method to distributed communication cases.

A. Convex Neighbor Polygon Based Security Mechanism

As introduced before, during the formation control process, each robot must take action based on the information from its in-neighbors, and intuitively they exhibit a sub-formation shape in real-time. This intrinsic feature essentially serves as a physical constraint of the MRN and can be used as an internal criterion for attack countermeasure design.

To begin with, we define the CNP of a robot as below.

Definition 1 (Convex neighbor polygon of a robot)**.** *Given $v \in \mathcal{V}$ and $\mathcal{C}_v^p \subseteq \{v \cup \mathcal{N}_v^{out}\}$, $\mathcal{P}_c(\mathcal{C}_v^p)$ is called a convex neighbor polygon of r_v, if i) $\mathcal{P}_c(\mathcal{C}_v^p)$ is the maximum convex polygon region with the positions of some nodes in $\mathcal{C}_v^p$ being the vertexes, satisfying $\mathbf{z}_i \in \mathcal{P}_c(\mathcal{C}_v^p), \forall i \in \mathcal{C}_v^p$, and ii) $\mathbf{z}_v$ is a vertex of $\mathcal{P}_c(\mathcal{C}_v^p)$.*

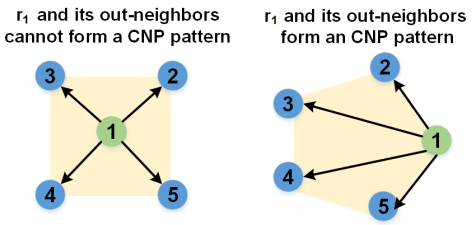


Fig. 2. Examples of the proposed CNP pattern by taking r_1 as r_v. In the left case, $\{r_1, \mathcal{N}_1^{out}\}$ constitute a quadrangle region, but r_1 is inside the region and not a vertex of the polygon, thus $\mathcal{P}_c(\mathcal{C}_v^p)$ is absent. In the right case, $\{r_1, \mathcal{N}_1^{out}\}$ constitute a pentagon region with r_1 being a vertex, thus the CNP pattern $\mathcal{P}_c(\mathcal{C}_v^p)$ exists.

The CNP pattern $\mathcal{P}_c(\mathcal{C}_v^p)$ presents a special sub-formation shape of a robot and its out-neighbors (Fig. 2 provides an illustrative example). If r_v has only one out-neighbor, then $\mathcal{P}_c(\mathcal{C}_v^p)$ always exists and the region covered by $\mathcal{P}_c(\mathcal{C}_v^p)$ is the segment between r_v and the neighbor. Based on the CNP, we propose a physical proximity based security mechanism to counter the replacement attack. Recall that r_a's attack needs to be within the trigger range of r_v's obstacle-avoidance behavior (i.e., $\|\mathbf{z}_a - \mathbf{z}_v\| \leq R_o$). Considering r_a and r_v are both within the communication range of r_i and r_i cannot discriminate against them due to their proximity, the following rule is used to update the state of a non-victim robot r_i

$$\dot{\tilde{z}}_i = \begin{cases} \sum\limits_{\ell \in \mathcal{N}_i^{in} \setminus \{v\}} a_{i\ell}(\tilde{z}_\ell - \tilde{z}_i - h_{i\ell}) + a_{iv}\bar{z}_v + u_i, & i \in \mathcal{N}_v^{out} \\ \sum\limits_{\ell \in \mathcal{N}_i^{in}} a_{i\ell}(\tilde{z}_\ell - \tilde{z}_i - h_{i\ell}) + u_i, & i \notin \mathcal{N}_v^{out} \end{cases}, \quad (10)$$

where $\bar{z}_v = (1-\alpha)z_v + \alpha z_a - \tilde{z}_i - h_{iv}$ and $\alpha \in [0,1]$ is a weight parameter. From defense perspective, (10) reversely utilizes r_a's proximity limitation $\|\mathbf{z}_a - \mathbf{z}_v\| \leq R_o$ to avoid confusion of selecting the true r_v, while alleviating the risk directly influenced by r_a.

Based on (10), we present the security rules for the system to determine whether r_a is the true r_v. Given the preset time interval $t_l > 0$ for validation, r_a will be taken as the true r_v at time t_0 by the MRN when the following conditions hold

- **CNP-based security rules**

$$\forall i \in \mathcal{N}_v^{out},\ d_{i,a}(t) < d_{i,v}(t),\ t \geq t_0 - t_l, \quad (11a)$$
$$\mathbf{z}_a(t) \in \mathcal{P}(\mathbf{z}_s(t), R_c),\ t \geq t_0 - t_l, \quad (11b)$$
$$\mathbf{z}_v(t) \notin \mathcal{P}(\mathbf{z}_s(t), R_c),\ t \geq t_0. \quad (11c)$$

The first rule (11a) requires r_a to be closer to all robots in $\mathcal{N}_v^{out}$ than r_v, and the existence confirmation of this CNP pattern is conducted by the base station robot r_s. The latter two rules (11b)-(11c) illustrate that r_a needs to be within while r_v needs to be outside the communication range of r_s, respectively. Only when the three rules are satisfied together, r_s will convey the message that r_a is the true r_v to $\mathcal{N}_v^{out}$ (i.e., the replacement is achieved). The proposed CNP-based security mechanism is reasonable due to the intrinsic proximity characteristic of a formation shape and the distance limitation in their communication.

According to the above analysis, we summarize the whole proposed security mechanism as Algorithm 1. When r_v is under the replacement attack, all its out-neighbors cannot distinguish the true r_v and r_a due to the closeness between r_a and r_v, and they will adopt (10) for updating the states [Line 3-6]. The replacement for r_a can be achieved only when r_a is closer to $\mathcal{N}_v^{out}$ than r_v and r_a drives r_v out of the station robot's communication range [Line 8-13]. In the algorithm, the checking period $t_\Delta > 0$ is used to denote the validation moment, and the auxiliary variable t_c serves as a time counter to determine whether the proposed rules hold for at least a time period of t_l. Note that there is a special case where the base station robot r_s is selected as the victim. In this situation, the attacker r_a further needs to fake the confirmation message to other robotics when the CNP-based security rule is violated, which requires additional communication costs.

Algorithm 1 CNP-based Security Mechanism

Input: The initial time under attack t_a, the checking period t_Δ, and the preset time interval t_l.

1: Initialization: $t_c = 0$, $k = 0$, $t_k = t_a$;
2: **while** r_a is indistinguishable from r_v **do**
3: $\bar{z}_v(t_k) = \alpha z_v(t_k) + (1-\alpha)z_a(t_k) - \tilde{z}_i(t_k) - h_{iv}$;
4: **for** $i \in \mathcal{N}_v^{out}$ **do**
5: $\dot{\tilde{z}}_i(t_k) = \sum\limits_{\ell \in \mathcal{N}_i^{in} \setminus \{v\}} a_{i\ell}(\tilde{z}_\ell(t_k) - \tilde{z}_i(t_k) - h_{i\ell}) + a_{iv}\bar{z}_v(t_k) + u_i(t_k)$;
6: **end for**
7: **if** $\forall i \in \mathcal{N}_v^{out}$, $d_{i,a}(t_k) \leq d_{i,v}(t_k)$, $\mathbf{z}_a(t_k) \in \mathcal{P}(\tilde{\mathbf{z}}_i(t_k), R_c)$ **then**
8: $t_c = t_c + 1$;
9: **if** $t_c \geq \lceil \frac{t_l}{t_\Delta} \rceil$ **then**
10: **if** $\mathbf{z}_v(t_k) \notin \mathcal{P}(\mathbf{z}_s(t_k), R_c)$ **then**
11: break;
12: **end if**
13: **end if**
14: **else**
15: $t_c = 0$;
16: **end if**
17: $t_{k+1} = t_k + t_\Delta$, $k = k + 1$;
18: **end while**
19: r_a is recognized as the true r_v;

Next, we characterize the proximity properties of a CNP.

Theorem 1. *Given the positions of r_v and $\mathcal{N}_v^{out}$, define the intersected region $\mathcal{P}_v^f = \bigcap\limits_{i \in \mathcal{N}_v^{out}} \mathcal{P}(\mathbf{z}_i, d_{i,v})$. If the CNP pattern $\mathcal{P}_c(\mathcal{C}_v^p)$ exists, then it holds that for arbitrary $\mathbf{z} \in \mathcal{P}_v^f$,*

$$\|\mathbf{z} - \mathbf{z}_i\|_2 < \|\mathbf{z}_v - \mathbf{z}_i\|_2,\ i \in \mathcal{N}_v^{out}. \quad (12)$$

Proof. The key to proving this result relies on demonstrating the non-emptiness of $\mathcal{P}_v^f$.

Note that this property obviously holds when the cardinality number $|\mathcal{C}_v^p| = 2$, thus we consider nontrivial cases where $|\mathcal{C}_v^p| \geq 3$. First, when $|\mathcal{C}_v^p| = 3$, there are only two vertexes adjacent to v, denoted as i_1 and i_2, respectively. In this case, $\mathcal{P}(\mathbf{z}_{i_1}, d_{i_1,v}) \cap \mathcal{P}(\mathbf{z}_{i_2}, d_{i_2,v}) = \emptyset$ if and only if the scalar states z_v, z_{i_2} and z_{i_2} are linearly dependent, which contradicts with the convex vertex condition of v. Therefore, it follows that

$$\mathcal{P}_v(d_{i_1,v}, d_{i_2,v}) \triangleq \mathcal{P}(\mathbf{z}_{i_1}, d_{i_1,v}) \cap \mathcal{P}(\mathbf{z}_{i_2}, d_{i_2,v}) \neq \emptyset. \quad (13)$$

Next, we turn to the general situation when $|\mathcal{C}_v^p| > 3$. For $i_3 \in \{\mathcal{C}_v^p \setminus \{v \cup i_1 \cup i_2\}\}$, it follows from (13) that

$$\mathcal{P}_v(d_{i_1,v}, d_{i_3,v}) \neq \emptyset,\ \mathcal{P}_v(d_{i_2,v}, d_{i_3,v}) \neq \emptyset. \quad (14)$$

Then, we need to prove

$$\mathcal{P}_v(d_{i_1,v}, d_{i_3,v}) \cap \mathcal{P}_v(d_{i_2,v}, d_{i_3,v}) \neq \emptyset. \tag{15}$$

We use the proof by contradiction and suppose $\mathcal{P}_v(d_{i_1,v}, d_{i_3,v}) \cap \mathcal{P}_v(d_{i_2,v}, d_{i_3,v}) = \emptyset$. This case is equivalent to the states of four vertexes satisfying

$$z_{i_3} - z_v = \beta_1(z_{i_1} - z_v) + \beta_2(z_{i_2} - z_v), \tag{16}$$

where $\beta_1 \leq 0$, $\beta_2 \leq 0$, $\beta_1\beta_2 = 0$, and they are not both zero at the same time. Note that $\beta_1 = 0$ (or $\beta_2 = 0$) indicates i_1 (or i_2), i_3 and v are on the same line. Consequently, we have

$$z_v = \frac{z_{i_3} - \beta_1 z_{i_1} - \beta_2 z_{i_2}}{1 - \beta_1 - \beta_2}, \tag{17}$$

which means z_v is a convex combination of $\{z_{i_1}, z_{i_3}\}$ or $\{z_{i_2}, z_{i_3}\}$, and thus contradicts with the convex vertex condition of v in $\mathcal{P}_c(\mathcal{C}_v^p)$. Therefore, if $\mathcal{P}_v(d_{i_1,v}, d_{i_3,v}) \neq \emptyset$ and $\mathcal{P}_v(d_{i_2,v}, d_{i_3,v}) \neq \emptyset$, then $\mathcal{P}_v(d_{i_1,v}, d_{i_3,v}) \cap \mathcal{P}_v(d_{i_2,v}, d_{i_3,v}) \neq \emptyset$. By analogy, we can continue to use this transitivity property and obtain

$$\mathcal{P}_v^{f_0} \triangleq \bigcap_{j \in \{\mathcal{C}_v^p \setminus v\}} \mathcal{P}(\mathbf{z}_j, d_{j,v}) \neq \emptyset. \tag{18}$$

If $\exists i \in \mathcal{N}_v^{out} \setminus \mathcal{C}_v^p$, then its position also locates in the convex polygon region formed by $\mathcal{C}_v^p$, i.e., $\mathbf{z}_i \in \mathcal{P}_c(\mathcal{C}_v^p)$. Then, it follows from the convex property that

$$\mathcal{P}(\mathbf{z}_i, d_{i,v}) \cap \mathcal{P}_v^{f_0} \neq \emptyset. \tag{19}$$

Summing up (13), (18) and (19) leads to

$$\mathcal{P}_v^f = \bigcap_{i \in \mathcal{N}_v^{out}} \mathcal{P}(\mathbf{z}_i, d_{i,v}) \neq \emptyset. \tag{20}$$

Finally, applying the definition of $\mathcal{P}(\mathbf{z}_i, d_{i,v}), i \in \mathcal{N}_v^{out}$ to $\mathbf{z} \in \mathcal{P}_v^f$ further yields (12). The proof is completed. □

Theorem 1 illustrates that if r_v and its out-neighbors constitute a CNP pattern $\mathcal{P}_c(\mathcal{C}_v^p)$, it will provide an entry point for r_a to directly meet the proximity rule (11a), incurring a severe security vulnerability. In this situation, r_a can keep $\mathbf{z}_a \in \mathcal{P}_v^f$ and make the CNP-based security rules easier to meet. Nevertheless, we observe that the attack entry provided by the CNP pattern is only an external beneficial factor for r_a. Even if the CNP for r_v is absent at the beginning of the attack, it is still possible that r_a can make $\{v \cup \mathcal{N}_v^{out}\}$ form the CNP pattern during the attack process, and further deactivate the CNP-based security mechanism.

B. System Evolution Under Attack

In this part, we demonstrate that the implicit interaction characteristic of formation robots is also dominant for the proposed security mechanism, apart from the explicit proximity properties of the CNP. Specifically, we propose the indirect controllability between two robots to analyze the system performance under the replacement attack, supporting the following security design.

First, we illustrate how the system will evolve when the attack is absent.

Lemma 1. *Considering the tracking input $u_i = 0, \forall i \in \mathcal{V}\setminus\{n\}$ and Assumption 1 holds, there exists bounded error between the asymptotic and desired formation shapes, i.e.,*

$$\lim_{t\to\infty} |z_i(t) - z_j(t) - h_{ij}| = |s_i - s_j|, \tag{21}$$

where s_i (s_j) is the $i(j)$-th element of $s = \sum_{\ell=2}^{n} \frac{1}{\lambda_\ell} q_\ell p_\ell^\intercal u$. Meanwhile, all nodes have the same velocity as $t \to \infty$, i.e.,

$$\lim_{t\to\infty} |\dot{z}_i(t) - \dot{z}_j(t)| = 0. \tag{22}$$

Proof. The key point to proving this theorem is to establish the explicit expression about t. First, by simple integration, the global solution to (4) is written as

$$z(t) = e^{-Lt} z(0) + \int_0^t e^{-L(t-\tau)}(Lh + u)\mathrm{d}\tau. \tag{23}$$

Note that under Assumption 1, the eigenvalue $\lambda_1 = 0$ is unique. Similar to the proof of [40, Theorem 2.2], we consider that all other eigenvalues $\{\lambda_i, i \in \mathcal{V}\setminus\{1\}\}$ are of order one for simple expressions (the general case follows similarly). Then, the Jordan decomposition of L and e^{-Lt} can be respectively written as

$$\begin{cases} L = M\,\mathrm{diag}\{\lambda_1, \cdots, \lambda_n\}M^{-1} = \sum_{i=1}^{n} q_i p_i^\intercal \lambda_i \\ e^{-Lt} = M\,\mathrm{diag}\{e^{-\lambda_1 t}, \cdots, e^{-\lambda_n t}\}M^{-1} = \sum_{i=1}^{n} q_i p_i^\intercal e^{-\lambda_i t} \end{cases}, \tag{24}$$

where $M = [q_1, \cdots, q_n]$ and $(M^{-1})^\intercal = [p_1, \cdots, p_n]$ are the transformation matrices consisting of the right and left eigenvectors of L, respectively. It is straightforward to obtain that $MM^{-1} = \sum_{i=1}^{n} q_i p_i^\intercal = I$. Substituting (24) into (23), $z(t)$ is further written as

$$\begin{aligned} z(t) =& \left[q_1 p_1^\intercal z(0) + \sum_{i=2}^{n} q_i p_i^\intercal e^{-\lambda_i t} z(0)\right] \\ &+ \left[q_1 p_1^\intercal t + \sum_{i=2}^{n} \frac{1}{\lambda_i} q_i p_i^\intercal (1 - e^{-\lambda_i t})\right] \left(\sum_{i=1}^{n} \lambda_i q_i p_i^\intercal h + u\right) \\ =& q_1 p_1^\intercal u \cdot t + q_1 p_1^\intercal z(0) + (I - q_1 p_1^\intercal)h + \sum_{i=2}^{n} \frac{1}{\lambda_i} q_i p_i^\intercal u \\ &+ \sum_{i=2}^{n} e^{-\lambda_i t} q_i p_i^\intercal (z(0) - h) - \sum_{i=2}^{n} \frac{e^{-\lambda_i t}}{\lambda_i} q_i p_i^\intercal u, \end{aligned} \tag{25}$$

where the first equality utilizes the fact that $\lambda_1 = 0$ and $\int_0^t e^{-\lambda_i t} = (1 - e^{-\lambda_i t})/\lambda_i$ ($i \in \mathcal{V}\setminus\{1\}$), and the second equality utilizes the properties $\sum_{i=1}^{n} q_i p_i^\intercal = I$ and $\sum_{i=2}^{n} \left(q_i p_i^\intercal (\sum_{i=1}^{n} q_i p_i^\intercal)\right) = \sum_{i=2}^{n} q_i p_i^\intercal$. It is clear that the terms in the last row of (25) are all exponentially decaying with t, which is the key to achieving asymptotic convergence of the system.

Recall that under Assumption 1, the left and right eigenvectors of $\lambda_1 = 0$ can be determined as $p_1 = [0, \cdots, 0, 1]^\intercal$ and $q_1 = \mathbf{1}$, thus yielding that

$$\begin{cases} q_1 p_1^\intercal u = \mathbf{1}([0, \cdots, 0, 1] \cdot [0, \cdots, 0, u_c]^\intercal) = u_c \mathbf{1} \\ q_1 p_1^\intercal z(0) = \mathbf{1}([0, \cdots, 0, 1] \cdot [z_1, \cdots, z_n]^\intercal) = z_n(0)\mathbf{1} \end{cases}. \tag{26}$$

Therefore, substituting (26) into (25), it follows that

$$\lim_{t\to\infty}(z(t)-u_c\mathbf{1}\cdot t-z_n(0)\mathbf{1}-(I-q_1p_1^\intercal)h) = \sum\nolimits_{i=2}^{n}\frac{1}{\lambda_i}q_ip_i^\intercal u = s. \tag{27}$$

Apparently, one can conduct element-wise operation on (27) and obtain $\lim_{t\to\infty}(z_i(t)-u_ct-z_n(0)-[(I-q_1p_1^\intercal)h]_i) = s_i$. Then, subtract arbitrary two elements of (27) and one has

$$\begin{aligned}&\lim_{t\to\infty}|z_i(t)-[(I-q_1p_1^\intercal)h]_i-(z_j(t)-[(I-q_1p_1^\intercal)h]_j)|\\ =&\lim_{t\to\infty}|z_i(t)-z_j(t)-h_{ij}| = |s_i-s_j|,\end{aligned} \tag{28}$$

where the fact $[(I-q_1p_1^\intercal)h]_i = h_i - h_n$ is adopted. The first statement is proved. Meanwhile, it is deduced that the derivative of $z(t)$ satisfies

$$\lim_{t\to\infty}\dot{z}(t) = q_1p_1^\intercal u = u_c\mathbf{1}, \tag{29}$$

which means that the velocities of all robots are identical and completes the proof. □

Lemma 1 shows that even if the velocity tracking is absent in the control input of a follower, the formation robots can still keep the same velocity asymptotically but with bounded pattern error. Indeed, the resultant velocity synchronization is exactly maintained by the pattern error between robots, i.e., the pattern error works as an input to keep the velocity. Therefore, the larger the leading velocity u_c is, the larger the pattern error will be. Another common case is that each follower is aware of the leader's velocity u_c [41], i.e., $u = u_c\mathbf{1}$. In this situation, one can adopt the proof of Lemma 1 and easily obtain the pattern error vector $s = \mathbf{0}$.

Remark 1. *We observe that the absence of the velocity tracking input will not affect the final velocity synchronization. The tracking input can be regarded as an estimator of the leading velocity, which aims to eliminate the formation shape error s asymptotically. Considering exponential decaying characteristic of reaching a stable formation in both situations of $u = [0,\cdots,0,u_c]^\intercal$ and $u = u_c\mathbf{1}$, the performance of a velocity tracking term can also be bounded by an exponential decaying term, i.e., $\exists\bar{\lambda}\in(0,1)$ and $\gamma\in\mathbb{R}$, such that*

$$|u_i(t)| \le u_c(1+\gamma e^{-\bar{\lambda}t}). \tag{30}$$

Since the major impact of the attack is put on the velocity, we temporarily drop the velocity tracking term u_i ($i\in\mathcal{V}\backslash\{n\}$) for legibility.

When r_v is under attack, it can be seen as a malicious node affected by r_a. For simple expression, denote by $u_e = g(z_a(u_a)-z_v, z_{v*}-z_v, v_v)$ the velocity of r_v under attack. Notice that the primary goal of r_v's obstacle-avoidance mechanism is to produce desired u_e to keep itself from colliding with r_a. Even if there is communication-keeping considerations in $g(\cdot)$, it will not affect the replacement attack, where r_a can still expel r_v from the formation by solving the attack design problem (9). During the attack process, the attack impact on r_v will spread to $\mathcal{N}_v^{out}$, and this indirect influence is critical for physically disabling the communication between two robots. To formally describe this property, we introduce the following definition.

Definition 2 (Indirect controllability)**.** *Given a robot pair (r_i,r_j) and a target state z_c^*, r_i is indirectly controllable by r_j, if there exists a bounded control u_e on r_j such that r_i can reach z_c^* in finite time.*

Note that from a graphical perspective, the proposed indirect controllability has a close relationship with the well-documented structural controllability [42]–[44], since they both need the targeted node to be reachable from the input node. However, the indirect controllability does not require all the nodes to be reachable, which is necessary for the structural controllability. In this sense, the indirect controllability can be regarded as a weak version of the structural controllability for two nodes in a network system.

Next, we illustrate under what conditions a robot r_i is indirectly controllable by r_v. Since the dynamics of r_v is directly determined by $u_e = g(\cdot)$ when it is under attack, r_v can be treated as another source node. Then, the new adjacency matrix $\tilde{A}(v) = [\tilde{a}_{ij}(v)]$ in this situation is given by

$$\tilde{a}_{ij}(v) = \begin{cases}0\ , & \text{if } i=v,\ j\in\mathcal{N}_v^{in}\\ a_{ij}\ , & \text{otherwise}\end{cases}, \tag{31}$$

and the corresponding Laplacian $\tilde{L}(v)$ is given by

$$\tilde{L}(v) = \text{diag}\{\tilde{A}(v)\cdot\mathbf{1}\} - \tilde{A}(v). \tag{32}$$

The parentheses notation (v) in $\tilde{A}(v)$ and $\tilde{L}(v)$ is omitted in the following when no confusion is caused. Notice that there are two all-zero rows in $\tilde{A}$. Thus, the algebraic multiplicity of the zero eigenvalues for $\tilde{L}$ is $\mu = 2$. The following lemma shows how $\tilde{L}$ will influence the convergence process.

Lemma 2 (Proposition 1 in [45])**.** *Considering a dynamic model $\dot{z}^o = -\tilde{L}z^o$ with arbitrary initial state $z^o(0)$, the limit state of z^o is given by*

$$\lim_{t\to\infty}z^o(t) = \sum\nolimits_{\ell=1}^{\mu}\tilde{q}_\ell\tilde{p}_\ell^\intercal z^o(0), \tag{33}$$

where $\tilde{p}_\ell$ ($\tilde{q}_\ell$) are the distinct and linearly independent left (right) eigenvectors associated with the zero eigenvalues of $\tilde{L}$. Specifically, the vectors satisfy

$$\tilde{p}_\ell^\intercal\mathbf{1} = 1(\ell=1,\cdots,\mu),\ \text{and}\ \sum\nolimits_{\ell=1}^{\mu}\tilde{q}_\ell = \mathbf{1}. \tag{34}$$

Note that the global asymptotic state (33) in Lemma 2 does not imply all the nodes reach the same value. Next, for simple analysis, we temporarily consider that r_a makes r_v move in a constant velocity u_e, and demonstrate the conditions of r_v's indirect controllability on r_i. Recall that there are two all-zero rows in $\tilde{A}$ and the Laplacian $\tilde{L}$ is still a row-stochastic matrix. A subset $\mathcal{V}_r(i)\in\mathcal{V}$ is called a reach, if node $i\in\mathcal{V}$ is the root node of $\mathcal{V}_r(i)$, and $\mathcal{V}_r(i)\backslash\{i\}$ consists of all nodes that can be reached from node i. Then, for the graph associated with $\tilde{A}$, there are two reaches $\mathcal{V}_r(v)\in\mathcal{V}$ and $\mathcal{V}_r(n)\in\mathcal{V}$, and for ease of distinction, they are treated as the first and second reaches in $\tilde{A}$, respectively.

Theorem 2. *Consider that r_v obeys $\dot{z}_v = u_e$ and other robots obey* (10)*. Given desired state z_c^* and arbitrary initial state z_i^0, r_i is indirectly controllable by r_v if and only if*

$$\tilde{q}_1^{[i]}|u_e| > \tilde{q}_2^{[i]}|u_c|, \tag{35}$$

where the right eigenvectors $\tilde{q}_1$ and $\tilde{q}_2$ correspond to the first and second reaches, respectively, and their i-th elements satisfy $\tilde{q}_1^{[i]} > 0$, $\tilde{q}_2^{[i]} \geq 0$ and $\tilde{q}_1^{[i]} + \tilde{q}_2^{[i]} = 1$.

Proof. The basic idea of this proof is to demonstrate that r_n's influence on the motion of r_i should be larger than r_v's influence on r_i. Since the dynamic evolution of robots is determined by the velocity of r_n (as stated in Lemma 1), we directly focus on the velocities of the robots.

To begin with, we need to obtain the values of left eigenvectors associated with the zero eigenvalues of $\tilde{L}$, i.e., $\tilde{p}_\ell (\ell = 1, 2)$. By leveraging the conclusion in [46, Theorem 5.2], we have that $\tilde{p}_1^{[i]} > 0$ (or $\tilde{p}_2^{[i]} > 0$) if and only if node i is a root node of the reach $\mathcal{V}_r(v)$ (or $\mathcal{V}_r(n)$). Based on this property, the vector $\tilde{p}_1$ can be obtained by solving the following equations

$$\begin{cases} \tilde{p}_1^{[i]} = 0, \text{ if } i \text{ is not the root node in } \mathcal{V}_r(v) \\ \tilde{p}_1^\intercal \mathbf{1} = 1 \end{cases}. \tag{36}$$

Similarly, the vector $\tilde{p}_2$ satisfies the same equations in (36) except that $\mathcal{V}_r(v)$ is replaced by $\mathcal{V}_r(n)$. Hence, the solutions to $\tilde{p}_1$ and $\tilde{p}_2$ are given by

$$\begin{cases} \tilde{p}_1^{[i]} = 1, & i = v \\ \tilde{p}_1^{[i]} = 0, & i \in \mathcal{V} \backslash \{v\} \end{cases}, \text{ and } \begin{cases} \tilde{p}_2^{[i]} = 1, & i = n \\ \tilde{p}_2^{[i]} = 0, & i \in \mathcal{V} \backslash \{n\} \end{cases}. \tag{37}$$

Next, we give the explicit expressions of r_i's state evolution under attack, where dynamics of the MRN is determined by $\tilde{L}$. Recall that (25) gives the state of r_i when there is no attack. When there are two source nodes, it resembles Lemma 2 that the system evolution under attack is given by

$$\begin{aligned} \tilde{z}(t) = &\sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u} \cdot t + \sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{z}(0) + (I - \sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal) h \\ &+ \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u}}{\tilde{\lambda}_\ell} + \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal (\tilde{z}(0) - h)}{e^{\tilde{\lambda}_\ell t}} - \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u}}{\tilde{\lambda}_\ell e^{\tilde{\lambda}_\ell t}}, \end{aligned} \tag{38}$$

where $\tilde{\lambda}_\ell, \ell = 1, \cdots, n$ are the eigenvalues of $\tilde{L}$ ($\tilde{\lambda}_1 = \tilde{\lambda}_2 = 0$), and $\tilde{u} = [0, \cdots, 0, u_e, 0, \cdots, 0, u_c]^\intercal$ is the formation input under attack. Then, the derivative of $\tilde{z}(t)$ is given by

$$\dot{\tilde{z}}(t) = \sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u} + \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal (\tilde{u} - \tilde{\lambda}_\ell \tilde{z}(0) + \tilde{\lambda}_\ell h)}{e^{\tilde{\lambda}_\ell t}}. \tag{39}$$

Since the second sum term in (39) is exponentially decaying with t, its influence will be extremely small as the system runs and can be neglected. Consequently, we have

$$\lim_{t \to \infty} \dot{\tilde{z}}(t) = \sum\nolimits_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u}. \tag{40}$$

Then, substituting the solution (37) into (40), the velocity of r_i as $t \to \infty$ is given by

$$\begin{aligned} \dot{\tilde{z}}_i^\infty &= \tilde{q}_1^{[i]} (\tilde{p}_1^{[v]} u_e + \tilde{p}_1^{[n]} u_c) + \tilde{q}_2^{[i]} (\tilde{p}_2^{[v]} u_e + \tilde{p}_2^{[n]} u_c) \\ &= \tilde{q}_1^{[i]} u_e + \tilde{q}_2^{[i]} u_c, \end{aligned} \tag{41}$$

which is a convex combination of u_e and u_c.

Finally, we show that the condition (35) is sufficient and necessary for the indirect controllability based on (41). **i) Sufficiency**: when $(z_c^* - z_i^0) u_c > 0$, it means that the attack of r_a aims to strengthen the movement in the direction of the leader. In this trivial situation, arbitrary u_e satisfying $u_e u_c > 0$ is available to meet the indirect controllability. When $(z_c^* - z_i^0) u_c < 0$, it means that r_a aims to strengthen the movement in the opposite direction of u_c. Then, substituting $\tilde{q}_1^{[i]}|u_e| > \tilde{q}_2^{[i]}|u_c|$ into (41) yields that

$$(\tilde{q}_1^{[i]} u_e + \tilde{q}_2^{[i]} u_c) u_c < 0, \tag{42}$$

which indicates that r_i will directly run to z_c^*, and thus is indirectly controllable by r_v. **ii) Necessity**: when $(z_c^* - z_i^0) u_c > 0$, this case is trivial as r_v will directly go to z_c^*. When $(z_c^* - z_i^0) u_c < 0$, if r_i is directly controllable by r_v, it means that the influence of the original leadership u_c is counteracted by r_a (i.e., $\dot{\tilde{z}}_i u_e > 0$). Hence, it follows from $(\tilde{q}_1^{[i]} u_e + \tilde{q}_2^{[i]} u_c) u_e > 0$ that

$$\left| \tilde{q}_1^{[i]} u_e \right| > \left| \tilde{q}_2^{[i]} u_c \right|, \tag{43}$$

yielding (35). Finally, based on the arbitrariness of $(z_c^* - z_i^0)$, (35) is sufficient and necessary for the indirect controllability. The proof is completed. □

Theorem 2 points out the sufficient and necessary condition for indirect controllability in terms of the global topology structure of the MRN, where the eigenvectors $\{\tilde{q}_\ell\}$ play a critical role. To meet the condition, r_a is required to have the global topology of the MRN, such that the eigenvectors $\{\tilde{q}_\ell\}$ can be calculated from $\tilde{L}$.

C. Vulnerability Under the Replacement Attack

As indicated in Section III-A, if the CNP pattern for a selected r_v exists when the attack begins, r_v is highly vulnerable to the replacement attack. This is because r_a only needs to take major efforts to break the communication conditions (11b)-(11c), which can easily achieved by leveraging by the obstacle-avoidance mechanism. In this part, we focus on the vulnerability analysis when the CNP pattern is absent initially.

First, note that the exact state evolution of the MRN is affected by many practical factors (e.g., the moving strategies of the attacker and obstacles in the environment), and the formation shape may fluctuate in the non-steady stage. It is extremely hard to find an explicit dynamic model such that all the above factors could be covered. Therefore, to avoid tedious considerations, we characterize the feasibility of the replacement attack by introducing the following definition.

Definition 3 (Dominantly feasible replacement)**.** *When the CNP pattern for r_v is absent initially, an attack strategy for replacing r_v is called dominantly feasible if $\exists i \in \mathcal{N}_v^{out}$, such that r_i is not indirectly controllable by r_v, i.e., $\tilde{q}_1^{[i]}|u_e| \leq \tilde{q}_2^{[i]}|u_c|$.*

An illustration of Definition 3 along with its relationships with the CNP and the indirect controllability is shown in Fig. 3. By defining the dominantly feasible replacement, we do not mean that the strategy can achieve the replacement attack

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2025.3526002

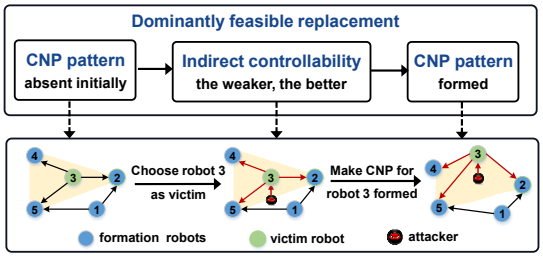


Fig. 3. Relationships between dominantly feasible replacement, CNP, and the indirect controllability. Note that the dominantly feasible replacement aims at the attack process of making the absent CNP pattern for r_v formed. The weaker the indirect controllability of r_v on $\mathcal{N}_v^{out}$ is, the better for the attack.

definitely, but enlighten the great possibility of achieving the attack. Definition 3 can be inversely interpreted as the more r_i $(i \in \mathcal{N}_v^{out})$ is indirectly controllable by r_v, the harder r_v can bypass r_i and become a vertex in a CNP pattern. In other words, if all robots in $\mathcal{N}_v^{out}$ are indirectly controllable by r_v, it is very hard for r_a to achieve the replacement attack.

Remark 2. *Although we previously suppose that the resultant obstacle-avoidance behavior $u_e = g(\cdot)$ is constant, it will not affect the existence of dominantly feasible replacement if u_e is time-varying, as long as the bound $u_e^{\max}$ and the topology satisfy*

$$\tilde{q}_1^{[i]}|u_e^{\max}| \leq \tilde{q}_2^{[i]}|u_c|. \tag{44}$$

Besides, although the robots may adjust the formation shape due to some practical constraints, this process is only transient and will not affect the existence of dominantly feasible replacement for our analysis.

Notice that (35) reveals the conditions for indirect controllability from a global topology perspective. For an external attacker, obtaining the global topology in some situations can be hard due to the capability constraints (e.g., limited observation range). Despite this drawback, we point out that the attacker is still likely to determine the indirect controllability of a robot merely by local knowledge about the MRN. The following theorem demonstrates how the dominantly feasible replacement can be countered.

Theorem 3. *Considering the CNP pattern for r_v is absent initially, there exists no dominantly feasible replacement against r_v if $\forall i \in \mathcal{N}_v^{out}$, such that*

$$a_{iv}|u_e| > \bar{a}_i|u_c|, \tag{45}$$

where $\bar{a}_i = \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij}$.

Proof. The key point of this theorem is to analyze the influence of the in-neighbors of r_i and prove (45) is sufficient for the indirect controllability.

First, we will show that when r_v is under attack, the state of its out-neighbor, $\tilde{z}_i(t)$, can be bounded by other two state evolution processes, whose velocities are the same as r_v. During the attack process, the state of r_i is updated by (10) and r_a satisfies $\mathbf{z}_a(t) \in \mathcal{P}(\mathbf{z}_v(t), R_o)$. Consider the extreme cases $z_v^a(t) = z_v(t)$ and $z_v^b(t) = z_v(t) - R_o$, and the resulting state evolution is denoted by $\dot{\tilde{z}}^a(t)$ and $\dot{\tilde{z}}^b(t)$, respectively. Resembling the velocity evolution of the MRN under attack described by (39), $\dot{\tilde{z}}^a(t)$ and $\dot{\tilde{z}}^b(t)$ can be written as

$$\dot{\tilde{z}}^a(t) = \sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u} + \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal (\tilde{u} - \tilde{\lambda}_\ell \tilde{z}(0) + \tilde{\lambda}_\ell h)}{e^{\tilde{\lambda}_\ell t}}, \tag{46}$$

$$\dot{\tilde{z}}^b(t) = \sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u} + \sum_{\ell=3}^{n} \frac{\tilde{q}_\ell \tilde{p}_\ell^\intercal (\tilde{u} - \tilde{\lambda}_\ell \tilde{z}(0) + \tilde{\lambda}_\ell h')}{e^{\tilde{\lambda}_\ell t}}, \tag{47}$$

where $h' = h - R_0 \boldsymbol{e}_v$, and $\boldsymbol{e}_v \in \mathbb{R}^n$ is the unit vector with the entry for r_v being one. It is clear from (46) and (47) that the offset in h' will not affect the velocity term $\sum_{\ell=1}^{2} \tilde{q}_\ell \tilde{p}_\ell^\intercal \tilde{u}$ when $t \to \infty$, and

$$\tilde{z}_i(t) \in \left[\min\{\tilde{z}_i^a(t), \tilde{z}_i^b(t)\}, \max\{\tilde{z}_i^a(t), \tilde{z}_i^b(t)\}\right]. \tag{48}$$

Therefore, using an arbitrary state that locates in $[z_v(t) - R_o, z_v(t)]$ to update r_i's state will make no difference when analyzing the indirect controllability. In the sequel, we directly consider that r_i uses $z_v(t)$ to update $\tilde{z}_i(t)$.

Next, we turn to prove the indirect controllability of r_v on r_i. For legibility, we begin with the simple case where only one directed path exists from r_v to r_i. In this situation, other in-neighbors of r_i will not receive information from r_v, i.e., $\forall j \in \mathcal{N}_i^{in} \setminus \{v\}$, $a_{jv} = 0$. Based on the global state evolution (25), define the exponentially decaying term under attack as

$$f(t) = \sum_{i=2}^{n} e^{-\tilde{\lambda}_i t} \tilde{q}_i \tilde{p}_i^\intercal (\tilde{z}(0) - h) - \sum_{i=2}^{n} \frac{e^{-\tilde{\lambda}_i t}}{\tilde{\lambda}_i} \tilde{q}_i \tilde{p}_i^\intercal \tilde{u}. \tag{49}$$

Then, one can rewrite the dynamics of r_i under attack as

$$\begin{aligned}
\dot{\tilde{z}}_i =& a_{iv}(z_v - \tilde{z}_i - h_v + h_i) + \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij}(\tilde{z}_j - \tilde{z}_i - h_j + h_i) \\
=& a_{iv}(u_e t + \tilde{z}_v^0 - \tilde{z}_i - h_{iv}) \\
&+ \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij}(u_c t + \tilde{z}_j^0 + f_j(t) - \tilde{z}_i - h_{ij}),
\end{aligned} \tag{50}$$

where $\tilde{z}_v^0$ and $\tilde{z}_j^0$ are the constant offsets of r_v and r_j, which can be computed similar to the constant vector in (25). Let $b_{iv} = a_{iv}(\tilde{z}_v^0 - h_{iv})$, $\bar{b}_i = \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij}(\tilde{z}_j^0 - h_{ij})$, $\bar{a}_i = \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij}$, and $\bar{f}_i = \sum_{j \in \mathcal{N}_i^{in} \setminus \{v\}} a_{ij} f_j$. Then, (50) is further rewritten as

$$\begin{aligned}
\dot{\tilde{z}}_i &= a_{iv} u_e t - a_{iv} \tilde{z}_i + b_{iv} + \bar{a}_i u_c t - \bar{a}_i \tilde{z}_i + \bar{b}_i + \bar{f}_i \\
&= (a_{iv} u_e + \bar{a}_i u_c) t - (a_{iv} + \bar{a}_i) \tilde{z}_i + (b_{iv} + \bar{b}_i) + \bar{f}_i \\
&= b_1 t - b_2 \tilde{z}_i + b_3 + \bar{f}_i,
\end{aligned} \tag{51}$$

where $b_1 = a_{iv} u_e + \bar{a}_i u_c$, $b_2 = a_{iv} + \bar{a}_i$ and $b_3 = b_{iv} + \bar{b}_i$. Note that (51) is a first-order constant coefficient non-homogeneous linear equation. Leveraging the constant variation method and superposition principle, the solution of (51) is given by

$$\tilde{z}_i(t) = \frac{b_1}{b_2} t + \left(\frac{b_1}{b_2^2} - \frac{b_3}{b_2}\right)(e^{-b_2 t} - 1) + \int_0^t \bar{f}_i(t) \mathrm{d}t. \tag{52}$$

Then, the explicit expression of r_i's velocity is represented as

$$\dot{\tilde{z}}_i(t) = \frac{b_1}{b_2} + \left(b_3 - \frac{b_1}{b_2}\right) e^{-b_2 t} + \bar{f}_i. \tag{53}$$

Note that the last two terms in (53) all decay exponentially with t and will not determine the final motion of r_i. Therefore, we directly focus on the asymptotic velocity

$$\dot{\tilde{z}}_i^{\infty} = \frac{b_1}{b_2} = a_{iv}u_e + \bar{a}_i u_c, \tag{54}$$

where $b_2 = 1$ is used. The next step is the same as the proof of Theorem 2. It follows that whether r_i is indirectly controllable by r_v is determined by $(z_c^* - z_i^0)u_c$ and $a_{iv}u_e + \bar{a}_i u_c$. Accordingly, one infers that r_i is indirectly controllable by r_v when (45) is satisfied.

Finally, consider the general case where other directed paths exist from r_v to r_i apart from the one r_v directly reaching r_i. In this situation, $\exists j \in \mathcal{N}_i^{in}\backslash\{v\}$, $a_{jv} > 0$, i.e., r_i will also be indirectly influenced by r_v. Then, one has

$$\dot{\tilde{z}}_i = a_{iv}(z_v - \tilde{z}_i - h_{iv}) + \sum_{j\in\mathcal{N}_i^{in}\backslash\{v\}} a_{ij}(\tilde{z}_j(z_v) - \tilde{z}_i - h_{ij}), \tag{55}$$

where $\tilde{z}_j(z_v)$ indicates that $\tilde{z}_j$ is influenced by z_v. Given the same local interaction weights $\{a_{ij}, j \in \mathcal{N}_i^{in}\}$, (55) indicates that $\tilde{z}_i(t)$ is much more influenced by r_a, compared with the case (50) where $a_{jv} = 0$. Recalling the explicit asymptotic form of $\dot{\tilde{z}}_i$ given by (41) (i.e., $\dot{\tilde{z}}_i^{\infty} = \tilde{q}_1^{[i]}u_e + \tilde{q}_2^{[i]}u_c$), one directly derives from this expression and (54) that

$$\tilde{q}_1^{[i]} > a_{iv} \Rightarrow (\frac{1}{\tilde{q}_1^{[i]}} - 1) < (\frac{1}{a_{iv}} - 1) \Rightarrow \frac{\tilde{q}_2^{[i]}}{\tilde{q}_1^{[i]}} < \frac{\bar{a}_i}{a_{iv}}. \tag{56}$$

Therefore, if $\frac{|u_e|}{|u_c|} > \frac{\bar{a}_i}{a_{iv}}$, then $\frac{|u_e|}{|u_c|} > \frac{\tilde{q}_2^{[i]}}{\tilde{q}_1^{[i]}}$ also holds, which is sufficient to satisfy the global condition in Theorem 2. The proof is completed. □

Theorem 3 is a sufficient condition for the indirect controllability of r_v. Meanwhile, it characterizes the severe vulnerability of r_v under the replacement attack, even if r_a does not have the global but only local topology knowledge about the MRN. Hence, the topology design is critical in determining the security performance under the replacement attack.

D. Topology Parameter Design

In this part, we present the initiatory topology design that enhances the performance of the CNP-based mechanism. First, we characterize the replacement risk of a robot.

Corollary 1. *Given the victim r_v and the maximum $u_e^{\max}$, there exist no dominantly feasible replacement under Algorithm 1, if the eigenvectors of $\tilde{L}$ satisfy*

$$\tilde{q}_1^{[i]}|u_e^{\max}| > \tilde{q}_2^{[i]}|u_c|, \ \forall i \in \mathcal{N}_v^{out}. \tag{57}$$

Proof. This result can be easily proved by utilizing Theorem 2 to ensure each robot in $\mathcal{N}_v^{out}$ is indirectly controllable by r_a. Due to the space limit, the details are omitted here. □

Theoretically, Corollary 1 points out the direction to design the global topology parameters to disable dominantly feasible replacement. It is worth noting that the non-existence of the dominantly feasible replacement is based on the negative proposition of Definition 3, and the condition in Theorem 3 is only sufficient to meet Corollary 1.

Next, we show how the topology parameters can be designed. Since the formation shape of the MRN is initially preset when it is deployed, the interaction relationships between robots are not arbitrary. Therefore, we directly suppose that the edge set $\mathcal{E}$ and the interaction weight bounds $0 < a_{lb} \le a_{up} \le 1$ are given by the user. Then, for a selected r_v, the topology design that defends r_v and its out-neighbors from the replacement attack is to make the condition (57) easy to meet by minimizing $\tilde{q}_2^{[i]}/\tilde{q}_1^{[i]}$, formulated as

$$\min_{A} \quad \sum_{i\in\mathcal{N}_v^{out}} \frac{\tilde{q}_2^{[i]}(v)}{\tilde{q}_1^{[i]}(v)} \tag{58a}$$

$$\text{s.t.} \quad a_{lb} \le a_{ij} \le a_{up}, \ \text{if } (i,j) \in \mathcal{E}, \tag{58b}$$

$$a_{ij} = 0, \ \text{if } (i,j) \notin \mathcal{E}, \tag{58c}$$

$$\sum\nolimits_{j=1}^{n} a_{ij} = 1, \tag{58d}$$

$$\text{(31) and (32).} \tag{58e}$$

Notice that the constraint (58e) represents the construction of new adjacency matrix $\tilde{A}(v)$ and its corresponding Laplacian matrix $\tilde{L}(v)$ when the victim r_v is determined. Specifically, $\tilde{L}(v)$ is directly used to calculate the terms $\tilde{q}_1^{[i]}(v)$ and $\tilde{q}_2^{[i]}(v)$ in the objective function (58a). If we do not merely put the security consideration on a single robot r_v but the whole MRN, then the objective function in (58) can be further revised as

$$\min_{A} \quad \sum_{v=1}^{n} \sum_{i\in\mathcal{N}_v^{out}} \frac{\tilde{q}_2^{[i]}(v)}{\tilde{q}_1^{[i]}(v)}, \tag{59}$$

along with the same constraints in (58).

It should be noted that the optimal solution of (58) or (59) is hard to obtain because the objective function is built on the eigenvectors of $\tilde{L}(v)$, which is highly nonlinear about A. To overcome this issue, we are inspired by Theorem 3 to provide a conservative but more computation-efficient version of (58), given by

$$\min_{A} \quad \sum_{i\in\mathcal{N}_v^{out}} \frac{\bar{a}_i}{a_{iv}} = \sum_{i\in\mathcal{N}_v^{out}} \frac{1}{a_{iv}} - |\mathcal{N}_v^{out}| \tag{60a}$$

$$\text{s.t.} \quad \text{(58b)} - \text{(58e)}. \tag{60b}$$

It is clear from (60) that when the CNP pattern for a selected r_v (e.g., the robot with maximum out-neighbor number $|\mathcal{N}_v^{out}|$) is absent initially, the larger each weight a_{iv} is (not necessarily $\sum_{i\in\mathcal{N}_v^{out}} a_{iv}$), the more immune r_v is to its replacement. Note that if the victim selection criteria of r_a is taken into consideration in the topology parameter design, then it will introduce a tradeoff between the attack risk and the topology design. For example, supposing the robot with maximum out-degree summation is always selected as the victim, the original problem (60) needs to include a new constraint

$$v = \arg\max_{\tilde{v}\in\mathcal{V}} \sum_{i\in\mathcal{N}_{\tilde{v}}^{out}} a_{i\tilde{v}}. \tag{61}$$

In this case, solving the revised problem would be more difficult because r_v needs to be determined. Similar to (59), if

TABLE II
ATTACK CAPABILITY LIMITATION UNDER DIFFERENT SCENARIOS

Capability to Realize Attack		**With Knowledge of the Topology**		**Without Knowledge of the Topology**
		Global	Local	
Whether the CNP $\mathcal{P}_c(\mathcal{C}_v^p)$ exists	Yes	$\tilde{q}_1^{[s]}\|u_e\| \le \tilde{q}_2^{[s]}\|u_c\|$	$a_{sv}\|u_e\| \le \bar{a}_s\|u_c\|$ (weak)	fail in high probability
	No	$\tilde{q}_1^{[i]}\|u_e\| \le \tilde{q}_2^{[i]}\|u_c\|, i \in \mathcal{N}_v^{out}$ and make $\mathcal{P}_c(\mathcal{C}_v^p)$ form	$a_{iv}\|u_e\| \le \bar{a}_i\|u_c\|, i \in \mathcal{N}_v^{out}$ (weak) and make $\mathcal{P}_c(\mathcal{C}_v^p)$ form	fail in high probability

we consider the replacement risk for all robots, the objective function in (60) is revised as

$$\min_{A} \quad \sum_{v=1}^{n} \sum_{i \in \mathcal{N}_v^{out}} \frac{1}{a_{iv}} = \sum_{v=1}^{n} \sum_{j \in \mathcal{N}_v^{in}} \frac{1}{a_{vj}} \tag{62}$$

which can be decomposed into minimizing $\sum_{j \in \mathcal{N}_v^{in}} \frac{1}{a_{vj}}$ for each $v \in \{1, \cdots, n\}$ independently. Notice that the conservativeness of (60) or (62) lies in that we directly optimize the local topology weights instead of the global one, and this operation can be regarded as minimizing the upper bound of (58a) or (59). Despite the conservativeness, designing the topology parameters by solving (60) or (62) is feasible and computation-tractable, and the optimality is guaranteed under the constraints (58b)-(58e).

Finally, we observe that for the situation where the CNP pattern for r_v exists (either it exists initially or is formed later), the proximity rule (11a) can be easily satisfied, and realizing the replacement attack mainly lies in meeting the communication conditions (11b)-(11c). In this regard, similar to r_v's indirect controllability on its out-neighbors, the indirect controllability of r_v on r_s is the key to the defense, where the smaller $\tilde{q}_2^{[s]}/\tilde{q}_1^{[s]}$ is, the harder (11b)-(11c) can be met.

E. Extensions and Discussions

In this part, we first consider the situation where no central robot is available and extend the CNP-based security mechanism to distributed cases. Then, a brief summary and application discussions are presented.

By calling distributed here, we mainly indicate that the interactive communication between robots is distributed, and a robot can only have information from its neighbors. In this situation, the requirement that $\forall i \in \mathcal{N}_v^{out}$ in the original rules (11) can be relaxed, as all the robots in the MRN independently and directly communicate with each other. Therefore, given a time slot t_l and $i \in \mathcal{N}_v^{out}$ initially, r_i will take r_a as the real r_v at moment t_0 if and only if the following three conditions are satisfied.

- **Distributed CNP-based security rules**

$$d_{i,a}(t) < d_{i,v}(t), \ t \ge t_0 - t_l, \tag{63a}$$
$$\mathbf{z}_a(t) \in \mathcal{P}(\tilde{\mathbf{z}}_i(t), R_c), \ t \ge t_0 - t_l, \tag{63b}$$
$$\mathbf{z}_v(t) \notin \mathcal{P}(\tilde{\mathbf{z}}_i(t), R_c), \ t \ge t_0. \tag{63c}$$

The above rules can be regarded a special one-on-one case of the rules in (11), as r_i is unaware of other members in $\mathcal{N}_v^{out}$. Since there is no central robot in the formation, the out-neighbors $\mathcal{N}_v^{out}$ only need to independently confirm the rules in (63). It is worth noting that in distributed cases, achieving the replacement attack is easier than that in centralized cases. This is because r_a can be recognized as the true r_v by r_v's out-neighbors in sequence (as indicated by (63)), which is more flexible to implement for r_a than being recognized by r_v's out-neighbors simultaneously. From the defense perspective, there is no need to solve the problem (60) for topology design, and one can change the objective function as a min-max version, given by

$$\min_{A} \max_{i \in \mathcal{N}_v^{out}} \quad \frac{1}{a_{iv}}. \tag{64}$$

In summary, the distributed CNP-based security mechanism enjoys lower computation burdens, but is more vulnerable than the centralized one.

Next, we summarize the attack capability limit under the CNP-based mechanism in Table II. It is clear that whether the replacement attack against an MRN can be realized is largely determined by the prior system knowledge mastered by r_a, which corresponds to the common intuition. First, without the topology knowledge of the MRN, r_a can hardly achieve the attack as the strategy cannot be designed accordingly. Second, if a CNP pattern $\mathcal{P}_c(\mathcal{C}_v^p)$ is unavailable, r_a needs to find extra attack strategies to make it present, which essentially requires much higher attack costs. Once the bound of the obstacle-avoidance mechanism $|g(\cdot)|$ and u_c are fixed, the attack feasibility is directly constrained by the indirect controllability. Specifically, for the local topology case, the condition $a_{iv}|u_e| \le \bar{a}_i|u_c|$ needs to hold for all $i \in \mathcal{N}_v^{out}$, which is utilized to increase the attack costs of r_a. Here we describe the condition as weak because it is not sufficient to meet the global condition, which can be explicitly demonstrated by

$$\frac{\bar{a}_i}{a_{iv}} \ge \frac{\tilde{q}_2^{[i]}}{\tilde{q}_1^{[i]}} \text{ and } \frac{\bar{a}_i}{a_{iv}} \ge \frac{|u_e^{\max}|}{|u_c|} \not\Rightarrow \frac{\tilde{q}_2^{[i]}}{\tilde{q}_1^{[i]}} \ge \frac{|u_e^{\max}|}{|u_c|}. \tag{65}$$

Therefore, we observe that the best way to actively strengthen the system security is to jointly design the formation shape configuration, the communication topology, and the reaction input magnitude, making it extremely hard for the attacker to successfully achieve the attack.

Another issue concerning the communication is that a non-victim robot may lose connection with the other robots during the attack process. Since this robot is not a spoofing target for the attacker (i.e., there is no communication package with its faked ID in cyberspace), it can adopt some predicting-and-tracking control strategies that leverage the historical data of its neighbors to reconnect the formation [18]. This point is not the focus of this paper and is omitted here.

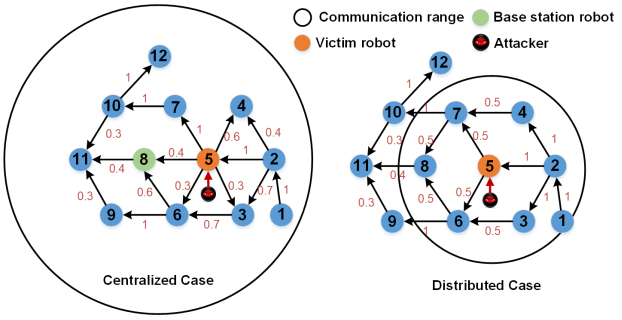


Fig. 4. An MRN of 12 robots with the specified formation shapes and topologies. The red arrow is the attack direction.

Finally, it is worth mentioning that the proposed defense method mainly builds on two fundamental factors in formation control: the topology structure and the geometric shape, and has little dependence on the specific robot control models. In this sense, the method can be well applied to real-world scenarios. Although the velocity of the formation leader is simplified to be constant, it can be time-varying. This relaxation will mainly affect obtaining the explicit critical conditions for defense as in (45), and will not change the topology parameter design. Specifically, if the time-varying leader velocity can reach a stable state or is strictly bounded, the critical condition analysis can still be applied by using the maximum velocity magnitude and work in a conservative sense.

IV. Simulation Results

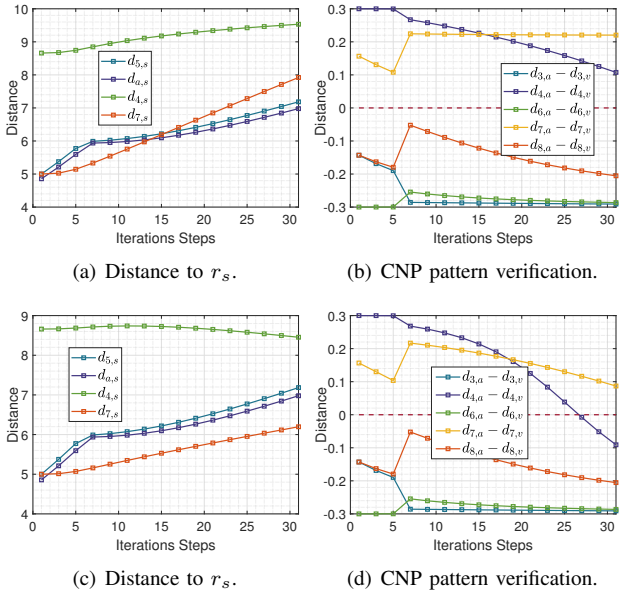
In this section, we provide representative simulation examples to demonstrate the performance of the proposed CNP security mechanism against the replacement attack.

We use an MRN of 12 robots for performance evaluation, considering both the centralized and distributed communication cases. Two representative scenarios with different topology settings and communication ranges are provided, as shown in Fig. 4. In the centralized case, r_8 is the central robot r_s with communication range $R_{c,1} = 14.5$m, while r_5 is selected as the victim robot. Note that here the CNP pattern for r_5 does not exist initially. In the distributed case, each robot communicates with its neighbors independently with communication range $R_{c,2} = 7$m and r_5 being the victim robot. Note that the CNP pattern for r_5 is presented in this case. The velocity setting is $u_c = 0.1$ m/s and $u_e^{\max} = 1$ m/s. As for the strategy design of the attack (9), here we set the attack cost function as

$$F(u_a(t), z_v(t), R_c) = -\sum_{j \in \mathcal{N}_v^{out}} (\|\hat{\boldsymbol{z}}_{\boldsymbol{j}}(t) - \mathbf{z}_v(u_a(t))\|_2 - R_c)^2,$$

which aims to maximize the sum of the distance of r_v and $\mathcal{N}_v^{out}$, such that their connection breaks. At each step, the attack input is obtained by the heuristic search method.

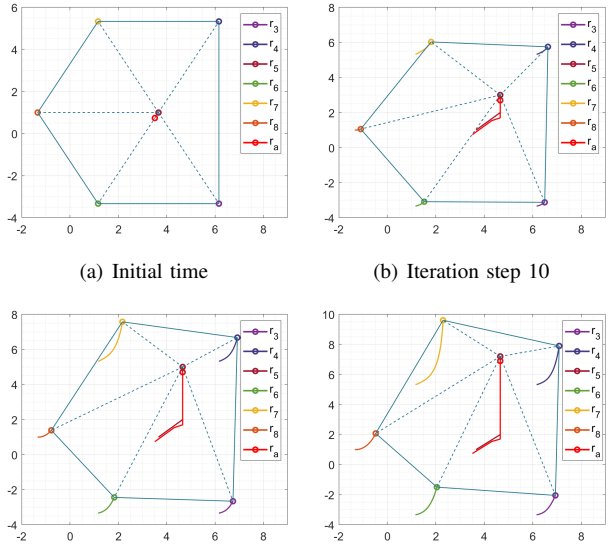
First, we present the results for the centralized communication case. For comparisons, we denote the topology of the centralized case in Fig. 4 as topology setting 1, and additionally use a topology setting 2 which inherits the former one with slight modifications ($a_{42} = 0.8, a_{45} = 0.2, a_{75} = 0.8, a_{78} = 0.2$ in this setting). Fig. 5(a)-5(b) show the example under topology setting 1, where Fig. 5(a) plots the distances between



(a) Distance to r_s. (b) CNP pattern verification.

(c) Distance to r_s. (d) CNP pattern verification.

Fig. 5. Examples of the CNP-based security mechanism in centralized communication case. (a)(b) the replacement attack fails in the topology setting 1. (c)(d): the replacement attack succeeds in the topology setting 2.

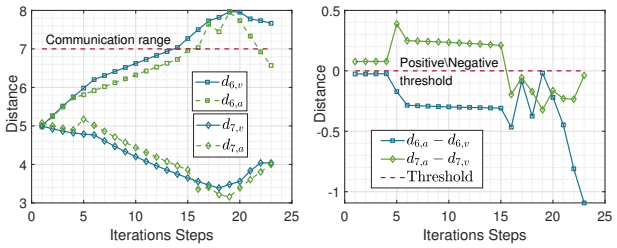


(a) Initial time (b) Iteration step 10

(c) Iteration step 20 (d) Iteration step 30

Fig. 6. Snapshots of the evolution for r_5 and $\mathcal{N}_5^{out}$ in topology setting 1.

$\{r_a, r_v, r_4, r_7\}$ and r_8 to verify the communication condition (11b)-(11c), while Fig. 5(b) plots the distance difference of $\mathcal{N}_5^{out}$-r_a and $\mathcal{N}_5^{out}$-r_5 to verify the CNP condition (11a). Regardless of the communication condition, it is clear to see that $d_{7,a} - d_{7,v}$ will not decrease as the attack continues, and the curves associated with $\mathcal{N}_5^{out}$-r_5 cannot reach below zero simultaneously, indicating the CNP pattern for r_5 cannot be formed and the replacement attack fails. This is because r_5 has total indirect controllability on r_7, which moves along with r_5 to prevent r_5 from becoming a vertex of a CNP pattern. Fig. 5(c) and Fig. 5(d) show the example under topology setting 2, and have the same meanings as those of Fig. 5(a) and Fig. 5(b). By contrast, the CNP condition in this example is easy to meet as the attack continues and the replacement attack will succeed. Specifically, r_5's indirect controllability on r_7 is



(a) Distance to $\mathcal{N}_5^{out}$. (b) CNP pattern verification.

Fig. 7. Example of the CNP-based security mechanism under the replacement attack in distributed communication case.

larger than that on r_4 (i.e., $\tilde{q}_1^{[7]}/\tilde{q}_2^{[7]} > \tilde{q}_1^{[4]}/\tilde{q}_2^{[4]}$), $d_{7,a}-d_{7,v} < 0$ is slower to meet than $d_{4,a} - d_{4,v} < 0$, corresponding to the topology parameter design idea that larger $\tilde{q}_1^{[i]}/\tilde{q}_2^{[i]}$ $(i \in \mathcal{V}_v^p)$ is better for defense. Furthermore, to intuitively show the attack process, the snapshots of the evolution process of r_5 and $\mathcal{N}_5^{out}$ in topology setting 1 are given in Fig. 6, which also verifies that the CNP pattern for r_5 cannot be formed. The process plots for topology setting 2 are likewise and omitted here.

Next, we move on to the distributed communication case shown in Fig. 7, where r_5 has two out-neighbors and the CNP pattern for r_5 exists initially. Fig. 7(a) describes the distances of $\mathcal{N}_5^{out}$-r_a and $\mathcal{N}_5^{out}$-r_5 , where d_6 increases while $d_{v,7}$ decreases during the attack process. Note that this effect arises from that r_a tries to meet the communication conditions (63b)-(63c) for r_6 first. After this is achieved, r_a continues to meet the above conditions for r_7. Fig. 7(b) illustrates that the proximity condition (63a) for the two neighbors can be met sequentially, indicating the replacement attack can be easily achieved. This result corresponds to the previous conclusion that the MRN is more vulnerable to the replacement attack in distributed cases than in centralized cases.

V. Conclusion

In this paper, we investigated the problem of securing the formation control of MRNs against the replacement attack, where an external attack robot aimed to replace a victim robot in the MRN. First, we proposed the notion of CNP to exploit the intrinsic physical characteristic of a formation shape, and designed a CNP-based security mechanism to counter the attack. Then, we introduced the indirect controllability between two robots to explicitly characterize the feasibility conditions from the attack perspective. We demonstrated that the topology structure and formation shape of the MRN largely constrained the attack capability. Finally, we provided the initiatory topology design to enhance the defense performance, along with extensions to the distributed communication cases and application discussions. Simulations illustrated the effectiveness of the proposed method. Future directions include i) extending the method to more general system models, and ii) investigating the detection and defense designs for situations when the attack is successfully realized.

References

[1] K.-K. Oh, M.-C. Park, and H.-S. Ahn, "A survey of multi-agent formation control," *Automatica*, vol. 53, pp. 424–440, 2015.

[2] F. Bullo, J. Cortes, and S. Martinez, *Distributed control of robotic networks: A mathematical approach to motion coordination algorithms*. Princeton University Press, 2009, vol. 27.

[3] H. Choi, W.-C. Lee, Y. Aafer, F. Fei, Z. Tu, X. Zhang, D. Xu, and X. Xinyan, "Detecting attacks against robotic vehicles: A control invariant approach," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 801–816.

[4] Y. Yang, Y. Xiao, and T. Li, "Attacks on formation control for multiagent systems," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 12 805–12 817, 2021.

[5] L. Zhou and P. Tokekar, "Multi-robot coordination and planning in uncertain and adversarial environments," *Current Robotics Reports*, vol. 2, pp. 147–157, 2021.

[6] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[7] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2015.

[8] Z. Feng and G. Hu, "Secure cooperative event-triggered control of linear multiagent systems under DoS attacks," *IEEE Transactions on Control Systems Technology*, vol. 28, no. 3, pp. 741–752, 2020.

[9] W. Liu, J. Sun, G. Wang, F. Bullo, and J. Chen, "Data-driven resilient predictive control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 68, no. 8, pp. 4722–4737, 2023.

[10] S. Gil, M. Yemini, A. Chorti, A. Nedić, H. V. Poor, and A. J. Goldsmith, "How physicality enables trust: A new era of trust-centered cyberphysical systems," *arXiv:2311.07492*, 2023.

[11] A. Mitra, J. A. Richards, S. Bagchi, and S. Sundaram, "Resilient distributed state estimation with mobile agents: Overcoming Byzantine adversaries, communication losses, and intermittent measurements," *Autonomous Robots*, vol. 43, no. 3, pp. 743–768, 2019.

[12] A. Prorok, M. Malencia, L. Carlone, G. S. Sukhatme, B. M. Sadler, and V. Kumar, "Beyond robustness: A taxonomy of approaches towards resilient multi-robot systems," *arXiv:2109.12343*, 2021.

[13] M. A. Kamel, X. Yu, and Y. Zhang, "Formation control and coordination of multiple unmanned ground vehicles in normal and faulty situations: A review," *Annual Reviews in Control*, vol. 49, pp. 128–144, 2020.

[14] W. He, W. Xu, X. Ge, Q.-L. Han, W. Du, and F. Qian, "Secure control of multiagent systems against malicious attacks: A brief survey," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3595–3608, 2022.

[15] K. Saulnier, D. Saldana, A. Prorok, G. J. Pappas, and V. Kumar, "Resilient flocking for mobile robot teams," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1039–1046, 2017.

[16] M. Santilli, M. Franceschelli, and A. Gasparri, "Secure rendezvous and static containment in multi-agent systems with adversarial intruders," *Automatica*, vol. 143, p. 110456, 2022.

[17] Z. Feng, C. Sun, and G. Hu, "Robust connectivity preserving rendezvous of multirobot systems under unknown dynamics and disturbances," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 4, pp. 725–735, 2017.

[18] Y. Li, H. Wang, J. He, and X. Guan, "Optimal topology recovery scheme for multi-robot formation control," in *IEEE 28th International Symposium on Industrial Electronics*, 2019, pp. 1847–1852.

[19] K. S. Engin and V. Isler, "Establishing fault-tolerant connectivity of mobile robot networks," *IEEE Transactions on Control of Network Systems*, vol. 8, no. 2, pp. 667–677, 2021.

[20] H. Park and S. Hutchinson, "Robust rendezvous for multi-robot system with random node failures: An optimization approach," *Autonomous Robots*, vol. 42, no. 8, pp. 1807–1818, 2018.

[21] R. Wehbe and R. K. Williams, "Probabilistic security for multirobot systems," *IEEE Transactions on Robotics*, vol. 37, no. 1, pp. 146–165, 2021.

[22] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim, "Rocking drones with intentional sound noise on gyroscopic sensors," in *24th USENIX Security Symposium*, 2015, pp. 881–896.

[23] G. Bianchin, Y.-C. Liu, and F. Pasqualetti, "Secure navigation of robots in adversarial environments," *IEEE Control Systems Letters*, vol. 4, no. 1, pp. 1–6, 2019.

[24] Z. Ju, H. Zhang, and Y. Tan, "Deception attack detection and estimation for a local vehicle in vehicle platooning based on a modified UFIR estimator," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3693–3705, 2020.

[25] Y. Li, J. He, L. Cai, and X. Guan, "Local topology inference of mobile robotic networks under formation control," *IEEE Transactions on Automatic Control*, vol. 68, no. 11, pp. 6450–6465, 2023.

[26] Y. Li, J. He, C. Chen, and X. Guan, "Intelligent physical attack against mobile robots with obstacle-avoidance," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 253–272, 2023.

[27] 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, no. 4, pp. 766–781, 2013.

[28] H. Rezaee, T. Parisini, and M. M. Polycarpou, "Resiliency in dynamic leader-follower multiagent systems," *Automatica*, vol. 125, p. 109384, 2021.

[29] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.

[30] H. Zhang and S. Sundaram, "Robustness of information diffusion algorithms to locally bounded adversaries," in *American Control Conference*, 2012, pp. 5855–5861.

[31] D. Zhao, Y. Lv, X. Yu, G. Wen, and G. Chen, "Resilient consensus of higher order multiagent networks: An attack isolation-based approach," *IEEE Transactions on Automatic Control*, vol. 67, no. 2, pp. 1001–1007, 2022.

[32] J. Usevitch and D. Panagou, "Resilient trajectory propagation in multirobot networks," *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 42–56, 2022.

[33] Y. Cao and W. Ren, "Distributed coordinated tracking with reduced interaction via a variable structure approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 33–48, 2011.

[34] I. Jawhar, N. Mohamed, J. Wu, and J. Al-Jaroodi, "Networking of multi-robot systems: Architectures and requirements," *Journal of Sensor and Actuator Networks*, vol. 7, no. 4, p. 52, 2018.

[35] M. Mohanan and A. Salgoankar, "A survey of robotic motion planning in dynamic environments," *Robotics and Autonomous Systems*, vol. 100, pp. 171–185, 2018.

[36] M. Lalou, M. A. Tahraoui, and H. Kheddouci, "The critical node detection problem in networks: A survey," *Computer Science Review*, vol. 28, pp. 92–117, 2018.

[37] M. Agarwal, S. Biswas, and S. Nandi, "Advanced stealth man-in-the-middle attack in WPA2 encrypted Wi-Fi networks," *IEEE Communications Letters*, vol. 19, no. 4, pp. 581–584, 2015.

[38] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man-in-the-middle attacks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2027–2051, 2016.

[39] H. Pirayesh and H. Zeng, "Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 767–809, 2022.

[40] F. L. Lewis, H. Zhang, K. Hengster-Movric, and A. Das, *Cooperative control of multi-agent systems: Optimal and adaptive design approaches*. Springer Science & Business Media, 2013.

[41] W. Ren and R. W. Beard, *Distributed consensus in multi-vehicle cooperative control*. Springer, 2008.

[42] C.-T. Lin, "Structural controllability," *IEEE Transactions on Automatic Control*, vol. 19, no. 3, pp. 201–208, 1974.

[43] A. Rahmani, M. Ji, M. Mesbahi, and M. Egerstedt, "Controllability of multi-agent systems from a graph-theoretic perspective," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 162–186, 2009.

[44] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, no. 7346, pp. 167–173, 2011.

[45] S. Monaco and L. R. Celsi, "On multi-consensus and almost equitable graph partitions," *Automatica*, vol. 103, pp. 53–61, 2019.

[46] F. Bullo, *Lectures on Network Systems*. Kindle Direct Publishing, Edition 1.6, 2022.

Yushan Li (Member, IEEE) received the B.E. degree in automatic control from Huazhong University of Science and Technology, Wuhan, China, in 2018, and the Ph.D. degree in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2024.

He is currently a Postdoctoral Researcher with the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include inference and secure control of network systems and multi-robot systems.

Jianping He (Senior Member, IEEE) received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2013. From 2013 to 2017, he was a Research Fellow with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada.

He is currently an Associate Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His research interests mainly include the distributed learning, control and optimization, security, and privacy in network systems.

Cailian Chen (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in automatic control from Yanshan University, China, in 2000 and 2002, respectively, and the Ph.D. degree in control and systems from the City University of Hong Kong, Hong Kong, SAR, in 2006.

She has been with the Department of Automation, Shanghai Jiao Tong University, since 2008, where she is currently a Distinguished Professor. Before that, she was a postdoctoral research associate in University of Manchester, U.K. (2006-2008). Her research interests include industrial wireless networks and computational intelligence and the Internet of Vehicles.

Xinping Guan (Fellow, IEEE) received the B.S. degree in mathematics from Harbin Normal University, Harbin, China, in 1986, and the Ph.D. degree in control science and engineering from the Harbin Institute of Technology, Harbin, in 1999.

He is currently a Chair Professor with Shanghai Jiao Tong University, Shanghai, China, where he is the Dean of the School of Electronic Information and Electrical Engineering, and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. Before that, he was the Professor and Dean of Electrical Engineering, Yanshan University, Qinhuangdao, China. His current research interests include industrial cyber-physical systems, wireless networking and applications in smart factories, and underwater networks.

Lin Cai (Fellow, IEEE) received her M.A.Sc. and Ph. D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2002 and 2005, respectively. Since 2005, she has been with the Department of Electrical & Computer Engineering at the University of Victoria, and she is currently a Professor.

Dr. Cai is an NSERC E.W.R. Steacie Memorial Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and an IEEE Fellow. She has been elected to serve the board of the IEEE Vehicular Technology Society, 2019-2024, and as its VP in Mobile Radio. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic and the Internet of Things.