MULTI-ROBOT SYSTEMS IN ADVERSARIAL SETTINGS: ADVERSARY DETECTION, RESILIENT COORDINATION AND COOPERATION

by

Mohammad (Rayan) Bahrami

A DISSERTATION

Submitted to the Faculty of the Stevens Institute of Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Mohammad (Rayan) Bahrami, Candidate

ADVISORY COMMITTEE

Prof. Hamid Jafarnejad Sani, Chairman Date

Prof. Brendan Englot Date

Prof. Long Wang Date

Prof. Yi Guo Date

STEVENS INSTITUTE OF TECHNOLOGY
Castle Point on Hudson
Hoboken, NJ 07030
2024

MULTI-ROBOT SYSTEMS IN ADVERSARIAL SETTINGS: ADVERSARY DETECTION, RESILIENT COORDINATION AND COOPERATION

ABSTRACT

Networked autonomous mobile robots, such as unmanned aerial and ground vehicles, represent a burgeoning class of cyber-physical systems (CPS) within critical infrastructure sectors. This dissertation addresses the imperative to ensure the safe and secure cooperation of these systems in the face of adversarial challenges. The adversaries are a class of worst-case scenario vulnerabilities in wireless communication networks of multi-robot systems and their perceptual sensing modalities, such as cameras. Such vulnerability can be perceived as the dynamical blind spots for multi-robot systems in the sense that adversarial attacks can be crafted based on the dynamics of the system so as to compromise, severely and shortly, not only the system's operation but also information confidentiality, integrity, and availability, while remaining stealthy (unnoticeable in the monitoring data) until a critical failure.

In the first part of this dissertation, we propose three principled algorithmic frameworks that allow for the detection and mitigation of adversarial attacks on multi-robot coordination. Our results extend the resilient consensus (coordination) of multi-agent (robot) systems to the case of time-varying communication topology with intermittent connections and provide theoretical stability and performance analysis in the continuous-time domain. This is addressed by, first, characterizing control-theoretic and graph-theoretic conditions under which specific classes of adversarial attacks on the communication networks exist, second, by developing the theoretical conditions that determine the degree to which a multi-robot system maintains a certain level of communication-related performance in a cooperative task while enduring

a specific number of adversarial/compromised robots in a given network, and third, by developing decentralized and distributed attack detection frameworks that allow for resilient coordination of the remaining uncompromised robots. We validate our theoretical findings and illustrate their performance through various tests in numerical simulations, high-fidelity simulations, and real-world experiments.

In the second part of this dissertation, we consider multi-robot (quadrotor) coordination with adversarial perception. We demonstrate that a class of adversarial image attacks on the robots' perception modules cause categorically similar effects, including misclassification and mislocalization, which can be formulated as sporadic (intermittent) and spurious data measurements. We propose a framework that allows for state estimation and perception-based relative localization in the presence of intermittent and spurious measurements caused by adversarial image attacks on the perception module. Additionally, we present two open-source vision-enabled multi-robot (quadrotor) platforms, together with developed software packages. We demonstrate the capability of these platforms and the resilience of our framework through experiments on perception-based multi-robot coordination under adversarial image attacks targeting their learned perception modules.

By providing principled algorithms and open-source software, this dissertation contributes to advancing the resilience and security of autonomous multi-robot systems in safety- and time-critical applications, with potential implications for enhancing operational safety across various sectors.

Author: Mohammad (Rayan) Bahrami
Advisor: Prof. Hamid Jafarnejad Sani
Date: August 15th, 2024
Department: Mechanical Engineering
Degree: Doctor of Philosophy

to my loved ones.

Acknowledgments

I would like to extend my appreciation to my academic advisor, Dr. Hamid Jafarnejad Sani, for his support during my time at Stevens. This dissertation would not have been possible without his support.

I would also like to extend my gratitude to Prof. Brendan Englot, Prof. Long Wang, and Prof. Yi Guo for serving on my dissertation committee and for their valuable input and insightful discussions.

I would like to acknowledge and express my appreciation to the funding sources that supported this research, including the Stevens Provost Doctoral Fellowship, Prof. Hamid Jafarnejad Sani's faculty startup fund, the U.S. National Science Foundation (NSF, award no. 2137753), and the Stevens 2023 Fernando Fernandez Ph.D. Robotics and Automation Summer Term Fellowship.

I would also like to thank the members of the Safe Autonomous Systems Lab at Stevens—Bijay Gaudel, Peter Dunphy, Isaac Van Benthuysen, and Patrick Zielinski—for their friendship, assistance with some experiments, and stimulating discussions.

Finally, I would like to extend my deepest gratitude to my loved ones for their unconditional love and unwavering support.

Table of Contents

List of Tables

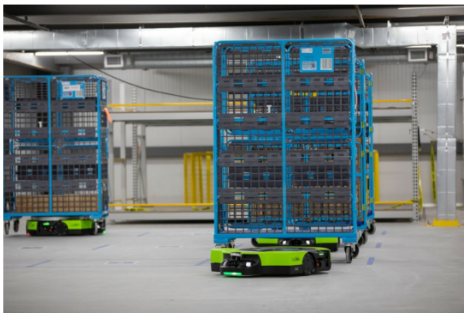
List of Figures

Chapter 1

Introduction

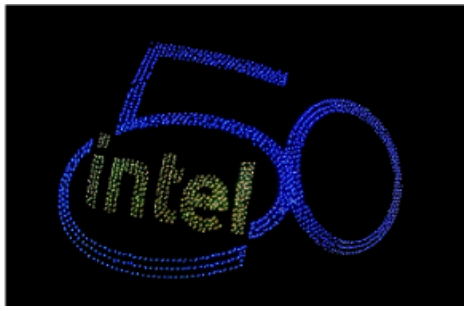
Cyber-physical systems (CPS), such as transportation networks, smart power grids, autonomous vehicles, and mobile sensor networks, are considered indispensable to the critical infrastructure of developed countries (see Fig. 1.1). The cooperation of a network of autonomous mobile robots, such as unmanned aerial or ground vehicles, is of paramount importance since network-enabled cooperation provides distributed coverage, reconfigurability, and mobility in a wide range of applications such as smart transportation, search and rescue missions, surveillance, wildfire monitoring, and delivery. Nonetheless, the deployment of such systems in critical infrastructures has been hampered by various challenges including assurance of safety and security. The safety and security challenges arise, in part, from the vulnerability of wireless communication networks as well as perceptual sensing modalities such as cameras on which networked autonomous mobile robots rely for information exchange, decision-making, and operation. It has been shown that the foregoing vulnerabilities can be exploited in an adversarial manner to design attacks on the transmitted or measured information so as to compromise, severely and shortly, the network-level system stability while remaining stealthy (unnoticeable in the monitoring data) until a critical breakdown.

This dissertation proposes solutions to bring about resilience in multi-robot systems cooperating in adversarial settings where there may exist cyberattacks on information exchanged over wireless communication networks as well as adversarial disruptions in perceptual sensing modalities such as cameras.



(a) Proteus, © 2022, Amazon

(b) Vertical farming, © 2021, AeroFarms & Nokia Bell Labs



(c) Digital firework © 2018, Intel

(d) Semi-autonomous platooning ©2018, Scania

Figure 1.1: Examples of deployment of autonomous and semi-autonomous multi-agent systems in various infrastructures sectors underscores the importance of safe and reliable operation. a) Proteus is Amazon's first fully autonomous warehouse robot. b) Autonomous monitoring of the vertical farming. c) small drones formation for digital fireworks. d) vehicle platooning in intelligent transportation systems.

1.1 Security and Resilience of Cyber-Physical System

The grand challenges of ensuring security and resilience for cyber-physical systems (CPS) have motivated the study and characterization of possible adversarial attacks against these complex systems. The seminal position paper in [18] initially put forth an interpretation of the security of CPS based on the traditional *security* goals, known as the CIA triad, allowing for the study and characterization of cyber threats (adversarial attacks). Such security specifications are defined as follows:

- *Integrity*: the trustworthiness of transmitted data.

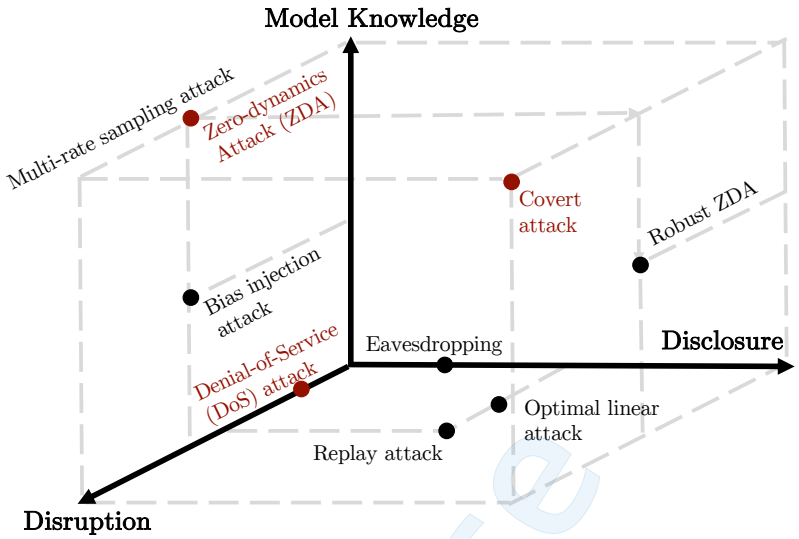


Figure 1.2: Cyber-physical attack space - Reproduced from [131, 96]. This dissertation addresses the adversarial attacks highlighted in red in the context of coordination of multi-agent (robot) systems.

- *Availability*: the availability of resources (e.g., information exchanged over communication networks) for CPS.

- *Confidentiality (privacy)*: the ability to prevent any disclosure of system information to unauthorized entities. Lack of confidentiality is associated with the breach of privacy and eavesdropping attacks.

Having specified the foregoing security goals, *adversarial* attacks on CPS can be classified as *deception* attacks, targeting the integrity[1]/trustworthiness of transmitted data, and *denial-of-service* (DoS) attacks, targeting the data availability upon demand [18]. There exists a vast literature on the study and characterization of CPS's vulnerability to various attacks and defense mechanisms against them [131, 77, 38]. Particularly for networked control systems, as a class of CPS, [131, 96] proposed a general framework to characterize adversarial attacks based on their level of disruption and the adversary's resources and knowledge of the targeted system (see Fig.

[1]The *deception* attacks are also referred to as *integrity* attacks [147].

1.2). Deception attacks such as Zero-dynamics Attacks (ZDA) and covert attacks are known as *stealthy* attacks with the highest level of disruption that require *a priori* full knowledge of the targeted system whereas DoS attacks have a relatively lower level of disruption with no need for *a priori* knowledge of the system. These adversarial attacks have been also studied in the context of multi-agent systems that are tasked with cooperation (e.g., formation, flocking, and swarming [113, 116, 144]) and coordination (e.g., consensus on a quantity of interest [98, 67, 78, 110, 144, 111, 14, 86]).

In terms of *resilience*, which was initially defined as the *survivability* of CPS [17], the property of interest is the system capability of *graceful degradation* in its operational goals (e.g., stability, safety, performance) when under attack. The resilience of multi-robot systems, however, is inherently contingent and poses difficulties in its quantification [102]. Therefore, resilience to specific classes of adversarial attacks is rigorously studied [38]. Examples of specific resilience problems related to the context of this dissertation are the resilience of network (graph) connectivity to node and link failures [146, 71], attack-resilient state estimation [94], resilient consensus [97, 67, 36, 37, 110, 144].

In this dissertation, we consider a class of deception and DoS attacks on multi-robot systems in consensus and consensus-based formation settings (cf. Fig. 1.2).

1.2 Network-enabled Coordination and Cooperation of Mutli-Robot Systems

Network-enabled coordination and cooperation of mobile autonomous robots are central to many safety-critical and time-critical applications such as search and rescue missions, surveillance and monitoring, platooning of autonomous Vehicles, motion and time coordination [108, 79, 32, 103, 55, 144]. These complex tasks often entail

consensus over a quantity of interest. For instance, coordination can be obtained by consensus over certain quantities such as time in time-coordination or inter-vehicle distances in formation control [92, 32, 103, 107, 109]. Also, a team of robots can achieve and maintain formation by exchanging their spatial information (e.g., position and velocity states) and in different settings such as leader-follower, leaderless consensus-based, and virtual structure approaches [91, 106]. Therefore, a reliable communication network for information exchange among robots is an integral part of the secure and safe operation of multi-robot systems.

However, the mobility and limited communication capabilities of mobile autonomous robots (e.g., small Unmanned Aerial Vehicles) give rise to an *ad hoc* and *intermittent* network connectivity of robots in distributed settings, posing practical and theoretical challenges to the safety-critical applications of such systems [55, 144]. The network connectivity challenge has motivated various studies of the connectivity maintenance in multi-robot cooperation subject to ad hoc or time-varying communication [79, 33, 53, 59, 39, 43].

This dissertation focuses on the consensus and consensus-based formation control of a team of robots with time-varying communication topology subject to intermittent connections.

1.3 Detection and Mitigation of Adversarial Attacks

As discussed in Section 1.1, the reliance on wireless communication networks renders the networked systems vulnerable to a large class of adversarial attacks, possibly introduced by a group of malicious (compromised) agents in the network, disrupting normal cooperation.

Detection of adversarial attacks, particularly deception attacks that are intro-

duced to the networked systems through a group of malicious or compromised agents (e.g., mobile robots), is inherently a challenging problem. First, a priori knowledge of the system dynamics can be exploited to design sophisticated deception attacks that are *stealthy* to the common anomaly detectors [131]. Examples of such attacks in decentralized and distributed settings are zero-dynamics attack (ZDA) [97, 78, 131], covert attack [42], and replay attack [110]. Second, the body of effective system-theoretic approaches (e.g., observer/model-based frameworks) developed for distributed detection of faults and stealthy attacks in spatially invariant systems such as power networks and smart grids [130, 98, 9, 42, 99] are premised on having *a priori* known and more often static communication topology which is not the case in a spatially distributed multi-robot system with a time-varying communication network that is subject to ad hoc connections [79, 59].

Considerable effort has been devoted to the detection and mitigation of adversarial attacks on multi-agent systems, allowing for a characterization of resilience in terms of a maximum number of malicious (compromised) agents tolerable in a given network. The prevailing approaches in multi-agent settings can be classified as graph-theoretic [101, 52] and system-theoretic approaches [97, 98, 99]. It has been shown that the resilience to a given type of adversarial attack and/or a group of malicious (compromised) agents in a network can be characterized through certain connectivity-related properties of the underlying communication network. Particularly for coordination/consensus of multi-agent systems with first-order dynamics, primary studies have characterized the worst-case bounds for the total number of malicious (non-cooperative) agents that can be detected and identified in a given static communication network with certain degree of vertex-connectivity [97].

Alternatively, to circumvent the detection and identification problems, a family of algorithms, known as Mean-Subsequence Reduced (MSR) algorithms, were devel-

oped [67] for resilient consensus. The MSR-like algorithms enable each agent (robot) to simply disregard part of the information received from its neighbors, ensuring the exclusion of malicious data to a certain degree. To ensure sufficient redundancy of exchanged information between agents, a connectivity-related notion of graph r-robustness was proposed in [67, 146] that allows for the quantification of resilience to a certain number of malicious agents in a given static network. The MSR-like algorithms have been also extended to the cases of multi-agent systems with double-integrator dynamics [36, 37], and higher-order dynamics [110]. See also the surveys in [52, 101].

However, the common challenge of the above-mentioned results in the context of mobile autonomous robots with time-varying communication networks is to maintain the connectivity constraints constantly throughout time. Although connectivity maintenance for multi-robot systems with limited communication capability has a rich history in the literature [79, 33, 53, 59, 39, 43, 55], the prior studies do not address the detection of attacks or non-cooperative robots in time-varying networks. Only a few studies have recently considered resilient consensus over time-varying communication networks. In [116], a hybrid controller was proposed for achieving resilient flocking. Consensus over networks with stochastic link failures and noisy communication was studied in [110]. Consensus and leader-follower consensus over periodically time-varying networks with intermittent communication were considered in [144, 135]. Another defense mechanism is the incorporation of strategically switching communication networks to minimize the space of possible stealthy attacks such as ZDA. We refer to [129, 78] and the references therein for a comprehensive review.

Finally, among the other recent but less prevalent approaches are the study in [14] proposing a decentralized framework to detect communication and sensor attacks that are stealthy to the current state-of-the-art residual-based methods. Also,

in [86], a control barrier function (CBF) approach was proposed for safety and objective specifications serving as metrics for the identification of adversarial agents and resilient control of multi-agent systems.

1.4 Statement of Contributions

This dissertation extends the prior results on adversary detection for resilient multi-agent (robot) systems to the case where the communication network is subject to topology switching and a priori unknown intermittent connections. Additionally, we consider adversarial image attacks on the robot's perception module on which the robot relies for localization in a map. More specifically, this dissertation considers coordination (e.g., consensus) and cooperation (e.g., formation) of multi-robot systems with second-order dynamics in adversarial settings where either or some of the CPS security goals, introduced in Section 1.1, are compromised (see also Fig. 1.2). In this dissertation, a group of malicious (compromised) robots introduces a class of deception attacks to disrupt the normal operation of the system. We consider vulnerabilities to data injection attacks and stealthy attacks, namely zero-dynamics attacks and covert attacks. We provide control-theoretic and graph-theoretic bounds that characterize the resilience of the multi-robot systems with the consensus (coordination) task to the foregoing adversarial attacks.

Chapter 3 considers the security goals of confidentiality and integrity for multi-agent systems (see Section 1.1). It presents[2] a two-layered decentralized attack detection framework to detect stealthy attacks, namely covert attacks and zero-dynamics attacks, on multi-agent control systems seeking consensus. The detection structure

[2]Chapter 3 is adapted from a publication by the author of this dissertation. © 2021 IEEE. Reprinted, with permission, from Bahrami, M., & Jafarnejadsani, H. (2021, December). Privacy-preserving stealthy attack detection in multi-agent control systems. In 2021 60th IEEE Conference on Decision and Control (CDC) (pp. 4194-4199).

consists of a global (central) observer and local observers for the multi-agent system partitioned into clusters. The proposed structure addresses the scalability of the approach and the privacy preservation of the multi-agent system's state information. The former is addressed by using decentralized local observers, and the latter is achieved by enforcing unobservability at the global level. Also, the communication graph model is subject to topology switching, triggered by local observers, allowing for the detection of stealthy attacks by the global observer. Theoretical conditions are derived for detectability of the stealthy attacks using the proposed detection framework.

Chapter 4 considers experimental studies of detecting stealthy attacks on networked UAVs in formation control settings. Compared to the results of Chapter 3, this chapter presents[3] an alternative local monitoring approach that allows for the distribution detection of stealthy attacks for relatively smaller networks of UAVs. The local detection framework, implemented onboard each UAV in the network, uses the model of networked UAVs and locally available measurements. Additionally, the software package developed for this dissertation has been released as an open-source project, which is available at `https://github.com/SASLabStevens/TelloSwarm`. Additionally, a video demonstration of our framework and experimental results is available at `https://www.youtube.com/watch?v=lVT_muezKLU`.

Chapter 5 considers[4] the security goals of availability and integrity for the multi-agent systems whose communication network is arbitrarily time-varying and

[3]Chapter 4 is adapted from a publication by the author of this dissertation. © 2022 IEEE. Reprinted, with permission, from Bahrami, M., & Jafarnejadsani, H. (2022, June). Detection of Stealthy Adversaries for Networked Unmanned Aerial Vehicles. In 2022 International Conference on Unmanned Aircraft Systems (ICUAS) (pp. 1111-1120).

[4]Chapter 5 is adapted from a publication by the author of this dissertation. © 2024 IEEE. Reprinted, with permission, from Bahrami, M., & Jafarnejadsani, H. (2024, August). Distributed Detection of Adversarial Attacks for Resilient Cooperation of Multi-Robot Systems with Intermittent Communication. Provisionally Accepted at IEEE Transactions on Control of Network Systems.

subject to intermittent connections, possibly imposed by denial-of-service (DoS) attacks. The results of this chapter extend observer-based approaches [97, 78] of adversary detection by relaxing their dependency on point-wise-in-time network connectivity/robustness and quantifying resilience to concurrent adversarial attacks. Specifically, this chapter presents explicit bounds for network connectivity in an integral sense that allows for the characterization of the system's resilience to certain classes of adversarial attacks. It will be shown that under connectivity in an integral sense uniformly in time, the system is finite-gain $\mathcal{L}$ stable and uniformly exponentially fast consensus and formation are achievable, provided malicious agents are detected and isolated from the network. This chapter also presents a distributed and reconfigurable framework with theoretical guarantees for detecting malicious agents, allowing for the resilient cooperation of the remaining cooperative agents. We have released our principled framework as an open-source project, which is available at `https://github.com/SASLabStevens/rescue`.

Chapter 6 presents a framework for perception-based multi-robot coordination subject to a class of adversarial image attacks. Specifically, the framework tackles adversarial image perturbations that lead to misclassification and mislocalization in the learned perception model, which performs object detection on onboard camera images to provide detection measurements for relative localization on a map. We propose that the effect of misclassification and mislocalization can be formulated as sporadic (intermittent) and spurious (false positive) measurement data. We propose a method for integrating data from Visual-Inertial Odometry (VIO) and the learned perception model to achieve robust relative localization and state estimation in the presence of sporadic and spurious measurements, which may be caused by adversarial image perturbations targeting the perception module. To test our proposed framework, we also present two multi-robot platforms equipped with open-source software packages for running

learned perception modules and wireless communication capabilities. Our packages are available at `https://github.com/SASLabStevens/TelloSwarm` and `https://github.com/SASLabStevens/AutonomyStack`.

Finally, Chapter 7 summarizes the findings of this dissertation and outlines future research directions.

Chapter 2

Preliminaries

2.1 Notation

We use $\mathbb{R}$, $\mathbb{R}_{>0}$, $\mathbb{R}_{\geq 0}$, $\mathbb{N}$, $\mathbb{Z}_{\geq 0}$, and $\mathbb{C}$ to denote the set of reals, positive reals, nonnegative reals, natural, nonnegative integers, and complex numbers, respectively. $\mathbf{1}_n$, $\mathbf{0}_n$, I_n and $\mathbf{0}_{n\times m}$ stand for the n-vector of all ones, the n-vector of all zeros, the identity n-by-n matrix, and the n-by-m zero matrix, respectively[1]. We use $\mathfrak{e}_n^i$ to denote the i-th canonical vector in $\mathbb{R}^n$, and $\|\cdot\|_p$ to denote the p-norm Euclidean (resp. infinity) norm of vectors and the induced norm of matrices. In addition, for any piecewise continuous, real-valued Lebesgue measurable signal $x(t) \in \mathbb{R}^n$, we use $\left\|(x)_{T_d}\right\|_{\mathcal{L}_p}$, were $1 \leq p \leq \infty$ and $T_d \in [0, \infty)$, to denote the $\mathcal{L}_p$ norm of its truncation signal that is defined as

$$(x)_{T_d} = \begin{cases} x(t), & 0 \leq t \leq T_d, \\ \mathbf{0}, & T_d < t. \end{cases}$$

The extended space $\mathcal{L}_{pe}$ consists of all measurable signals whose truncations belong to $\mathcal{L}_p$, that is $\mathcal{L}_{pe} = \{x(t) \mid (x)_{T_d} \in \mathcal{L}_p, \forall\, T_d \in [0, \infty)\}$. Also, the notation $x^{(\mathrm{m})}(t)$ denotes the m-th order time derivative of $x(t)$. The notations $\mathrm{spec}(\cdot)$ and $\lambda_i(\cdot)$ denote the spectrum of a matrix in the ascending order by magnitude and its i-th eigenvalue, respectively. We use $\mathrm{col}(\cdot)$ and $\mathrm{diag}(\cdot)$ to denote the column and diagonal concatenation of vectors or matrices, and $\otimes$ to denote the Kronecker product. The support of vector $x \in \mathbb{R}^n$ is the set of nonzero components defined as

[1] We may omit the subscripts when clear from the context.

$\text{supp}(x) = \{i \in \{1, \ldots, n\} \mid x_i \neq 0\}$. We also define the set of nonzero columns of the n-by-n matrix M by $\text{colsupp}(M) = \{i \in \{1, \ldots, n\} \mid [M]_{:,i} \neq \mathbf{0}_n\}$. Finally, for any set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality, and for any subset of indices $\mathcal{A} \subset \mathcal{V} = \{1, 2, \ldots, N\}$, where $N \in \mathbb{N}$, the binary matrix $I_{\mathcal{A}} \in \mathbb{R}^{N \times F}$ denotes the concatenation of the i-th columns of $I_{|\mathcal{V}|}$ where $i \in \mathcal{A}$ (i.e. $I_{\mathcal{A}} = \begin{bmatrix} \mathfrak{e}_N^{i_1} & \mathfrak{e}_N^{i_2} & \ldots & \mathfrak{e}_N^{i_{|\mathcal{A}|}} \end{bmatrix}$).

2.2 Graph Theory

We let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected graph with the set of nodes $\mathcal{V} = \{1, 2, \ldots, N\}$, where $N \in \mathbb{N}$, and the set of edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. For any pair of nodes $i, j \in \mathcal{V}$, $i \neq j$, the edge $(j, i) \in \mathcal{E}$ indicates a path from the j-th node to the i-th node. Accordingly, the symmetric adjacency matrix $\mathsf{A} \coloneqq [a_{ij}] \in \mathbb{R}_{\geq 0}^{N \times N}$ is defined such that $a_{ij} > 0$ if and only if $(j, i) \in \mathcal{E}$, and otherwise $a_{ij} = 0$. The Laplacian matrix $\mathsf{L} \coloneqq [l_{ij}] \in \mathbb{R}^{N \times N}$ is defined as $l_{ii} = \sum_{j \neq i} a_{ij}$ and $l_{ij} = -a_{ij}$ if $i \neq j$. $\mathcal{N}^{i(1)} = \{j \in \mathcal{V} \mid (j, i) \in \mathcal{E}\}$ denotes the set of (1-hop) neighbors of node i. The Laplacian matrix of an undirected graph is symmetric and its spectrum, $\text{spec}(\mathsf{L})$ admits $0 = \lambda_1(\mathsf{L}) \leq \lambda_2(\mathsf{L}) \leq \cdots \leq \lambda_N(\mathsf{L}) \leq N$ (see [48, Ch. 13]).

Definition 2.2.1. (Algebraic connectivity). The second-smallest eigenvalue λ_2 of the symmetric Laplacian matrix L of an undirected graph $\mathcal{G}$ is called the algebraic connectivity of $\mathcal{G}$. λ_2 is also referred to as the Fiedler eigenvalue.

An undirected graph $\mathcal{G}$ is *connected* if and only if its *algebraic connectivity* is positive i.e. $\lambda_2(\mathsf{L}) > 0$.

Definition 2.2.2. (Graph component [48, Ch. 1.2], [87, Ch. 6.12]**).** The components of an undirected graph $\mathcal{G}$ are its maximal connected induced subgraphs; that is there exists at least a path connecting every two nodes of a component but not from

a node in the component to any other nodes of the graph. A component is trivial if it has no edges; the special case where a singleton node of a graph is connected to no others is considered to be a component of size one.

Definition 2.2.3. (Vertex and edge connectivity). For a connected graph $\mathcal{G}$, a vertex cutset (resp. an edge cutset) is a set of vertices (resp. edges) whose deletion increases the number of connected components of $\mathcal{G}$. The vertex connectivity, denoted by $\boldsymbol{\kappa}(\mathcal{G})$, (resp. edge connectivity, denoted by $\boldsymbol{e}(\mathcal{G})$) is the minimum number of vertices (resp. edges) in a vertex cutset (resp. edge cutset). Accordingly, a graph is called $\boldsymbol{\kappa}$-vertex-connected (or simply $\boldsymbol{\kappa}$-connected) if $\boldsymbol{\kappa} \leq \boldsymbol{\kappa}(\mathcal{G}) \in \mathbb{R}_{>0}$.

Definition 2.2.4. (r-robust graph [67]**).** A static graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is r-robust, where $r = r(\mathcal{G}) \in \mathbb{Z}_{\geq 0}$ with $0 \leq r(\mathcal{G}) \leq \lceil |\mathcal{V}|/2 \rceil$, if for any pair of nonempty, disjoint subsets of $\mathcal{V}$, at least one of the subsets, denoted as $\mathcal{S}$, holds $|\mathcal{N}^{i(1)} \setminus \mathcal{S}| \geq r, \ \exists\, i \in \mathcal{S}$.

We make the convention that the time-varying graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ with a right-continuous switching signal $\sigma(t) : \mathbb{R}_{\geq 0} \to \{1, 2, \ldots, \mathbf{q}\} =: \mathcal{Q}$, where $\mathbf{q} \in \mathbb{N}$, denotes a finite[2] set of graphs, indexed by finite set $\mathcal{Q}$, that each holds all properties of graph $\mathcal{G}$. For instance, an undirected graph $\mathcal{G}_{\sigma(t)}$ is connected at *a given time instant* $t = t' \in \mathbb{R}_{\geq 0}$ if and only if its *algebraic connectivity* holds $\lambda_2(\mathsf{L}_{\sigma(t')}) > 0$. Also, $\mathcal{G}_{\sigma(t)}$ is $\boldsymbol{\kappa}$-vertex-connected at *a given time instant* $t = t' \in \mathbb{R}_{\geq 0}$ if $\boldsymbol{\kappa} \leq \boldsymbol{\kappa}(\mathcal{G}_{\sigma(t')}) \in \mathbb{R}_{>0}$.

We also make the convention that, in any active mode $\sigma(t) \in \mathcal{Q}$ (σ in short), $\mathcal{N}_\sigma^{i(k)} \subseteq \mathcal{V}$ denotes the set of k-hop neighbors of the agent $i \in \mathcal{V}$, that is for $j \in \mathcal{N}_\sigma^{i(k)}$ there exists a path of length k, where $k \in \mathbb{Z}_{\geq 0} \setminus \{0\}$, in mode σ, between the agents i and j.

[2]The set of possible communication graphs, $\mathcal{Q}$, is finite by $2^{\binom{N}{2}}$ possible cases because an undirected graph with N nodes at most is complete with $\binom{N}{2} = N(N-1)/2$ edges [139, Ch. 1, P. 11] .

2.3 Dynamical Systems Theory

A linear system of the form $\dot{x}(t) = Ax(t)+Bu(t)$ with the output $y(t) = Cx(t)+Du(t)$, where $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m, y(t) \in \mathbb{R}^p$, is represented by the tuple $\Sigma(A, B, C, D)$.

Definition 2.3.1. (Zeroing direction and zero-dynamics attack [152, Ch. 3], [78]**).** Scalar $\lambda_o \in \mathbb{C}$ is a zero of the tuple $\Sigma(A, B, C, D)$ if, and only if, there exists zeroing direction $\text{col}(\mathbf{x}_0, \mathbf{u}_0) \neq \text{col}(\mathbf{0}, \mathbf{0})$ associated with λ_o such that

$$\begin{bmatrix} \lambda_o I_n - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{2.3.1}$$

Then, the signal $u(t) = \mathbf{u}_0 e^{\lambda_o t}$ is a zero-dynamics attack that generates non-zero state trajectories $x(t) = \mathbf{x}_0 e^{\lambda_o t}$ while the output $y = Cx + Du$ satisfies $y(t) = \mathbf{0}$.

Lemma 2.3.1. ***(Observability of linear switched systems*** *[128]**).*** *Given a system of the form $\dot{\mathbf{x}} = \mathbf{A}_{\sigma(t)}\mathbf{x}$, with measurements $\mathbf{y} = \mathbf{C}\mathbf{x}$, ($\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^p$), over the interval $[t_0, t_\mathbf{m})$ that includes switching instances $\{t_k\}_{k=1}^{\mathbf{m}-1}$ for modes $\sigma(t) = k \in \mathcal{Q}$ with the dwell time $\tau_k = t_k - t_{k-1}$, the output of system is given by $\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}_k(t-t_{k-1})} \prod_{l=k-1}^{1} e^{\mathbf{A}_l(\tau_l)}\mathbf{x}(t_0)$, $t \in [t_{k-1}, t_k)$. Then, the following statements hold:*

1. *The system is observable and the initial condition $\mathbf{x}(t_0)$ is reconstructable from $\mathbf{y}(t)$ if, and only if, the matrix $\boldsymbol{\mathcal{O}}$ defined in* (2.3.2) *is full rank (i.e., $\boldsymbol{\mathcal{N}}_1^{\mathbf{m}} := \ker(\boldsymbol{\mathcal{O}}) = \{0\}$).*

2. *If the matrix $\boldsymbol{\mathcal{O}}$ in* (2.3.2) *is rank deficient, the unobservable subspace of the system for $t \in [t_0, t_\mathbf{m})$, which is the largest $\mathbf{A}_{\sigma(t)}$-invariant subspace contained in $\ker(\mathbf{C})$, can be recursively computed using* (2.3.3)-(2.3.4).

$$\boldsymbol{\mathcal{O}} = \mathrm{col}(\mathcal{O}_1, \mathcal{O}_2 e^{\mathbf{A}_1\tau_1}, \cdots, \mathcal{O}_{\mathbf{m}} \prod_{i=\mathbf{m}}^{1} e^{\mathbf{A}_i\tau_i}), \tag{2.3.2}$$

$$\boldsymbol{\mathcal{N}}_{\mathbf{m}}^{\mathbf{m}} = \ker(\mathcal{O}_{\mathbf{m}}), \tag{2.3.3}$$

$$\boldsymbol{\mathcal{N}}_{k}^{\mathbf{m}} = \ker(\mathcal{O}_k) \cap \left[\bigcap_{i=k+1}^{\mathbf{m}} \ker \left(\mathcal{O}_i \prod_{j=i-1}^{k} e^{\mathbf{A}_j\tau_j} \right) \right], \tag{2.3.4}$$

where

$$\mathcal{O}_k = \mathrm{col}\left(\mathbf{C}, \mathbf{C}\mathbf{A}_k, \ldots, \mathbf{C}\mathbf{A}_k^{2N-1}\right), \ 1 \leq k \leq \mathbf{m}-1, \tag{2.3.5}$$

$$\mathbf{A}_k = \mathbf{A}_{\sigma(t)}, \quad t \in [t_{k-1}, t_k). \tag{2.3.6}$$

Chapter 3

Privacy-Preservation and Stealthy Attack Detection for Multi-Agent Control Systems

Motivated by safety and security specifications concerning the vulnerability of wireless communication networks to cyberattacks such as data injection attacks [36, 78, 97], this section presents[1] an attack detection framework for consensus-based coordination of multi-agent systems subject to privacy-preservation constraints on the exchanged information.

Having the confidentiality of transmitted information including agents' initial condition (i.e., position and velocity states) and the final agreement value (consensus value) as a security goal, we propose enforced unobservability constraints on the network topology to preserve the privacy of state information at the global level (i.e. network-level dynamics). Second, we propose a glocal (global-local) attack detection framework for which the networked multi-agent system is partitioned into clusters (subsystems) with their respective globally and locally monitored agents that satisfy specific conditions related to the network privacy and the detectability of stealthy attacks, namely zero-dynamics attack and covert attack. Finally, we derive the theoretical conditions for topology switching (Theorem 3.2.4) under which local detectors trigger switches in the system's communication topology such that stealthy attacks become detectable for the global (centralized) observer. We further discuss different types of topology switching and their outcome for the detection of stealthy attacks.

[1]This chapter is adapted from a publication by the author of this dissertation.

3.1 Problem Formulation

3.1.1 System Dynamics and Communication Topology

Agent dynamics. Consider a multi-agent system consisting of $N \geq 3$ mobile agents with double-integrator dynamics as follows:

$$\Sigma_i : \begin{cases} \dot{\boldsymbol{p}}_i(t) = \boldsymbol{v}_i(t) \\ \dot{\boldsymbol{v}}_i(t) = \boldsymbol{u}_i(t) \end{cases}, \qquad i \in \mathcal{V} = \{1, \ldots, N\}, \tag{3.1.1}$$

in which $\boldsymbol{p}_i(t) \in \mathbb{R}$, and $\boldsymbol{v}_i(t) \in \mathbb{R}$ are the position and velocity states. $\boldsymbol{u}_i(t) \in \mathbb{R}$ denotes the control input[2] of each mobile agent to be computed given local information exchange with its 1-hop neighbors, $\mathcal{N}_\sigma^{i(1)}$, over a switching communication network $\mathcal{G}_{\sigma(t)}$ with finite number of modes $\sigma(t)$'s. Motivated by the vulnerability of wireless communication networks to deception and DoS attacks [18, 78, 97], we let an unknown subset of agents, denoted by $\mathcal{A} \subset \mathcal{V}$ and referred to as *malicious* agents, update their control inputs $\boldsymbol{u}_i$, $i \in \mathcal{A}$, such that $\boldsymbol{u}_i = \boldsymbol{u}_i^{\mathrm{n}} + \boldsymbol{u}_i^{\mathrm{a}}$ in (3.1.1), where $\boldsymbol{u}_i^{\mathrm{n}}$ is the *normal* control input (to be designed) and $\boldsymbol{u}_i^{\mathrm{a}}$ is an injected attack signal. We also refer to the rest of the agents, $\mathcal{V} \setminus \mathcal{A}$, as *cooperative* (or *normal*) agents. In this adversarial setting, the *cooperative* (resp. *malicious*) agents seek to achieve (resp. prevent) the cooperation objective that is defined as

$$\lim_{t\to\infty} \left| \boldsymbol{p}_i(t) - \boldsymbol{p}_j(t) - \boldsymbol{p}_{ij}^\star \right| = \boldsymbol{0}, \qquad \forall\, i, j \in \mathcal{V}, \tag{3.1.2a}$$

$$\lim_{t\to\infty} |\boldsymbol{v}_i(t)| = \boldsymbol{0}, \qquad \forall\, i \in \mathcal{V}, \tag{3.1.2b}$$

[2]For brevity, we may omit the time argument, t, from expressions whenever possible in the rest of this chapter.

where the predefined constants $\boldsymbol{p}_{ij}^{\star} = \boldsymbol{p}_i^{\star} - \boldsymbol{p}_j^{\star}$ are the desired relative positions for any pair of mobile agents in the cooperative settings (e.g., formation control). In this section, we consider coordination problems (e.g., the consensus of the system states) with $\boldsymbol{p}_{ij}^{\star} = 0$.

Having specified the cooperation and adversary objectives, we consider the following distributed control protocol

$$\boldsymbol{u}_i = \boldsymbol{u}_i^{\mathrm{n}} + \boldsymbol{u}_i^{\mathrm{a}}, \qquad i \in \mathcal{V}, \tag{3.1.3a}$$

$$\boldsymbol{u}_i^{\mathrm{n}} = -\alpha \sum_{j \in \mathcal{N}_\sigma^{i(1)}} a_{ij}^{\sigma}(\boldsymbol{p}_i - \boldsymbol{p}_j - \boldsymbol{p}_{ij}^{\star}) - \gamma \boldsymbol{v}_i, \tag{3.1.3b}$$

which relies only on communication with 1-hop neighbors $\mathcal{N}_\sigma^{i(1)}$. Also, the constants $\alpha, \gamma \in \mathbb{R}_{>0}$ are the control gains, and a_{ij}^{σ}'s are the entries of the symmetric adjacency matrix $\mathsf{A}_{\sigma(t)}$ associated with the graph $\mathcal{G}_{\sigma(t)}$ representing the switching communication network of agents Σ_i's in (3.1.1).

Communication topology. The switching communication network (topology) of $N \geq 3$ mobile agents, indexed by the set $\mathcal{V} = \{1, \ldots, N\}$, is described by a finite collection of undirected graphs $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$, where the edge set $\mathcal{E}_{\sigma(t)} \subset \mathcal{V} \times \mathcal{V}$ denotes the communication links. More specifically, an edge $(i, j) \in \mathcal{E}_{\sigma(t)}$ if and only if the i-th and j-th agents are adjacent neighbors exchanging information in the active communication mode $\sigma(t) \in \{1, 2, \ldots, \mathbf{q}\} =: \mathcal{Q}$, $\mathbf{q} \in \mathbb{N}$.

Assumption 3.1.1. *The agents Σ_i's in* (3.1.1) *initially communication in a normal mode of communication topology $\mathcal{G}_{\sigma(t)}$ specified by $\sigma(t) = 1 \in \mathcal{Q}$, for all $t \in [t_0, t_1)$, where $t_1 > t_0 \in \mathbb{R}_{\geq 0}$, until switching to a safe mode following the detection of an attack at a time instant $t_1 > t_{\mathrm{a}}$, where t_{a} is the attack's starting time. In the safe mode for $t \geq t_1$, the communication topology switching is specified by the switching*

signal $\sigma(t) = \{2, \ldots, \mathbf{q}\} \in \mathcal{Q}$ *whose switching policy will be determined later (See Section 3.2.5).*

Network-level dynamics. Given (3.1.1) and (3.1.3), the network-level dynamics of the multi-agent system can be represented by a family of linear switched systems as follows:

$$\Sigma_{\sigma(t)}: \begin{bmatrix} \dot{\boldsymbol{p}} \\ \dot{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I \\ -\alpha \mathsf{L}_\sigma & -\gamma I \end{bmatrix} \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{v} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ I_{\mathcal{A}} \end{bmatrix} \mathbf{u}_{\mathcal{A}} =: \mathbf{A}_\sigma \mathbf{x} + \mathbf{B}_{\mathcal{A}} \mathbf{u}_{\mathcal{A}}, \qquad \mathbf{x}_0 = \mathbf{x}(t_0), \tag{3.1.4a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} - \mathbf{u}_{\mathcal{S}}, \quad \mathbf{C} = \operatorname{diag}\{C_{\mathrm{p}}, C_{\mathrm{v}}\}, \tag{3.1.4b}$$

where $\mathbf{x} = \operatorname{col}(\boldsymbol{p}, \boldsymbol{v})$ with $\boldsymbol{p} \in \mathbb{R}^N$ and $\boldsymbol{v} \in \mathbb{R}^N$ being the stacked position states and velocity states of all N agents. L_σ is the Laplacian matrix of the network $\mathcal{G}_{\sigma(t)}$, encoding the communication links. $\mathbf{u}_{\mathcal{A}} = \operatorname{col}\left(\boldsymbol{u}_i^{\mathrm{a}}\right)_{i \in \mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and $I_{\mathcal{A}} = \left[\mathfrak{e}_N^{i_1}\ \mathfrak{e}_N^{i_2}\ \ldots\ \mathfrak{e}_N^{i_{|\mathcal{A}|}}\right] \in \mathbb{R}^{N \times |\mathcal{A}|}$, where $\mathfrak{e}_N^i$ specifies the input direction in $\mathbb{R}^N$, corresponding to the i-th malicious agent among N agents. The system measurements $\mathbf{y} = \operatorname{col}(\mathbf{y}_1, \cdots, \mathbf{y}_{|\mathcal{M}|})$ corresponding to the output matrix $\mathbf{C}$ such that:

$$\operatorname{colsupp}(C_{\mathrm{k}}) \in \mathcal{M}_{\mathrm{k}} \subset \mathcal{V}, \quad \mathrm{k} \in \{\mathrm{p}, \mathrm{v}\}, \quad \mathcal{M} = \{\mathcal{M}_{\mathrm{p}}, \mathcal{M}_{\mathrm{v}}\}, \tag{3.1.5}$$

where the set $\mathcal{M}$ (to be selected) represents the set of indices of agents that are monitored (e.g., by a ground control station). Finally, $\mathbf{u}_{\mathcal{S}} = \operatorname{col}(u_{\mathrm{s}_1}, \ldots, u_{\mathrm{s}_{|\mathcal{M}|}})$ is a vector of injected malicious signals in the compromised measurement sensor channels.

3.1.2 Adversary Model

Let $\mathcal{A} \subset \mathcal{V}$ denote the set of agents with a compromised (under attack) control channel, and $\mathcal{S} \subset \mathcal{M}$ represent the set of agents with compromised sensor channels. The dynamics of the adversarial attack are given by[3]

$$\Sigma_{\mathcal{A}} : \begin{cases} \dot{\tilde{\mathbf{x}}} = \tilde{\mathbf{A}}_{\sigma(t)}\tilde{\mathbf{x}} + \mathbf{B}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}}, \quad \tilde{\mathbf{x}}(t_{\mathrm{a}}) = \tilde{\mathbf{x}}_0, \\ \mathbf{u}_{\mathcal{S}} = \tilde{\mathbf{C}}\tilde{\mathbf{x}}, \end{cases} \tag{3.1.6}$$

where the vector attack $\mathbf{u}_{\mathcal{A}}$ is generally a function of disclosed information, i.e., $\mathbf{u}_{\mathcal{A}} \coloneqq f(t, \tilde{\mathbf{x}}, \mathbf{u}_i, \mathbf{y})$ by which the attacker steers the system towards undesired states, and $t_{\mathrm{a}} \geq t_0$ is the attack's starting time. For example, the attack signal is in the form of $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{u}_0 e^{\lambda_{\mathrm{o}}(t-t_{\mathrm{a}})}$ in the case of ZDA, where λ_{o} and $\mathbf{u}_0$ are introduced in Definition 2.3.1.

Assumption 3.1.2. ***(Disclosed information).*** *In the normal mode, where $\sigma(t) = 1 \in \mathcal{Q}$, $t \in [t_0, t_1)$, the attacker*

1. *has perfect knowledge of the system model, that is* $\Sigma_{\mathcal{A}}(\tilde{\mathbf{A}}_{\sigma(t)}, \mathbf{B}_{\mathcal{A}}, \tilde{\mathbf{C}}, \sigma = 1) = \Sigma_{\sigma(t)}(\mathbf{A}_{\sigma(t)}, \mathbf{B}_{\mathcal{A}}, \mathbf{C}, \sigma = 1)$,
2. *does not know the system's initial condition, i.e.,* $\tilde{\mathbf{x}}(t_{\mathrm{a}}) \neq \mathbf{x}(t_0)$, *and* $\tilde{\mathbf{x}}(t_{\mathrm{a}}) = \tilde{\mathbf{x}}_0 = \mathbf{0}$ *in a covert attack.*
3. *has no knowledge of the system switching time instants* $\{t_k\}_{k=1}^{\mathbf{m}-1}$, *where* $\mathbf{m} \in \mathbb{N}$, *associated with the safe mode when* $\sigma(t) = \{2, \ldots, \mathbf{q}\} \in \mathcal{Q}$, $t \in [t_1, \infty)$,
4. *starts the attack at* $t_{\mathrm{a}} \geq t_0 = 0$.

Assumption 3.1.3. ***(Defender's policy).*** *The defender*

[3]The matrix $\mathbf{B}_{\mathcal{A}}$ in (3.1.6) is the same as in (3.1.4).

1. *selects the monitored agents and designs the attack detection framework,*

2. *designs the communication topology for the safe mode and its corresponding switching policy.*

Proposition 3.1.4. ***(Stealthy attacks).*** *Consider system* (3.1.4)*, under the attack model* (3.1.6)*, and Assumptions 3.1.1 and 3.1.2, an attack is stealthy*[4] *if the system output in* (3.1.4) *satisfies*

$$\mathbf{y}(t; \mathbf{x}_0, \mathbf{u}_{\mathcal{A}}, \mathbf{u}_{\mathcal{S}}) = \mathbf{y}(t; \bar{\mathbf{x}}_0, \mathbf{0}, \mathbf{0}), \qquad \forall\, t \in [t_0,\, t_1), \tag{3.1.7}$$

where $\mathbf{x}_0$ *and* $\bar{\mathbf{x}}_0$ *are the actual and possible initial states, respectively. Then,* (3.1.7) *can be realized in two senses;*

1. *Covert Attack: Under Assumption 3.1.2, if the attacker sets the initial condition* $\tilde{\mathbf{x}}(t_{\mathrm{a}}) = \mathbf{0}$ *or alternatively* $\tilde{\mathbf{x}}(t_0) \in \boldsymbol{\mathcal{N}}_1^1 = \ker(\mathcal{O}_1)$ *in* (3.1.6)*, then the attack* $\mathbf{u}_{\mathcal{A}}$ *on* (3.1.4) *is covert, that is there exists a vector* $\mathbf{u}_{\mathcal{S}}$*, injected in* (3.1.4)*, canceling out the effect of* $\mathbf{u}_{\mathcal{A}}$ *on the system output* $\mathbf{y}(t)$*.*

2. *Zero-dynamics Attack (ZDA): the attacker can excite the zero dynamics of the system by an unbounded signal and remains stealthy with no need to alter the system measurements (i.e.,* $\mathbf{u}_{\mathcal{S}}(t) = \mathbf{0}$ *in* (3.1.4)*) if* $\tilde{\mathbf{x}}_0 \in \ker(\mathbf{C})$ *and* $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{u}_0 e^{\lambda_{\mathrm{o}}(t-t_{\mathrm{a}})}$*,* $t_{\mathrm{a}} = t_0$*, where* λ_{o}*,* $\tilde{\mathbf{x}}_0$ *and* $\mathbf{u}_0$ *are obtained using Definition 2.3.1.*

Proof. Clearly before an attack starts, (3.1.7) is met over $t \in [t_0,\, t_{\mathrm{a}})$. Consider $\mathbf{x}(t_{\mathrm{a}})$ as the system states when the attack starts,
(i): in the case of covert attack, the output of the system (3.1.4) with the initial

[4]The stealthy attacks defined by the condition (3.1.7) are also known as undetectable attacks in the literature [98].

normal mode $\sigma(t) = 1$ over $t \in [t_\mathrm{a},\, t_1)$ is given by

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}_1(t-t_\mathrm{a})}\mathbf{x}(t_\mathrm{a}) + \mathbf{C}\int_{t_\mathrm{a}}^{t} e^{\mathbf{A}_1(t-\boldsymbol{\tau})}\mathbf{B}\mathbf{u}_{\mathcal{A}}(\boldsymbol{\tau})d\boldsymbol{\tau} - \mathbf{u}_{\mathcal{S}}(t), \tag{3.1.8}$$

and the last term which is the output of the attacker's model (3.1.6) is given by

$$\mathbf{u}_{\mathcal{S}}(t) = \tilde{\mathbf{C}}e^{\tilde{\mathbf{A}}_1(t-t_\mathrm{a})}\tilde{\mathbf{x}}(t_\mathrm{a}) + \tilde{\mathbf{C}}\int_{t_\mathrm{a}}^{t} e^{\tilde{\mathbf{A}}_1(t-\boldsymbol{\tau})}\mathbf{B}\mathbf{u}_{\mathcal{A}}(\boldsymbol{\tau})d\boldsymbol{\tau}. \tag{3.1.9}$$

Substituting (3.1.9) into (3.1.8) and considering Assumption 3.1.2 yields

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}_1(t-t_\mathrm{a})}(\mathbf{x}(t_\mathrm{a}) - \tilde{\mathbf{x}}(t_\mathrm{a})), \quad t \in [t_\mathrm{a},\, t_1). \tag{3.1.10}$$

The measurement (3.1.10) matches the attack-free response if the attacker simply sets $\tilde{\mathbf{x}}(t_\mathrm{a}) = \mathbf{0}$. Also, in the case $\tilde{\mathbf{x}}(t_\mathrm{a}) \neq \mathbf{0}$, $t_\mathrm{a} = t_0 = 0$, it is immediate from lemma 2.3.1 that if $\tilde{\mathbf{x}}(t_0) \in \boldsymbol{\mathcal{N}}_1^1 \neq \{0\} \implies \mathbf{C}e^{\mathbf{A}_1(t-t_\mathrm{a})}\tilde{\mathbf{x}}(t_0) = \mathbf{0}$, $t_\mathrm{a} = t_0 = 0$ in (3.1.10), and thus $\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}_1(t-t_\mathrm{a})}\mathbf{x}(t_\mathrm{a})$, $t \in [t_\mathrm{a},\, t_1)$. In both of the cases, condition (3.1.7), guaranteeing the covertness of the attack, is met. We, however, focus on the first case under Assumption 3.1.2-(ii), therefore the system state $\mathbf{x}(t)$, without any jump, continuously holds the following

$$\mathbf{x}(t) = \bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(t), \tag{3.1.11}$$

where

$$\tilde{\mathbf{x}}(t) = \mathbf{0} \implies \mathbf{x}(t) = \bar{\mathbf{x}}(t), \qquad \forall\, t \in [t_0,\, t_\mathrm{a}), \tag{3.1.12}$$

$$\tilde{\mathbf{x}}(t) = \int_{t_\mathrm{a}}^{t} e^{\mathbf{A}_1(t-\boldsymbol{\tau})}\mathbf{B}\mathbf{u}_{\mathcal{A}}(\boldsymbol{\tau})d\boldsymbol{\tau}, \qquad \forall\, t \in [t_\mathrm{a},\, t_1), \tag{3.1.13}$$

with $\bar{\mathbf{x}}(t)$, $\forall\, t \in [t_0,\, t_1)$ denoting the state of the system in (3.1.4) in the absence of covert attack (i.e. $\dot{\bar{\mathbf{x}}} = \mathbf{A}_1\bar{\mathbf{x}}$, $\bar{\mathbf{x}}_0 = \mathbf{x}_0$).

(ii): In the case of ZDA, let $t_{\mathrm{a}} = t_0 = 0$ for simplicity, and $\bar{\mathbf{x}}_0 = \mathbf{x}_0 - \tilde{\mathbf{x}}_0$. Under Assumption 3.1.2 and using Definition 2.3.1, the attacker can solve the following:

$$\begin{bmatrix} \lambda_{\mathrm{o}} I - \mathbf{A}_1 & -\mathbf{B}_{\mathcal{A}} \\ \mathbf{C} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_0 \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{3.1.14}$$

to design the ZDA signal $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{u}_0 e^{\lambda_{\mathrm{o}} t}$ causing unbounded system states

$$\mathbf{x}(t) = \bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}_0 e^{\lambda_{\mathrm{o}} t}, \tag{3.1.15}$$

while (3.1.7) is met, where $\bar{x}(t)$ is the state of the system in (5) assuming the initial condition $\bar{x}_0$ and no attack signal. The second equation in (3.1.14), $\mathbf{C}\tilde{\mathbf{x}}_0 = \mathbf{0}$, implies $\tilde{\mathbf{x}}_0 \in \ker(\mathbf{C})$. It is an immediate result from Definition 2.3.1 that the attack signal $\mathbf{u}_a(t) = \mathbf{u}_0 e^{\lambda_0 t}$ results in $\mathbf{u}_{\mathcal{S}}(t) = \mathbf{C}\tilde{\mathbf{x}}(t) = \mathbf{0}$ in (3.1.6) while the system states $\tilde{\mathbf{x}}(t) = \tilde{\mathbf{x}}_0 e^{\lambda_{\mathrm{o}} t} \in \ker(\mathbf{C})$, $\forall\, t \in [t_0, t_1)$ is unboundedly increasing. Consider (3.1.15) and the superposition principle in linear systems, then injecting the designed ZDA signal $\mathbf{u}_a(t)$ in (3.1.4) yields the solution $\mathbf{y} = \mathbf{C}\mathbf{x}(t) = \mathbf{C}\bar{\mathbf{x}}(t) + \mathbf{C}\tilde{\mathbf{x}}_0 e^{\lambda_{\mathrm{o}} t}$, which by considering (3.1.14) is equivalent to (3.1.7), guaranteeing the stealthiness of ZDA for (3.1.4).

3.1.3 Problem Statement

Given the system and attack models in the previous section, we now state the two problems which this chapter aims to address in the following:

Problem 3.1.5. ***(Privacy-preserving average consensus).*** *Given the switching*

consensus system (3.1.4)*, we seek to preserve the following privacy requirements:*

1. *neither the system's initial states* $\mathbf{x}(t_0)$ *nor the final agreement values (*$\boldsymbol{p}^* = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{p}_i(t_0)$*,* $\boldsymbol{v}^* = 0$*) should be revealed or be reconstructable.*

2. *the system's communication topology* $\mathcal{G}_{\sigma(t)}$ *should not be reconstructable.*

Problem 3.1.6. ***(Scalable attack detection).*** *Given the system in* (3.1.4) *under the attack model* (3.1.6)*, we seek to develop a stealthy attack detection framework such that:*

1. *it features a decentralized and scalable structure.*

2. *it satisfies the privacy-preserving requirements defined in Problem 3.1.5.*

3.2 Privacy Preservation and Attack Detection

In this section, we describe the attack detection framework and characterize the conditions required to address Problems 3.1.5 and 3.1.6.

3.2.1 Attack Detection Scheme

The proposed framework, depicted in Fig. 3.1, is a two-level attack detection framework. It is privacy-preserving and relies on topology switching generating model discrepancy between the attacker model (3.1.6) and the actual system (3.1.4). The system is decomposed into a set of subsystems based on the characteristics of its communication topology such as sparsity. Then, a set of monitored agents will be characterized such that each subsystem (the dynamics of agents within a cluster) is fully observable with respect to its locally available measurements while the main system (3.1.4) is partially observable with respect to its globally available measurements (3.1.5). We show how unobservability and system clustering can be used respectively

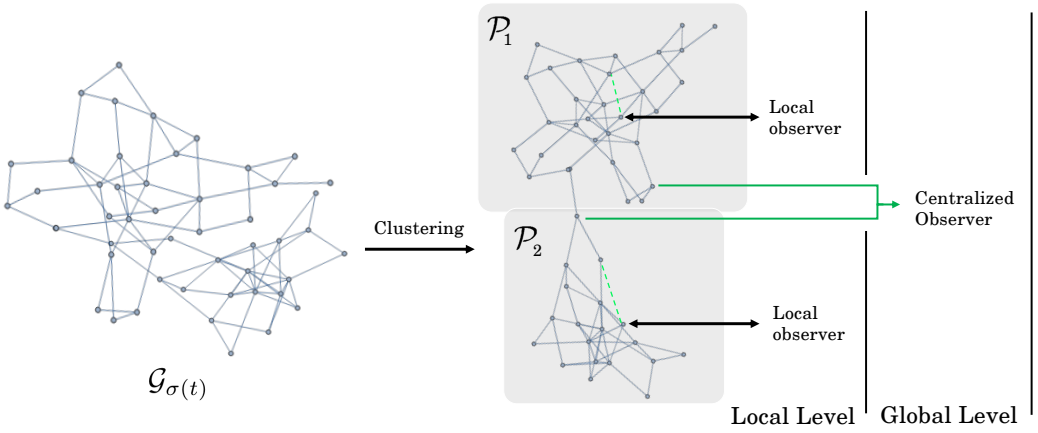


Figure 3.1: Attack detection architecture. It includes a centralized observer and a set of local observers. The centralized observer only monitors some of the agents from a ground station with bandwidth limitation. The local observers deployed onboard allow for local monitoring and local decision-making for network topology switches, enabling the detection of stealthy attacks by the centralized observer of the ground station.

to address Problem 3.1.5 and 3.1.6. Building upon global and (private) local measurements, the attack detection framework consists of a centralized observer, implemented in the control center, and local observer(s) in each cluster ($\mathcal{P}_\mathrm{i}$, $\mathrm{i} \in \{1, 2\}$ in Fig. 3.1). As increasing data transmission between agents and the centralized observer in the control center raises scalability and privacy concerns (cf. Problem 3.1.6), local observers play a vital role in our attack detection framework. They are hidden from the attacker because they are distributed among clusters of the multi-agent system, and their output is not sent to the control center but kept locally for attack detection. If a local observer detects a stealthy attack, it triggers a network topology switch whereby the stealthy attack becomes detectable in the global measurements available for the centralized observer. The local decision-making for network topology switches and indirect communication with the control center allow for agile reconfigurability in autonomous multi-agent systems (e.g., networks of autonomous aerial or ground vehicles) while eliminating the need for additional data exchange at the global level, which otherwise is required for monitoring and stealthy attack detection.

3.2.2 Privacy Preservation

Problem 3.1.5 on privacy preservation can be addressed by imposing an unobservability constraint on system (3.1.4). Indeed, one can select the set of monitored agents $\mathcal{M}$ in (3.1.5) such that $(\mathbf{A}_{\sigma(t)}, \mathbf{C})$ is not an observable pair on $t \in [t_0, \infty)$, making the globally available measurement $\mathbf{y}$ in (3.1.5) insufficient to reconstruct either the entire system states' information or the system's switching structure (cf. privacy requirements in Problem 3.1.5).

The following lemma provides sufficient conditions to determine whether the global system measurement (3.1.5) is consistent with the privacy requirements.

Lemma 3.2.1. ***(Invariant unobservable subspace of system*** (3.1.4)***).*** *The subspace* $\operatorname{span}\left\{ \begin{smallmatrix} \mathbf{1}_N \\ \mathbf{0}_N \end{smallmatrix} \right\}$ *is an* $\mathbf{A}_{\sigma(t)}$*-invariant unobservable subspace of the switching system in* (3.1.4) *provided that it lies in* $\ker(\mathbf{C})$ *and* $\mathcal{G}_{\sigma(t)}$ *features only connected undirected (or strongly connected and balanced directed) graphs.*

Proof. See Appendix A.2.

Remark 3.2.1. ***(Generality of Lemma 3.2.1).*** *The result suggests that monitoring only the agents' velocity causes the agents' positions not to be reconstructable independently for system* (3.1.4). *This is a generic solution to Problem 3.1.5 that holds for all undirected graphs. It is also worth noting that the monitored agents corresponding to set* $\mathcal{M}$ *in* (3.1.5) *can also be selected differently from the results in Lemma 3.2.1 for any particular graph.*

We next introduce the system partitioning method followed by observer design to address Problem 3.1.6.

3.2.3 System Partitioning

Consider the communication graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_\sigma(t))$ of the system (3.1.4), let the set of agents $\mathcal{V}$ be partitioned into disjoint clusters $\boldsymbol{\mathcal{P}} \coloneqq \{\mathcal{P}_1, \ldots, \mathcal{P}_{|\boldsymbol{\mathcal{P}}|}\}$ such that $\cup_{\mathrm{i}=1}^{|\boldsymbol{\mathcal{P}}|} \mathcal{P}_\mathrm{i} = \mathcal{V}$ with $\mathcal{P}_\mathrm{i} \in \mathbb{R}^{N_\mathrm{i}}$ and inter-cluster couplings

$$\mathcal{E}_{\text{cut}} \coloneqq \{(j,i)) \mid i \in \mathcal{P}_\mathrm{i},\, j \in \mathcal{P}_\mathrm{j},\, \mathcal{P}_\mathrm{i} \cap \mathcal{P}_\mathrm{j} = \emptyset\}\,.$$

Accordingly, after relabeling the system states, the system (3.1.4) is partitioned into $|\boldsymbol{\mathcal{P}}|$ subsystems described as

$$\Sigma_{\mathcal{P}_\mathrm{i}} : \begin{cases} \dot{\mathbf{x}}_\mathrm{i} = \mathbf{A}^{\mathrm{i}}_{\sigma(t)}\mathbf{x}_\mathrm{i} + \sum_{\mathrm{j}\in\mathcal{N}_{\mathcal{P}_i}} \mathbf{A}^{\mathrm{ij}}_{\sigma(t)}\mathbf{x}_\mathrm{j} + \mathbf{B}^{\mathrm{i}}_{\mathcal{A}}\mathbf{u}_{a_\mathrm{i}}, \\ \mathbf{y}_{\mathrm{i}_i} = \mathbf{C}_{\mathrm{i}_i}\mathbf{x}_\mathrm{i}, \quad i \in \mathcal{M}_\mathrm{i} \subset \mathcal{P}_\mathrm{i}, \\ \mathbf{x}_\mathrm{i}(0) = \mathbf{x}_{0_\mathrm{i}}, \qquad \mathrm{i} \in \{1, \cdots, |\boldsymbol{\mathcal{P}}|\}\,, \end{cases} \tag{3.2.1}$$

with

$$\mathbf{A}^{\mathrm{i}}_{\sigma(t)} = \begin{bmatrix} 0 & I \\ -\alpha \mathsf{L}^{\mathrm{i}}_{\sigma(t)} & -\gamma I \end{bmatrix}, \qquad \mathbf{A}^{\mathrm{ij}}_{\sigma(t)} = \begin{bmatrix} 0 & 0 \\ -\alpha \mathsf{L}^{\mathrm{ij}}_{\sigma(t)} & 0 \end{bmatrix}, \tag{3.2.2}$$

$$\mathsf{L}_{\sigma(t)} = \begin{bmatrix} \mathsf{L}^{1}_{\sigma(t)} & \cdots & \mathsf{L}^{1,|\boldsymbol{\mathcal{P}}|}_{\sigma(t)} \\ \vdots & \ddots & \vdots \\ \mathsf{L}^{|\boldsymbol{\mathcal{P}}|,1}_{\sigma(t)} & \cdots & \mathsf{L}^{|\boldsymbol{\mathcal{P}}|}_{\sigma(t)} \end{bmatrix}, \qquad \mathbf{B}^{\mathrm{i}}_{\mathcal{A}} = \begin{bmatrix} 0 \\ I_{\mathcal{A}_\mathrm{i}} \end{bmatrix}, \tag{3.2.3}$$

where $\mathbf{x}_\mathrm{i} \coloneqq \begin{bmatrix} (\boldsymbol{p})^\top_\mathrm{i} & (\boldsymbol{v})^\top_\mathrm{i} \end{bmatrix}^\top \in \mathbb{R}^{2N_\mathrm{i}}$ with $(\boldsymbol{p})_\mathrm{i}$ and $(\boldsymbol{v})_\mathrm{i}$ representing the vectors of position and velocity states belonging to cluster $\mathcal{P}_\mathrm{i} \subset \mathcal{V}$. Also, $\mathbf{u}_{a_\mathrm{i}}$ associated with the set $\mathcal{A}_\mathrm{i}$ is the vector-valued attack on actuator channels in the cluster as defined in (3.1.4). The output signal $\mathbf{y}_{\mathrm{i}_i}(t)$, associated with the output matrix $\mathbf{C}_{\mathrm{i}_i}$, denotes

the *local* measurements that are available at node i in cluster $\mathcal{P}_{\mathrm{i}}$. Finally, $\mathcal{N}_{\mathcal{P}_{\mathrm{i}}} := \{\mathrm{j} \in \{1, \cdots, |\boldsymbol{\mathcal{P}}|\} \mid \exists\, (j,i) \in \mathcal{E}_{\mathrm{cut}}, i \in \mathcal{P}_{\mathrm{i}}, j \in \mathcal{P}_{\mathrm{j}}\}$ denotes the index set of the neighboring clusters of cluster $\mathcal{P}_{\mathrm{i}}$.

We note that the decomposition of (3.1.4) into (3.2.1) leads to a concatenated set $\bar{\mathcal{M}} := \{\mathcal{M}, \mathcal{M}_1, \ldots, \mathcal{M}_{|\boldsymbol{\mathcal{P}}|}\}$, where the set $\mathcal{M}$ is associated with *global* measurements (3.1.5) available for the control center and sets $\mathcal{M}_{\mathrm{i}}$'s, $\mathrm{i} \in \boldsymbol{\mathcal{P}}$ are associated with the *local* measurements $\mathbf{y}_{\mathrm{i}_i}$ available at a node i in respective clusters $\mathcal{P}_1, \ldots, \mathcal{P}_{|\boldsymbol{\mathcal{P}}|}$ in (3.2.1).

We make the following assumptions:

Assumption 3.2.2. ***(Local information).***

1. local knowledge*: in each cluster, the agent $i \in \mathcal{P}_{\mathrm{i}}$ serves as the local control center that has the local system model of the cluster (matrices $\mathbf{A}^{\mathrm{i}}_{\sigma(t)}$, $\mathbf{A}^{\mathrm{i}j}_{\sigma(t)}$ and $\mathbf{C}_{\mathrm{i}_i}$) and the local measurement $\mathbf{y}_{\mathrm{i}_i}(t)$.*

2. local measurements*: the measured output $\mathbf{y}_{\mathrm{i}_i}(t)$ in (3.2.1) is locally available at the node i and, unlike global measurements, it is not sent to the control center to keep the output secure and inaccessible to the attacker.*

3. cross-cluster communication*: every local control center, i.e., the node i in cluster $\mathcal{P}_{\mathrm{i}}$, considers coupling terms $\sum_{\mathrm{j} \in \mathcal{N}_{\mathcal{P}_{\mathrm{i}}}} \mathbf{A}^{\mathrm{ij}}_{\sigma(t)} \mathbf{x}_{\mathrm{j}}$ as unknown inputs to $\Sigma_{\mathcal{P}_{\mathrm{i}}}$. Moreover, inter-cluster couplings do not change, i.e., $\mathbf{A}^{\mathrm{ij}}_{\sigma(t)} = \mathbf{A}^{\mathrm{ij}}_1$, $\forall\, t \in [t_0, \infty)$. Thus there is no need for the exchange of $\mathbf{x}_{\mathrm{j}}$'s information between local control centers.*

The assumption 3.2.2-1 is common in the literature (cf. [42]) as the model-based detection of cyber attacks on exchanged data over a network requires augmented

knowledge of the neighboring agents' model to estimate their states and further compare them with the received data. Minimizing the local information exchange affects the scalability and depends on the sparsity of the communication network as well as on applications.

3.2.4 Observer Design and Attack Detectability Analysis

As described in Section 3.2.1, the attack detection framework is composed of a centralized observer for monitoring the system (3.1.4) from the control center, and a set of local observers in clusters, that serve as local attack detectors and trigger for communication topology switching. In what follows, we describe the observer design procedure based on the conditions derived in the previous section.

Decentralized observer. Consider the dynamics of the system partitions described in (3.2.1) and Assumption 3.2.2, we use the unknown input observer (UIO) scheme in [24] to estimate the cluster state $\hat{\mathbf{x}}_{\mathrm{i}}$ independent of the states $\mathbf{x}_{\mathrm{j}}$'s of the neighboring clusters (i.e. $\mathrm{j} \in \mathcal{N}_{\mathcal{P}_{\mathrm{i}}}$). This is achieved by considering the interconnection of local models as unknown inputs and rewriting them such that

$$\sum_{\mathrm{j}\in\mathcal{N}_{\mathcal{P}_{\mathrm{i}}}} \mathbf{A}^{\mathrm{ij}}_{\sigma(t)}\mathbf{x}_{\mathrm{j}} \coloneqq \mathbf{E}^{\mathrm{i}}\mathbf{x}^{d}_{\mathrm{i}}, \quad \sigma(t)=1, \quad \forall\, t\in[t_0,\,\infty), \tag{3.2.4}$$

where $\mathbf{E}^{\mathrm{i}}$ is a full column rank[5] matrix and $\mathbf{x}^{d}_{\mathrm{i}}$ is a vector of the states of neighboring clusters that are received by cluster $\mathcal{P}_{\mathrm{i}}$. Now, introducing the UIO state $\mathbf{z}_{\mathrm{i}} = \hat{\mathbf{x}}_{\mathrm{i}} - \mathbf{h}^{\mathrm{i}}\mathbf{y}_{\mathrm{i}_i}$,

[5]The columns of $\mathbf{E}^{\mathrm{i}}$ for cluster $\mathcal{P}_{\mathrm{i}}$ are corresponding to the edge-cuts connecting $\mathcal{P}_{\mathrm{i}}$ to its neighboring clusters.

the dynamics of the local UIO is given by

$$\Sigma_{\mathcal{O}}^{\mathcal{Z}_{\mathrm{i}}}:\begin{cases}\dot{\mathbf{z}}_{\mathrm{i}}=\mathbf{F}_{\sigma(t)}^{\mathrm{i}}\mathbf{z}_{\mathrm{i}}+\left(\mathbf{K}_{\sigma(t)}+\bar{\mathbf{K}}_{\sigma(t)}\right)\mathbf{y}_{\mathrm{i}_{i}},\\ \hat{\mathbf{x}}_{\mathrm{i}}=\mathbf{z}_{\mathrm{i}}+\mathbf{h}^{\mathrm{i}}\mathbf{y}_{\mathrm{i}_{i}},\\ \hat{\mathbf{x}}_{\mathrm{i}}(0)=\mathbf{0},\quad \mathcal{P}_{\mathrm{i}}\subset\mathcal{V},\quad \mathrm{i}\in\{1,\cdots,|\boldsymbol{\mathcal{P}}|\},\end{cases}\tag{3.2.5}$$

where $\mathbf{F}_{\sigma(t)}^{\mathrm{i}},\mathbf{K}_{\sigma(t)},\bar{\mathbf{K}}_{\sigma(t)}$, and $\mathbf{h}^{\mathrm{i}}$ are matrices satisfying conditions

$$\mathbf{T}^{\mathrm{i}}=\left(I-\mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_{i}}\right),\quad\left(\mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_{i}}-I\right)\mathbf{E}^{\mathrm{i}}=0,\tag{3.2.6}$$

$$\mathbf{F}_{\sigma(t)}^{\mathrm{i}}=\left(\bar{\mathbf{A}}_{\sigma(t)}^{\mathrm{i}}-\bar{\mathbf{K}}_{\sigma(t)}\mathbf{C}_{\mathrm{i}_{i}}\right),\quad\mathbf{K}_{\sigma(t)}=\mathbf{F}_{\sigma(t)}^{\mathrm{i}}\mathbf{h}^{\mathrm{i}},\tag{3.2.7}$$

$$\bar{\mathbf{A}}_{\sigma(t)}^{\mathrm{i}}=\mathbf{A}_{\sigma(t)}^{\mathrm{i}}-\mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_{i}}\mathbf{A}_{\sigma(t)}^{\mathrm{i}}.\tag{3.2.8}$$

Furthermore, $\mathbf{F}_{\sigma(t)}^{\mathrm{i}}$ is Hurwitz stable over $t\in[t_0,\,t_{\mathbf{m}})$ for all *normal* and *safe modes*.

Consider (3.2.1), (3.2.5) and let $\mathbf{e}_{\mathrm{i}}:=\mathbf{x}_{\mathrm{i}}-\hat{\mathbf{x}}_{\mathrm{i}}$, one can use the conditions in (3.2.6)-(3.2.8) to obtain the error dynamics of UIO as follows

$$\Sigma_{\mathcal{O}}^{\mathbf{e}_{\mathrm{i}}}:\begin{cases}\dot{\mathbf{e}}_{\mathrm{i}}&=\mathbf{F}_{\sigma(t)}^{\mathrm{i}}\mathbf{e}_{\mathrm{i}}+\mathbf{T}^{\mathrm{i}}\mathbf{B}^{\mathrm{i}}\mathbf{u}_{a_{\mathrm{i}}},\quad\mathbf{e}_{\mathrm{i}}(0)=\mathbf{x}_{\mathrm{i}}(0),\\ \mathbf{r}_{\mathrm{i}_{i}}&=\mathbf{C}_{\mathrm{i}_{i}}\mathbf{e}_{\mathrm{i}},\quad\mathcal{P}_{\mathrm{i}}\subset\mathcal{V},\quad\mathrm{i}\in\{1,\cdots,|\boldsymbol{\mathcal{P}}|\}.\end{cases}\tag{3.2.9}$$

In the absence of adversarial attacks, $\mathbf{u}_{a_{\mathrm{i}}}=\mathbf{0}$, it is straightforward to show that $\lim_{t\to\infty}\mathbf{e}_{\mathrm{i}}(t)=\mathbf{0}$ as $\mathbf{F}_{\sigma(t)}^{\mathrm{i}}$ is Hurwitz stable in all modes. LMI-based approaches can be used to design (3.2.7) such that (3.2.9) remains stable under arbitrary switching [28].

Recall Assumption 3.2.2-2, unlike the case of global measurements (cf. Proposition 3.1.4-(i)), the local measurements $\mathbf{y}_{\mathrm{i}_{i}}$'s are hidden and thus cannot be altered by the attacker to cancel out the effect of the attack $\mathbf{u}_{a_{\mathrm{i}}}$ on the output of (3.2.1). This difference also manifests itself in the residual of local observer (3.2.9). Therefore, in

order to determine the stealthiness of attack $\mathbf{u}_{a_{\mathrm{i}}}$ with respect to the local residual signal $\mathbf{r}_{\mathrm{i}_i}$, it is necessary and sufficient to investigate whether the stealthiness conditions presented in Proposition 3.1.4 are satisfied for the system in (3.2.9).

In the following proposition, we formally characterize the conditions for the detection of stealthy attacks using the local observer in (3.2.5).

Proposition 3.2.3. ***(Attack detectability of local observers).*** *For a strongly connected cluster $\mathcal{P}_{\mathrm{i}}$ with $\boldsymbol{\mathcal{E}}$ inter-clustering edges and $|\mathcal{A}_{\mathrm{i}}|$ compromised agents, there exists a local observer given by* (3.2.5) *to locally detect the stealthy attacks if*

1. *there is a* $\mathbf{k}$*-connected node* $i \in \mathcal{P}_{\mathrm{i}}$ *as the local monitored agent such that* $\mathbf{k} \geq \boldsymbol{\mathcal{E}} + |\mathcal{A}_{\mathrm{i}}|$,
2. $\operatorname{rank}\left(\mathbf{C}_{\mathrm{i}_i}\mathbf{E}^{\mathrm{i}}\right) = \operatorname{rank}\left(\mathbf{E}^{\mathrm{i}}\right)$,
3. *the matrix pencil* $\mathbf{P}$ *in* (3.2.10) *is full (column) rank,*

$$\mathbf{P} = \begin{bmatrix} \lambda_{\mathrm{o}} I - \mathbf{A}^{\mathrm{i}}_{\sigma(t)} & \mathbf{B}^{\mathrm{i}}_{\mathcal{A}} & \mathbf{E}^{\mathrm{i}} \\ \mathbf{C}_{\mathrm{i}_i} & 0 & 0 \end{bmatrix}. \tag{3.2.10}$$

where the tuple $\left(\mathbf{A}^{\mathrm{i}}_{\sigma(t)}, \mathbf{B}^{\mathrm{i}}, \mathbf{C}_{\mathrm{i}_i}\right)$ *and matrix* $\mathbf{E}^{\mathrm{i}}$ *are defined in* (3.2.1) *and* (3.2.4)*, respectively.*

Proof. See Appendix A.3.

Remark 3.2.2. ***(Evaluation of the condition in*** (3.2.10)***).*** *Conditions* (1)-(3) *in Proposition 3.2.3 are equivalent to necessary and sufficient conditions for the existence of UIO in* (3.2.5) *[24]. It is worth noting that as matrix* $\mathbf{B}^{\mathrm{i}}$ *in* (3.2.10) *is unknown to the defender, it can be replaced with* I_{N_i}*, i.e., assuming all the nodes of the cluster are under attack, in analysis and selecting locally monitored agents associated*

with $\mathbf{C}_{\mathrm{i}_i}$. *This, however, may require further communication between agents within a cluster. Alternatively, as in a set cover problem setting, a set of local monitoring agents that each of them satisfies the conditions* (1)-(3) *for part of a cluster can be used to cover all of the nodes of the cluster [130]. Minimizing the number of local measurements versus the number of local observers is a trade-off problem that will be the subject of future work.*

Centralized observer. Consider the dynamical system (3.1.4), a Luenberger-type centralized observer, derived based on the normal mode $\sigma(t) = 1$, is given by

$$\Sigma_{\mathcal{O}}^{\mathcal{M}} : \begin{cases} \dot{\hat{\mathbf{x}}} = \mathbf{A}_{\sigma(t)}\hat{\mathbf{x}} + \mathbf{H}_{\sigma(t)}(\mathbf{y} - \hat{\mathbf{y}}), & \sigma(t) = 1, \\ \hat{\mathbf{y}} = \mathbf{C}\hat{\mathbf{x}}, & \hat{\mathbf{x}}(0) = \mathbf{0}, \\ \mathbf{r}_0 = (\mathbf{y} - \hat{\mathbf{y}}), & \text{residual}, \end{cases} \tag{3.2.11}$$

where $\mathbf{H}_{\sigma(t)}$ is the observer gain and $\mathbf{r}_0(t)$ denotes the residual signal available in the control center for monitoring purposes.

In order to design the observer gain $\mathbf{H}_{\sigma(t)}$, the partial observability of pair $(\mathbf{A}_{\sigma(t)}, \mathbf{C})$ imposed in Section 3.2.2 and the activated mode $\sigma(t)$ should be taken into account. An immediate solution is to define an LMI optimization problem finding a constant $\mathbf{H}_{\sigma(t)} \coloneqq \mathbf{H}$ by which $(\mathbf{A}_{\sigma(t)} - \mathbf{HC})$ is (Hurwitz) stable in all modes [29, 27].

From Assumption 3.1.2 and condition (3.1.7), it is straightforward to show that the attack $\mathbf{u}_a$ remains stealthy for the observer (3.2.11) in the normal mode over the time span $t \in [t_0, t_1)$ where $\mathbf{A}_{\sigma(t)} = \mathbf{A}_1$.

Recall (3.1.11) and (3.1.15), and let

$$\bar{\mathbf{e}} \coloneqq \bar{\mathbf{x}} - \hat{\mathbf{x}} \tag{3.2.12}$$

$$\mathbf{e} \coloneqq \mathbf{x} - \hat{\mathbf{x}} = \bar{\mathbf{x}} + \tilde{\mathbf{x}} - \hat{\mathbf{x}} = \bar{\mathbf{e}} + \tilde{\mathbf{x}} \tag{3.2.13}$$

be the estimation error of the states of an attack-free system ($\dot{\bar{\mathbf{x}}} = \mathbf{A}_{\sigma(t)}\bar{\mathbf{x}}$, $\mathbf{y} = \mathbf{C}\bar{\mathbf{x}}$) and the under attack system in (3.1.4), respectively. Then using (3.1.4) and (3.2.11), the error dynamics of the centralized observer is given by

$$\Sigma_{\mathcal{O}}^{\mathbf{e}} : \begin{cases} \dot{\mathbf{e}} = (\mathbf{A}_1 - \mathbf{HC})\mathbf{e} + (\mathbf{A}_{\sigma(t)} - \mathbf{A}_1)\mathbf{x} + \mathbf{H}\mathbf{u}_{\mathcal{S}} + \mathbf{B}\mathbf{u}_{\mathcal{A}}, \\ \mathbf{e}(0) = \mathbf{x}_0, \\ \mathbf{r}_0 = (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{C}\mathbf{e} - \mathbf{u}_{\mathcal{S}} = \mathbf{C}\bar{\mathbf{e}}, \quad \text{residual}, \end{cases} \tag{3.2.14}$$

where for measurement $\mathbf{y}$ in (3.2.11) we used the expression $\mathbf{y} = \mathbf{Cx} - \mathbf{u}_{\mathcal{S}}$ as defined in (3.1.4). Consider (3.1.6) and (3.1.7), $\mathbf{y}$ in (3.2.11) also satisfies $\mathbf{y} = \mathbf{Cx} - \mathbf{u}_{\mathcal{S}} = \mathbf{Cx} - \mathbf{C}\tilde{\mathbf{x}} = \mathbf{C}\bar{\mathbf{x}}$. Then using $\mathbf{y} = \mathbf{C}\bar{\mathbf{x}}$, (3.1.4), (3.1.6), (3.2.11), (3.2.12), the following dynamics is obtained

$$\Sigma_{\mathcal{O}}^{\bar{\mathbf{e}}} : \begin{cases} \dot{\bar{\mathbf{e}}} = (\mathbf{A}_1 - \mathbf{HC})\bar{\mathbf{e}} + (\mathbf{A}_{\sigma(t)} - \mathbf{A}_1)\bar{\mathbf{x}}, \\ \bar{\mathbf{e}}(0) = \bar{\mathbf{x}}_0, \\ \bar{\mathbf{r}}_0 = \mathbf{C}\bar{\mathbf{e}}, \quad \text{residual}. \end{cases} \tag{3.2.15}$$

Note that, during normal mode $\sigma(t) = 1$ over the time span $\forall\, t \in [t_0, t_1)$, the residual $\mathbf{r}_0$ in (3.2.14) is the same as that of (3.2.15) that is the dynamics of the estimation error of system states in the absence of attacks. This implies that, in the case of a covert attack with $\mathbf{u}_{\mathcal{S}} \neq 0$, as long as signal $\mathbf{u}_{\mathcal{S}}(t)$ cancels out the effect of $\mathbf{u}_a(t)$ on the output $\mathbf{y}(t)$, the residual $\mathbf{r}_0(t) = \mathbf{C}\bar{\mathbf{e}}(t)$ converges to zero as $t_1 \to \infty$, yielding the stealthiness of the covert attack, in the normal mode, for the centralized observer (3.2.11).

In the case of a ZDA, $\mathbf{u}_{\mathcal{S}} = 0$ in (3.2.14) although (3.1.7) still holds that leads to the stealthiness of a ZDA for the observer (3.2.11). To show this, one needs to verify the attack $\mathbf{u}_{\mathcal{A}}$ remains in the zeroing direction of (3.2.14). Using Definition 2.3.1 for

(3.2.14) in the normal mode, we obtain

$$\begin{bmatrix} \lambda_{\mathrm{o}} I - (\mathbf{A}_1 - \mathbf{HC}) & -\mathbf{B} \\ \mathbf{C} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}(0) \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{3.2.16}$$

where $\tilde{\mathbf{e}}(0) \coloneqq \mathbf{e}(0) - \bar{\mathbf{e}}(0) = \mathbf{x}_0 - \bar{\mathbf{x}}_0 = \tilde{\mathbf{x}}_0$. Recall $\tilde{\mathbf{x}}_0 \in \ker(\mathbf{C})$ in (3.1.14), then the second equation of (3.2.16) yields $\mathbf{C}\tilde{\mathbf{e}}(0) = \mathbf{C}\tilde{\mathbf{x}}_0 = \mathbf{0}$. Applying $\mathbf{C}\tilde{\mathbf{e}}(0) = \mathbf{0}$ into the first equation of (3.2.16) simplifies the matrix pencil in (3.2.16) into that of (3.1.14) over $t \in [t_0,\, t_1)$ where $\mathbf{A}_{\sigma(t)} = \mathbf{A}_1$. This ensures the stealthiness of ZDA in the normal mode for the observer (3.2.11).

The following Theorem provides conditions to address Problem 3.1.6-*2* by characterization of switching modes that lead to attack detection with respect to global measurements.

Theorem 3.2.4. ***(Attack detectability under switching communication).*** *Consider system* (3.1.4) *under the stealthy attacks modeled in* (3.1.6)*, and let intra-cluster topology switching satisfy*

1. $\mathrm{Im}(\Delta \mathsf{L}_{\mathbf{q}}) \cap \ker\left([\mathbf{C}_{\mathrm{p}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}\right) = \emptyset$,
2. $\mathsf{L}_{\mathbf{q}}$ *features distinct eigenvalues,*
3. $[\mathcal{U}_{\mathbf{q}}]_{i,\ell} - [\mathcal{U}_{\mathbf{q}}]_{j,\ell} \neq 0, \quad \forall\, \ell \in \mathcal{V} \setminus \{1\}, \forall\, i, j \in \mathcal{D}_{\mathrm{c}}, \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}$,

where $\Delta \mathsf{L}_{\mathbf{q}} \coloneqq \mathsf{L}_{\sigma(t)} - \mathsf{L}_1$*, with* $\sigma(t) = \mathbf{q} \in \mathcal{Q}$*,* $t \in [t_1, \infty)$*,* $\mathbf{C}_{\mathrm{x}}^{\top}$ *and* $\mathbf{C}_{\mathrm{v}}^{\top}$ *are given in* (3.1.4)-(3.1.5) *and* $\mathcal{D}_{\mathrm{c}} \subset \mathcal{V}$*, denotes the set of nodes in* c*-th connected component of* $\Delta \mathsf{L}_{\mathbf{q}}$ *corresponding to agents involved in connected switching links, and finally* $\mathcal{U}_{\mathbf{q}}$ *is a unitary matrix* ($\mathcal{U}_{\mathbf{q}}\mathcal{U}_{\mathbf{q}}^{\top} = I$) *diagonalizing Laplacian* $\mathsf{L}_{\mathbf{q}}$*.*

Then, ZDA and covert attacks undetectable for the centralized observer (3.2.11) *are impossible only if the topology switching satisfies conditions 1-3. If additionally the*

system is not at its exact consensus equilibrium when the attack is launched, conditions 1-3 are sufficient for the detection of ZDA.

Proof. See Appendix A.4.

Remark 3.2.3. ***(Safe topology switching).*** *For a given pair* $(\mathbf{A}_{\sigma(t)}, \mathbf{C})$ *in* (3.1.4)*, one can compute a set of switching modes by evaluating the conditions 1-3 of Theorem 3.2.4. This could be performed through iterative algorithms changing graph connections. Furthermore, if* $\boldsymbol{\mathcal{Z}}$ *be an unknown subspace associated with system states affected by stealthy attack* $\mathbf{u}_a(t)$ *i.e.* $\tilde{\mathbf{x}}(t) \in \boldsymbol{\mathcal{Z}}$*. Then, in view of* $\tilde{\mathbf{x}}_0 = \mathbf{x}_0 - \bar{\mathbf{x}}_0$ *(see Proposition 3.1.4), the discrepancy term* $(\mathbf{A}_{\sigma(t)} - \mathbf{A}_1)\mathbf{x}$ *in the dynamical system* (3.2.14) *will be bounded and vanishing if*

$$\boldsymbol{\mathcal{X}}_{\mathbf{q}} \cap \boldsymbol{\mathcal{Z}} = \emptyset. \tag{3.2.17}$$

Therefore, if condition (3.2.17) *holds,* $(\mathbf{A}_{\sigma(t)} - \mathbf{A}_1)\mathbf{x}$ *does not affect the stability of the system, as a consequence of input-to-state stability property of consensus systems [81]. It is also noteworthy that although identifying* $\boldsymbol{\mathcal{Z}}$ *beforehand is practically impossible as* $\mathbf{B}$ *and* $\tilde{\mathbf{x}}_0$ *in* (3.1.6) *are unknown to the defender, local observers detecting stealthy attacks in a cluster can locally identify and trigger a safe switching mode that satisfies* (3.2.17)*.*

3.2.5 Attack Detection Procedure

The results in the previous section provide conditions for the detectability of stealthy attacks locally, at the cluster level, and globally, at a ground control station equipped with a centralized observer. As described earlier, the attack detection framework relies on switching communication links generating a discrepancy between the attacker model (3.1.6) and the actual system (3.1.4). To this end, at the local level (clusters),

unknown-input observers in (3.2.5), satisfying conditions of Proposition 3.2.3, locally detect stealthy attacks. Followed by the detection, a local observer triggers a topology switching, $\mathcal{G}_{\sigma(t)}$, that satisfies conditions 1-3 of Theorem 3.2.4, yielding stealthy attack detection in the control center. This procedure is presented in Algorithm 1.

Algorithm 1 Topology switching for attack detection

procedure ATTACK DETECTION($\mathcal{G}_{\sigma(t)}$, Obs. in (3.2.11), (3.2.5))
  do run global observer (3.2.11) and local observers (3.2.5).
  if $\mathbf{r}_{\mathrm{i}_i}(t) >$ `threshold` **then**
    do Identify a safe mode $\sigma(t) = \mathbf{q} \in \mathcal{Q}$ for $\mathsf{L}_{\sigma(t)}$ that satisfies conditions 1-3 in Theorem. 3.2.4
    do Trigger an identified safe mode $\sigma(t) = \mathbf{q} \in \mathcal{Q}$
    if $\mathbf{r}_0(t) >$ `threshold` **then**
      Stealthy attack is detected.
    end if
  end if
end procedure

As presented in Algorithm 1, the observers (attack detectors) require an appropriate threshold for their residuals to avoid false attack detection. These thresholds can be designed by considering an upper bound on the estimation error of observers in the attack-free case. An analytical analysis, however, will be the subject of future work.

3.3 Simulation Results

We use a numerical example to validate the performance of the attack detection framework. We consider a network of $N = 19$ agents and investigate, in three cases,

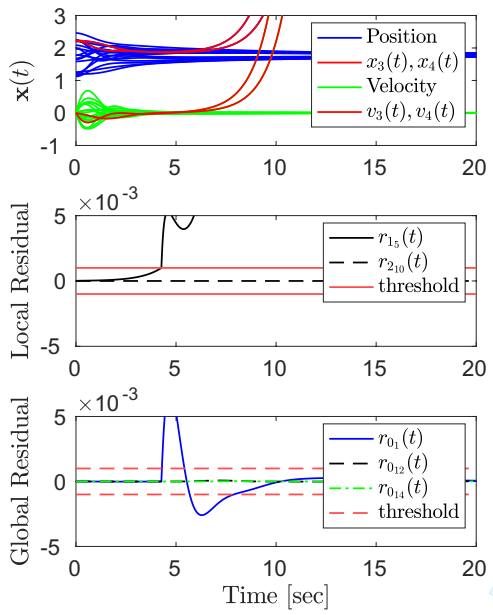
the effect conditions proposed in Proposition 3.2.3 and Theorem 3.2.4 on stealthy attack detection. It is assumed that the network has been partitioned into three clusters $\mathcal{P}_1 = \{1, \cdots, 7\}$, $\mathcal{P}_2 = \{8, \cdots, 12\}$, $\mathcal{P}_3 = \{13, \cdots, 19\}$. Each cluster is equipped with the local observer (3.2.5) (the nodes highlighted in blue in Fig. 3.2) whose local measurements are consistent with Assumption 3.2.2 and Proposition 3.2.3. More specifically, In cases 1 and 2, cluster $\mathcal{P}_1$ has two local observers that each has access to its neighboring agents' measurements. In cluster $\mathcal{P}_2$, however, we considered one local observer having more communication with other agents within the cluster for its realization (cf. Remark 3.2.2). Similar analysis is applied to case 3. Moreover, there is a centralized observer with global measurements as $\mathcal{M}_x = \emptyset$, $\mathcal{M}_v = \{1, 12, 14\}$ consistent with Lemma 3.2.1. In the simulations, the system's initial conditions are considered to be known for observers although this is not a requirement for the presented theoretical results. Also, the constant thresholds were selected by evaluating the observers' performance in different case studies.

In cases 1 and 2 (shown respectively in Figs. 3.2-(a) and 3.2-(b) with their communication topology in Fig. 3.2-(d)) a ZDA occurs in cluster $\mathcal{P}_1$ and particularly affects agents 3 and 4. As depicted, ZDA is stealthy in the global residuals $\mathbf{r}_{0_i}$'s, $i \in \{1, 12, 14\}$ before topology switching. It is, however, detectable in local residual $\mathbf{r}_{1_5}(t)$. The local control center, node 5, can trigger either of case 1's or case 2's switching topologies shown in Figs. 3.2-(d). While the conditions 1-3 of Theorem 3.2.4 are met in both cases, only case 2 meets (3.2.17) of remark 3.2.3. Consequently, the global residual $\mathbf{r}_{0_1}(t)$ for case 1 is bounded and vanishing after topology switching while that of case 2 is unbounded.

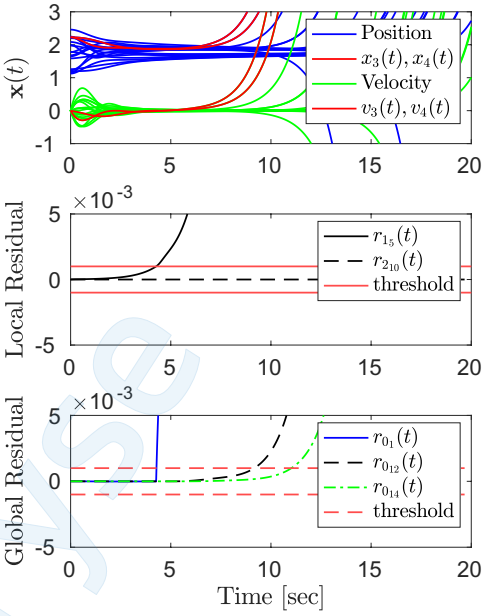
In cases 3 (shown in Fig. 3.2(c) with its communication topology in 3.2-(d)) a ZDA occurs in cluster $\mathcal{P}_2$ and particularly affects agents 11. Note that, unlike in cases 1 and 2, none of the Theorem 3.2.4's conditions are met in case 3, yielding the

global residuals $\mathbf{r}_{0_i}(t)$, $i \in \{1, 12, 14\}$ remain unaffected by the switching topology. Consequently, stealthy attack is not detectable.

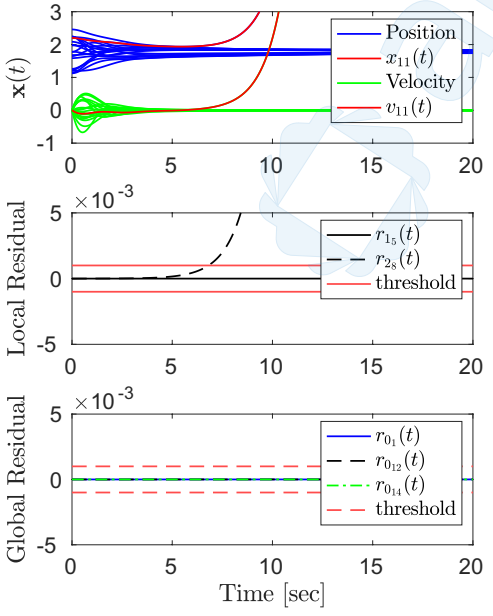
Moreover, comparing the bounded global residual in cases 1 with the unbounded global residual in case 2, suggests that meeting condition (3.2.17) presents a trade-off between a faster attack detection at a price of further exposing system states to ZDA and a slower detection by keeping uncompromised system states bounded.



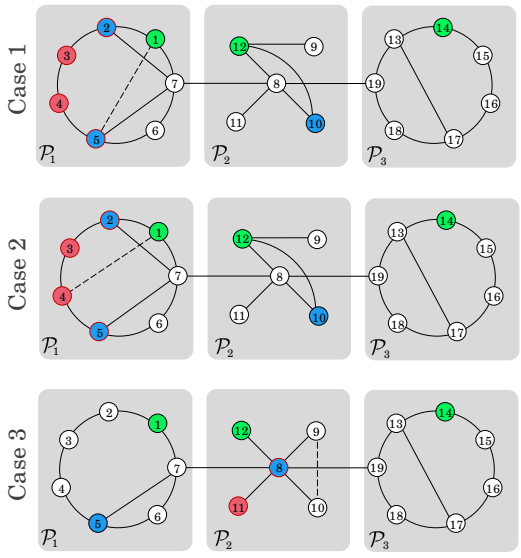
(a) Case 1: bounded residual



(b) Case 2: unbounded residual



(c) Case 3: no detection



(d) Communication topology

Figure 3.2: Simulation results of privacy-preserving stealthy attack detection for a multi-agent control system with 19 agents. [Continued on next page]

Figure 3.2: [cont'd]: The state trajectory $\mathbf{x}(t)$ consists of the agents' positions (in blue) and velocities (in green) as well as the red trajectories showing the affected agents by the stealthy attack (ZDA). (a)-(c) the results of attack detection for three cases with their respective network topology switching depicted in (d). In all cases of (d), the green nodes represent the agents globally monitored by the centralized observer, the blue nodes indicate the local control centers equipped with local observers, the red-bordered nodes show compromised agents, and the red-colored nodes represent compromised agents affected by the stealthy zero-dynamics attack (ZDA). Finally, the dashed lines (edges) represent the switching communication links. In the figures displaying local residuals, with a slight abuse of notation (cf. (3.2.9)), the scalar residual $\mathbf{r}_{\mathrm{i}_i}$ shows only the velocity estimation error of node i.

Chapter 4

Detection of Stealthy Attacks for Networked Unmanned Aerial Vehicles

This chapter[1] extends the previous chapter to the case of formation control, as a cooperative task, of unmanned aerial vehicles (UAVs). It presents model-based centralized and decentralized observer techniques for detecting a class of stealthy attacks, namely zero-dynamics and covert attacks, on networked UAVs in formation control settings. The centralized observer that runs in a control center leverages switching in the UAVs' communication topology for attack detection, and the decentralized observers, implemented onboard each UAV in the network, use the model of networked UAVs and locally available measurements. Experimental results are provided to show the effectiveness of the proposed detection schemes in different case studies.

4.1 Problem Formulation

4.1.1 Quadrotor's Dynamics

We consider a team of N homogeneous unmanned aerial vehicles (quadrotor UAVs) that cooperate to achieve a geometric shape/formation in $\mathbb{R}^2$. Attached to the center of mass of each quadrotor, the body frame $\{\mathcal{B}_i\}$ with unit axes $\{\vec{\mathfrak{b}}_1^i, \vec{\mathfrak{b}}_2^i, \vec{\mathfrak{b}}_3^i\}$, $i \in \{1, \ldots, N\} =: \mathcal{V}$, whose position and orientation with respect to the inertial global frame $\{\mathcal{W}\}$ with unit vectors $\{\vec{\mathfrak{e}}_x, \vec{\mathfrak{e}}_y, \vec{\mathfrak{e}}_z\}$ (see Fig. 4.1a) are, respectively, determined by a vector $p_i = \text{col}\left(p_i^x, p_i^y, p_i^z\right) \in \{\mathcal{W}\}$, $\forall\, i \in \mathcal{V}$ and a rotation matrix $R_i(\psi_i, \phi_i, \theta_i) \in \text{SO(3)}$ in the special orthogonal group with ψ_i, ϕ_i, and θ_i being the respective $z{-}x{-}y$

[1]This chapter is adapted from a publication by the author of this dissertation.

Euler angles. Then, the rigid body motion of the quadrotors follows [76]

$$\dot{p}_i = v_i, \qquad m\dot{v}_i = -mg\vec{\mathfrak{e}}_z + R_i f_i, \tag{4.1.1a}$$

$$\dot{R}_i = R_i \Omega_i^{\times}, \qquad J\dot{\Omega}_i = -\Omega_i \times J\Omega_i + \boldsymbol{\tau}_i, \tag{4.1.1b}$$

where $p_i \in \mathbb{R}^3$ and $R_i(\psi_i, \phi_i, \theta_i) \in \mathrm{SO}(3)$ are the position and orientation of the i-th quadrotor in the inertial frame $\{\mathcal{W}\}$, m is the mass of the quadrotor, g is the gravitational acceleration, and finally $f_i \in \{\mathcal{B}_i\}$ is the total thrust. Also, in the rotational dynamics, $\Omega_i \in \mathbb{R}^3$ is the angular velocity, $J \in \mathbb{R}^{3\times 3}$ is the inertia matrix, and $\boldsymbol{\tau}_i \in \mathbb{R}^3$ is the total torque, all expressed in respective body-fixed frames. Finally, the notation $\Omega_i^{\times}$ denotes the skew-symmetric matrix, such that $\Omega_i^{\times} r = \Omega_i \times r$ for any vector $r \in \mathbb{R}^3$ and the cross product $\times$.

4.1.2 Formation Control

The cooperative control of quadrotor UAVs, shown in Fig. 4.1b, follows a hierarchical structure, where at the high level, the UAVs coordinate with each other and their formation/position controller cooperatively generates the desired attitude/orientation and the desired total thrust for a low-level attitude controller. In this chapter, we focus on 2D formation in the $x-y$ plane, for which cooperative control protocols will be designed based on a linearized model of the UAVs' transnational dynamics in (4.1.1a) around a hovering state and under small-angle approximations as follows

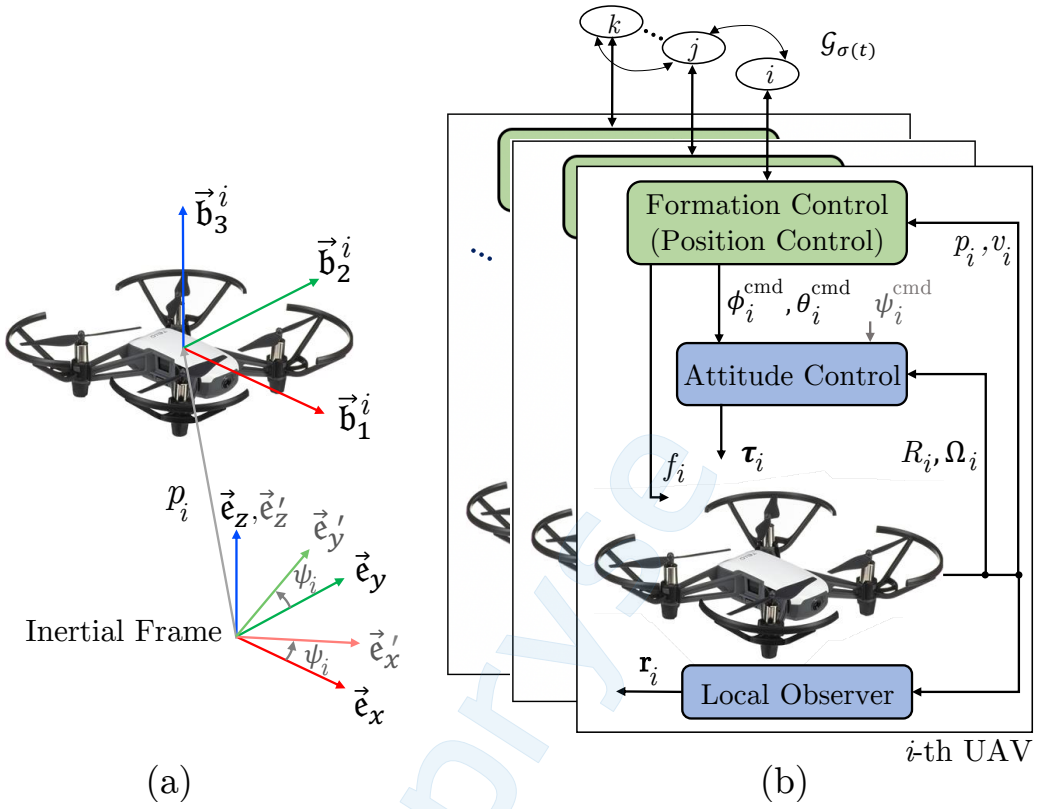


Figure 4.1: (a) Illustration of reference frames. (b) The coordination control architecture.

[76]:

$$\ddot{p}_i^x = g(\Delta\theta_i \cos(\psi_i) + \Delta\phi_i \sin(\psi_i)), \tag{4.1.2a}$$

$$\ddot{p}_i^y = g(\Delta\theta_i \sin(\psi_i) - \Delta\phi_i \cos(\psi_i)), \tag{4.1.2b}$$

$$\ddot{p}_i^z = -g + f_i/m, \tag{4.1.2c}$$

in which $\Delta\theta_i$ and $\Delta\phi_i$ denote, respectively, the deviation of pitch and roll angles of the i-th quadrotor from their equilibrium point $\theta_i = \phi_i = 0$. Associated with each UAV, we define an intermediary frame $\{\mathcal{I}'\}$ with unit axes $\{\vec{\mathfrak{e}}_x', \vec{\mathfrak{e}}_y', \vec{\mathfrak{e}}_z'\}$ and orientation $R_z(\psi_i)$ such that $p_i = R_z(\psi_i)p_i'$ for vectors $p_i \in \{\mathcal{W}\}$ and $p_i' \in \{\mathcal{I}'\}$ (see Fig. 4.1a). Assuming all UAVs have consensus on a desired yaw angle $\psi_i = \psi^*$, $\forall\, i \in \mathcal{V}$, $\{\mathcal{I}'\}$

will be the common reference frame of the UAVs in which the linearized equation of motion in (4.1.2) can be represented by

$$\ddot{p}_i^{x'} = +g\Delta\theta_i, \tag{4.1.3a}$$

$$\ddot{p}_i^{y'} = -g\Delta\phi_i, \tag{4.1.3b}$$

$$\ddot{p}_i^{z'} = -g + f_i/m, \tag{4.1.3c}$$

that shows the decoupled dynamics in the x_i', y_i', z_i' directions[2]. We let the reference commands for pitch and roll angles in (4.1.3a)-(4.1.3b) be

$$\theta_i^{cmd} = +u_i^x/g, \qquad \phi_i^{cmd} = -u_i^y/g, \tag{4.1.4}$$

where u_i^x and u_i^y are the formation control inputs to be designed, respectively, in the x and y directions. It is also necessary to mention that θ_i^{cmd} and ϕ_i^{cmd} in (4.1.4) will be desired setpoints for each UAV's low-level (on-board) attitude controller, and that we use independent PID controllers to stabilize the altitude of quadrotors (z_i-dynamics in (4.1.2c)) around a desired hovering point. Therefore, the altitude dynamics in (4.1.2c) and the rotational dynamics in (4.1.1b) are dropped from the high-level state space of networked UAVs and the reduced-order planar dynamics is obtained by substituting (4.1.4) for $\Delta\theta_i$ and $\Delta\phi_i$ in (4.1.3a) and (4.1.3b) as follows:

$$\Sigma_i : \begin{cases} \dot{\boldsymbol{p}}_i(t) = \boldsymbol{v}_i(t) \\ \dot{\boldsymbol{v}}_i(t) = \boldsymbol{u}_i(t) \end{cases}, \qquad i \in \mathcal{V} = \{1, \ldots, N\}, \tag{4.1.5}$$

in which $\boldsymbol{p}_i(t) \coloneqq \mathrm{col}(p_i^x, p_i^y) \in \mathbb{R}^2$, and $\boldsymbol{v}_i(t) \coloneqq \mathrm{col}(\dot{p}_i^x, \dot{p}_i^x) \in \mathbb{R}^2$ are the stacked positions and velocities in the x and y directions, and $\boldsymbol{u}_i(t) \coloneqq \mathrm{col}(u_i^x, u_i^y) \in \mathbb{R}^2$

[2]We will omit the superscript $'$ in the rest of chapter for notational simplicity.

denotes their corresponding control input for each UAV.

Desired formation reference. We define a desired configuration (formation shape) by specifying a set of N desired setpoints $\boldsymbol{p}_1^\star, \boldsymbol{p}_2^\star, \ldots, \boldsymbol{p}_N^\star$ in $\mathbb{R}^2$ that form the desired relative positions[3] $\{\boldsymbol{p}_{ij}^\star = \boldsymbol{p}_i^\star - \boldsymbol{p}_j^\star \in \mathbb{R}^2 \mid \forall\, i, j \in \mathcal{V},\ i \neq j\}$, all expressed in the UAVs' common frame. The formation references are transmitted to the UAVs from a ground control center. We follow the consensus-based formation settings [109] where the UAVs coordinate their relative positions to reach the desired relative positions $\boldsymbol{p}_{ij}^\star$'s, which is formulated as

$$\lim_{t\to\infty} \left|\boldsymbol{p}_i(t) - \boldsymbol{p}_j(t) - \boldsymbol{p}_{ij}^\star\right| = \mathbf{0}, \qquad \forall\, i, j \in \mathcal{V}, \tag{4.1.6a}$$

$$\lim_{t\to\infty} |\boldsymbol{v}_i(t)| = \mathbf{0}, \qquad \forall\, i \in \mathcal{V}. \tag{4.1.6b}$$

Inter-UAV communication. We model the switching inter-UAV communication by an undirected graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$, where the vertex set $\mathcal{V} = \{1, \ldots, N\}$ represents the index set of N UAVs in (4.1.1) (with their respective reduced models in (4.1.5)), and the edge set $\mathcal{E}_{\sigma(t)} \subset \mathcal{V} \times \mathcal{V}$ represents the communication links such that an edge $(i, j) \in \mathcal{E}_{\sigma(t)}$ implies information exchange between the i-th and j-th UAV in a given active mode determined by the right-continuous switching signal $\sigma(t) : \mathbb{R}_{\geq 0} \to \mathcal{Q} := \{1, 2, \ldots, \mathbf{q}\}$, $\mathbf{q} \in \mathbb{N}$, at time t, with $\mathcal{Q}$ being the finite index set of possible communication graphs.

Assumption 4.1.1. *Throughout this chapter, we assume the inter-UAV's communication graphs $\mathcal{G}_{\sigma(t)}$'s are connected in all modes $\sigma(t) \in \mathcal{Q}$.*

To meet the formation constraints in (4.1.6), we use the following consensus-

[3] In the context of formation control, these reference states are called *formation states* [109] or *shape vectors* [134] depending on the design methods and their underlying assumptions.

based distributed control protocol:

$$\boldsymbol{u}_i = \boldsymbol{u}_{\mathrm{n}_i} + \boldsymbol{u}_{\mathrm{a}_i}, \quad i \in \mathcal{V}, \tag{4.1.7}$$

$$\boldsymbol{u}_{\mathrm{n}_i} = -\alpha \sum_{j \in \mathcal{N}_\sigma^{i(1)}} a_{ij}^{\sigma(t)} (\boldsymbol{p}_i - \boldsymbol{p}_j - \boldsymbol{p}_{ij}^\star) - \gamma \boldsymbol{v}_i \tag{4.1.8}$$

where $\boldsymbol{u}_{\mathrm{n}_i} \in \mathbb{R}^2$ denotes the nominal control input with $a_{ij}^{\sigma(t)}$ being the entry of the symmetric adjacency matrix associated with the UAVs' switching communication graph $\mathcal{G}_{\sigma(t)}$. Also, $\alpha \in \mathbb{R}_{>0}$ and $\gamma \in \mathbb{R}_{>0}$ are the control gains. $\boldsymbol{u}_{\mathrm{a}_i} \in \mathbb{R}^2$ is the vector-valued malicious signal injected in the control channel of the i-th UAV.

We let an unknown subset $\mathcal{A} = \{i_1, i_2, \dots\} \subset \mathcal{V}$ denote the set of UAVs subject to attack $\boldsymbol{u}_{\mathrm{a}_i} \neq \mathbf{0}$, which we refer to as compromised UAVs, and we refer to the rest of UAVs Σ_i's with $\boldsymbol{u}_{\mathrm{a}_i} = \mathbf{0}$, $\forall\, i \in \mathcal{V} \setminus \mathcal{A}$, in (4.1.5) as uncompromised UAVs.

Network-level dynamics. Given (4.1.5), (4.1.7) and (4.1.8), the dynamics of the networked UAVs can be represented by

$$\Sigma : \dot{\mathbf{x}} = \mathbf{A}_{\sigma(t)} \mathbf{x} + \mathbf{B}_{\sigma(t)}^{\boldsymbol{F}} \mathbf{x}^* + \mathbf{B}_{\mathcal{A}} \mathbf{u}_{\mathcal{A}}, \qquad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{4.1.9}$$

in which, the system states and matrices are given by

$$\mathbf{x}(t) = \mathrm{col}\,(\boldsymbol{p}_1, \dots, \boldsymbol{p}_N, \boldsymbol{v}_1, \dots, \boldsymbol{v}_N) \in \mathbb{R}^{4N}, \tag{4.1.10a}$$

$$\mathbf{x}^* = \mathrm{col}\,(\boldsymbol{p}_1^\star, \dots, \boldsymbol{p}_N^\star, \mathbf{0}_2, \dots, \mathbf{0}_2) \in \mathbb{R}^{4N}, \tag{4.1.10b}$$

$$\mathbf{A}_{\sigma(t)} = A_{\sigma(t)} \otimes I_2, \qquad \mathbf{B}_{\sigma(t)}^{\boldsymbol{F}} = -\mathbf{A}_{\sigma(t)}, \qquad \mathbf{B}_{\mathcal{A}} = B_{\mathcal{A}} \otimes I_2, \tag{4.1.10c}$$

$$A_{\sigma(t)} = \begin{bmatrix} 0_{N\times N} & I_N \\ -\alpha \mathsf{L}_{\sigma(t)} & -\gamma I_N \end{bmatrix}, \; B_{\mathcal{A}} = \begin{bmatrix} 0 \\ \mathrm{B}_{\mathcal{A}} \end{bmatrix}, \tag{4.1.10d}$$

$$\mathrm{B}_{\mathcal{A}} = \begin{bmatrix} \mathfrak{e}_{i_1} & \mathfrak{e}_{i_2} & \dots & \mathfrak{e}_{i_{|\mathcal{A}|}} \end{bmatrix}, \quad \mathbf{u}_{\mathcal{A}} = \mathrm{col}\,(\boldsymbol{u}_{\mathrm{a}_i})_{i \in \mathcal{A}}\,, \tag{4.1.10e}$$

where $\mathsf{L}_{\sigma(t)}$ is the Laplacian matrix of graph $\mathcal{G}_{\sigma(t)}$, encoding the inter-UAVs' communication links, defined as $\mathsf{L}_{\sigma(t)} \coloneqq [l_{ij}^{\sigma(t)}] \in \mathbb{R}^{N\times N}$ with $l_{ii}^{\sigma(t)} = \sum_{j\neq i} a_{ij}^{\sigma(t)}$ and $l_{ij}^{\sigma(t)} = -a_{ij}^{\sigma(t)}$ if $i \neq j$. $\mathfrak{e}_i$ is the i-th vector of the canonical basis in $\mathbb{R}^N$ corresponding to the i-th UAV compromised by attack $\boldsymbol{u}_{\mathrm{a}_i},\ i \in \mathcal{A}$.

Network-level measurements. We define the system measurements $\mathbf{y}$ to be composed of the position of a set of UAVs, indexed by $\mathcal{M}_{\mathrm{p}} = \{\mathrm{p}_1, \mathrm{p}_2, \dots\} \subset \mathcal{V}$, and/or the velocity of a set of UAVs, indexed by $\mathcal{M}_{\mathrm{v}} = \{\mathrm{v}_1, \mathrm{v}_2, \dots\} \subset \mathcal{V}$, that are transmitted to a ground control center for monitoring. More precisely,

$$\mathbf{y} = \mathbf{C}\mathbf{x} - \mathbf{u}_{\mathcal{S}}, \quad \mathbf{C} = C \otimes I_2, \quad \mathcal{M} = \{\mathcal{M}_{\mathrm{p}}, \mathcal{M}_{\mathrm{v}}\}, \tag{4.1.11a}$$

$$C = \operatorname{diag}\left(C_{\mathrm{p}}, C_{\mathrm{v}}\right), \tag{4.1.11b}$$

$$C_{\mathrm{p}} = \operatorname{col}\left(\mathfrak{e}_{\mathrm{p}_1}^\top, \mathfrak{e}_{\mathrm{p}_2}^\top, \dots, \mathfrak{e}_{\mathrm{p}_{|\mathcal{M}_{\mathrm{p}}|}}^\top\right) \in \mathbb{R}^{|\mathcal{M}_{\mathrm{p}}|\times N}, \tag{4.1.11c}$$

$$C_{\mathrm{v}} = \operatorname{col}\left(\mathfrak{e}_{\mathrm{v}_1}^\top, \mathfrak{e}_{\mathrm{v}_2}^\top, \dots, \mathfrak{e}_{\mathrm{v}_{|\mathcal{M}_{\mathrm{v}}|}}^\top\right) \in \mathbb{R}^{|\mathcal{M}_{\mathrm{v}}|\times N}, \tag{4.1.11d}$$

where $\mathbf{u}_{\mathcal{S}} = \operatorname{col}\left(\mathbf{u}_{\mathrm{s}_1}, \mathbf{u}_{\mathrm{s}_2}, \dots, \mathbf{u}_{\mathrm{s}_{|\mathcal{M}|}}\right) \in \mathbb{R}^{2|\mathcal{M}|}$ denotes the vector-valued sensory attacks on the measurements.

Proposition 4.1.2. ***(Formation convergence).*** *Assume that the formation configuration is feasible and that the communication graph is connected in each mode. Then, under the control protocol* (4.1.7)*, and in the absence of attacks, the states of the UAVs in* (4.1.5) *converge to the desired formation configuration in* (4.1.6)*.*

Proof. The proof follows a change of variables as in [106] and a convergence analysis similar to that in [78]. A more comprehensive proof follows from the proof of Proposition 5.2.6 of this dissertation.

Note that the UAV's dynamics in (4.1.5) as well as the control protocol (4.1.8) for the x and y directions are decoupled. Thus, for notational simplicity, we may use

the following

$$\Sigma : \dot{\mathrm{x}} = A_{\sigma(t)}\mathrm{x} + B^{\boldsymbol{F}}_{\sigma(t)}\mathrm{x}^{*} + B_{\mathcal{A}}\mathrm{u}_{\mathcal{A}}, \quad \mathrm{x}(t_0) = \mathrm{x}_0, \tag{4.1.12a}$$

$$\mathrm{y} = C\mathrm{x} - \mathrm{u}_{\mathcal{S}}, \tag{4.1.12b}$$

to represent the dynamics in (4.1.9) with its monitored states in (4.1.11b) in only one direction of the $x-y$ plane. Accordingly, $\mathrm{x} = \mathrm{col}\,(\mathrm{p}, \mathrm{v}) \in \mathbb{R}^{2N}$ in (4.1.12) denotes the stacked vector of all positions $\mathrm{p} = \mathrm{col}\,(\mathrm{p}_i)_{i\in\mathcal{V}}$ and velocities $\mathrm{v} = \mathrm{col}\,(\mathrm{v}_i)_{i\in\mathcal{V}}$ in one direction with their corresponding formation references $\mathrm{x}^{*} \in \mathbb{R}^{2N}$ as well as attack inputs $\mathrm{u}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and $\mathrm{u}_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{M}|}$, and other system matrices are given in (4.1.10d)-(4.1.10e) and (4.1.11b) with $B^{\boldsymbol{F}}_{\sigma(t)} = -A_{\sigma(t)}$.

4.1.3 Attack Stealthiness

We consider the worst-case scenario adversarial settings where an attacker leverages *a priori* system knowledge of the UAVs' coordination or a prerecorded sequence of sensory data to design sophisticated stealthy attacks implementable through actuator attacks $\boldsymbol{u}_{\mathrm{a}_i}(t)$'s, $i \in \mathcal{A}$ in (4.1.7) and sensor attacks $\mathbf{u}_{\mathcal{S}}(t)$ in (4.1.11a).

Here, *a priori* system knowledge refers to the initial configuration of the networked system (4.1.9) with the measurements (4.1.11) (or equivalently (4.1.12)), denoted by the tuple $\hat{\Sigma}(\hat{\mathbf{A}}_{\sigma(t)}, \hat{\mathbf{C}}, \sigma(t) = 1)$ with $\hat{\mathbf{A}}_{\sigma(t)}$ and $\hat{\mathbf{C}}$ being the approximations of their counterparts in (4.1.9) and (4.1.11). The amount of *a priori* system knowledge needed for designing *stealthy* attacks varies for different attacks [131], and will be quantified in Section 4.2.1.

Stealthy attacks refer to a class of adversarial attacks (cyber attacks [124, 82]) $\boldsymbol{u}_{\mathrm{a}_i}$'s, $i \in \mathcal{A}$ in (4.1.7) and $\mathbf{u}_{\mathcal{S}}$ in (4.1.11a) that disrupt the system's normal operation

while remaining stealthy in the monitored measurements (4.1.11), that is (cf. (3.1.7))

$$\mathbf{y}(t; \mathbf{x}_0, \mathbf{u}_{\mathcal{A}}, \mathbf{u}_{\mathcal{S}}) = \mathbf{y}^{\mathrm{n}}(t; \mathbf{x}_0^{\mathrm{n}}, \mathbf{0}, \mathbf{0}), \quad \forall\, t \in [t_0,\, t_d), \tag{4.1.13}$$

where $\mathbf{y}^{\mathrm{n}}(t; \mathbf{x}_0^{\mathrm{n}}, \mathbf{0}, \mathbf{0}) = \mathbf{C}\mathbf{x}^{\mathrm{n}}$ is the output associated with an attack-free system with the same dynamics as in (4.1.12a), and $\mathbf{x}_0$ and $\mathbf{x}_0^{\mathrm{n}}$ are the actual and a possible initial states, respectively. Also, t_0 is the initial time instant, and t_d is the attack detection time instant, i.e., the time instant at which condition in (4.1.13) no longer holds and attacks lose their stealthiness.

4.1.4 Problem Statement: Attack Detection

We consider the attack detection problem as a hypothesis testing problem with the null and alternative hypotheses

$$\mathcal{H}^{\mathrm{o}} : \text{attack-free}, \qquad \text{vs.} \qquad \mathcal{H}^{\mathrm{1}} : \text{attacked}, \tag{4.1.14}$$

for which we present detection frameworks in Section 4.2.2.

4.2 Observer Design and Analysis for Attack Detection

In this section, we characterize the models for stealthy attacks on the networked UAVs in (4.1.9) and develop centralized and decentralized detection schemes.

4.2.1 Realization of Stealthy Attacks

Given system in (4.1.9), let $\mathcal{M}$ in (4.1.11a) be a set of monitored states and let $\mathcal{A}$ be a set of compromised UAVs subject to attack $\boldsymbol{u}_{\mathrm{a}_i} \neq \mathbf{0}$ in (4.1.7). In what follows, we characterize stealthy attacks in terms of different realizations of (4.1.13).

Zero-dynamics attack (ZDA). ZDA refers to the class of attacks based on the zero dynamics of the system $(A_{\sigma(t)}, B_{\mathcal{A}}, C, \sigma(t) = 1)$ in (4.1.12) that are (nontrivial) state trajectories excited through input directions $B_{\mathcal{A}}$ and invisible at the output y, and that can be characterized by the rank deficiency of matrix pencil

$$P(\lambda_{\mathrm{o}}) = \begin{bmatrix} \lambda_{\mathrm{o}} I_N - A_1 & -B_{\mathcal{A}} \\ C & \mathbf{0} \end{bmatrix}, \tag{4.2.1}$$

for some $\lambda_{\mathrm{o}} \in \mathbb{R}_{>0}$ [78].

Proposition 4.2.1. *Assume the system in* (4.1.9) *in its initial active mode* $\sigma(t) = 1$ *has unstable zero dynamics, i.e., the matrix pencil* $P(\lambda_{\mathrm{o}})$ *in* (4.2.1) *is rank deficient for some* $\lambda_{\mathrm{o}}^x,\ \lambda_{\mathrm{o}}^y, \in \mathbb{R}_{>0}$, *and that the attacker's a priori knowledge of the system* $\hat{\Sigma}(\hat{\mathbf{A}}_{\sigma(t)}, \hat{\mathbf{C}}, \sigma(t) = 1) = (\mathbf{A}_{\sigma(t)}, \mathbf{C}, \sigma(t) = 1)$. *Then, there exists a stealthy attack policy*

$$\mathbf{u}_{\mathcal{A}} = \mathrm{col}\,(\boldsymbol{u}_{\mathrm{a}_i})_{i\in\mathcal{A}},\ \boldsymbol{u}_{\mathrm{a}_i} = [\mathrm{u}_{\mathrm{a}_i}^x(0)e^{\lambda_{\mathrm{o}}^x t}\ \ \mathrm{u}_{\mathrm{a}_i}^y(0)e^{\lambda_{\mathrm{o}}^y t}\ ]^\top, \tag{4.2.2}$$

in dynamics (4.1.9) *that causes part of system states exponentially deviate from the formation configuration in* (4.1.6) *while the condition in* (4.1.13) *holds. In this attack model, the measurement signals are not compromised, i.e.,* $\mathbf{u}_{\mathcal{S}} = \mathbf{0}$.

Proof. The proof follows from Proposition 3.1.4 and so is omitted here.

It is noteworthy that the assumption on *a priori* system knowledge in Proposition 4.2.1 can be relaxed. In the cases that only a subset of the system model as *a priori* is disclosed to the attacker, that is $\hat{\Sigma}(\hat{\mathbf{A}}_{\sigma(t)}, \hat{\mathbf{C}}, \sigma(t) = 1) \approx (\mathbf{A}_{\sigma(t)}, \mathbf{C}, \sigma(t) = 1)$, a ZDA can be realized that only affects the UAVs within the known subset of the system, which is known as local ZDA [131].

Covert attack [124]. Covert attacks are a class of intrusions through input channels $\mathbf{B}_{\mathcal{A}}$ whose covertness at the output is obtained by alteration of the measurement signals (4.1.11a) and whose realization requires perfect knowledge of the system i.e. $\hat{\Sigma}(\hat{\mathbf{A}}_{\sigma(t)}, \hat{\mathbf{C}}, \sigma(t) = 1) = (\mathbf{A}_{\sigma(t)}, \mathbf{C}, \sigma(t) = 1)$. Let attack policy $\mathbf{u}_{\mathcal{A}}(t) = \mathrm{col}\,(\boldsymbol{u}_{\mathrm{a}_i})_{i\in\mathcal{A}} : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}^2$ in (4.1.9) be any continuous signal initiated at time instant $t_{\mathrm{a}} \in \mathbb{R}_{\geq 0}$. Then, the attack is covert and (4.1.13) holds if the attacker alters the measurement (4.1.11a) by

$$\mathbf{u}_{\mathcal{S}}(t) = \mathbf{C}\int_{t_{\mathrm{a}}}^{t} e^{\mathbf{A}_1(t-\tau)}\mathbf{B}_{\mathcal{A}}\mathbf{u}_{\mathcal{A}}(\tau)d\tau. \tag{4.2.3}$$

We refer to Chapter 3 for the details of the derivation and proof.

Cooperative DoS and replay attack. It is shown that a denial-of-service (DoS), interfering in a UAV's communication, causes unstable and unsafe flights [25]. We formulate a scenario where replay attacks[4] [82] are implemented in cooperation with a DoS in order to keep the DoS stealthy in the networked-level measurements (4.1.11a). Here, the cooperatively-stealthy DoS and replay attack takes place when the UAVs have reached the formation configuration in (4.1.6) and thus are hovering only, giving the attacker the opportunity to record and store slow-varying measurements (4.1.11a) for a time interval $T_r \in \mathbb{R}_{>0}$ before starting the attacks $\mathbf{u}_{\mathcal{A}}$ and $\mathbf{u}_{\mathcal{S}}$ respectively in (4.1.9) and (4.1.11a) that is $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{0}$ and $\mathbf{u}_{\mathcal{S}}(t) = \mathbf{0}$, $\forall\, t \in [t_0\ t_{\mathrm{a}})$ where $t_{\mathrm{a}} > T_r$. Then, upon starting a DoS at a time instant $t_{\mathrm{a}} \in \mathbb{R}_{>0}$, causing one or more UAVs to lose their inter-communication and deviate from the equilibrium states (4.1.6), a

[4] A replay attack is the case that the attacker replays (periodically) a sequence of stored data as real-time measurements to conceal any deviation from a normal operation.

concurrent replay attack $\mathbf{u}_{\mathcal{S}}$ in (4.1.11a) given by

$$\mathbf{u}_{\mathcal{S}}(t) = \mathbf{C}\mathbf{x}(t) - \mathbf{y}(t - nT_r),\ n \in \mathbb{N},\quad t \geq t_{\mathrm{a}}, \tag{4.2.4}$$

causes the stealthiness condition in (4.1.13) holds.

We note that *a priori* system knowledge is not required for the cooperative DoS and replay attack that is $\hat{\Sigma}(\hat{\mathbf{A}}_{\sigma(t)}, \hat{\mathbf{C}}, \sigma(t)) = \emptyset$.

4.2.2 Observer-based Detection Framework

We present centralized and decentralized detection schemes to address the attack detection problem formulated as in (4.1.14).

Centralized detection scheme. In the centralized detection scheme, we leverage switching links in the inter-UAVs' communication topology to generate model discrepancy rendering the stealthy attacks detectable in the measurements (4.1.11) monitored in a ground control center. Note that the UAVs' communication may be subject to switching connections in two senses. First, a communication link failure induced due to operation in uncertain environments, and second, a planned switch (addition or removal of connections) triggered for security and performance reasons. Regardless of the underlying causes of switching links in the inter-UAVs' communication, we investigate their effect on the detection of stealthy attacks.

Consider the dynamical system in (4.1.12), a centralized attack detection monitor (central monitor), derived based on the initial (normal) communication mode of UAVs ($\sigma(t) = 1$), is given by

$$\Sigma_{\mathcal{O}}^{\mathcal{M}}: \begin{cases} \dot{\hat{\mathrm{x}}} = A_{\sigma(t)}\hat{\mathrm{x}} + B_{\sigma(t)}^{\boldsymbol{F}}\mathrm{x}^{*} + H(\mathrm{y} - \hat{\mathrm{y}}), & \sigma(t) = 1, \\ \hat{\mathrm{y}} = C\hat{\mathrm{x}}, & \hat{\mathrm{x}}(0) = \mathbf{0}, \\ \mathrm{r}_0 = \mathrm{y} - \hat{\mathrm{y}}, & \text{central residual}, \end{cases} \tag{4.2.5}$$

where H is an observer gain such that $(A_{\sigma(t)} - HC)$ is stable in all modes and $\lim_{t\to\infty} \mathrm{r}_0 = \mathbf{0}$ in the absence of attacks (See Section 3.2.4)). Also, we let $\mathrm{r}_0^j(t) = C^j(\mathrm{x} - \hat{\mathrm{x}})$ denote the j-th component of the residual r_0 with C^j being the j-th row vector of matrix C in (4.1.11b). Then, in the absence of attacks, an upper bound on the residuals is obtained as follows:

$$|\mathrm{r}_0^j(t)| \leq \bar{k}_j e^{-\bar{\lambda}_j t}\bar{\omega} + \epsilon_0 =: \epsilon_0^j, \tag{4.2.6}$$

where $\bar{k}_j$ and $\bar{\lambda}_j$ are positive constants such that $|C^j e^{(A_1 - HC)t}| \leq \bar{k}_j e^{-\bar{\lambda}_j t}$, $\bar{\omega}$ is an upper bound such that $|\mathrm{e}(0)| = |\mathrm{x}(0) - \hat{\mathrm{x}}(0)| = |\mathrm{x}(0)| \leq \bar{\omega}$, and $\epsilon_0 \in \mathbb{R}_{>0}$ is a sufficiently small constant to account for measurement noises.

Given the central monitor (4.2.5) and its corresponding thresholds in (4.2.6), the hypothesis testing problem in (4.1.14) can be quantified either by

$$\mathcal{H}^{0} : \textit{attack-free}, \qquad \text{if} \qquad |\mathrm{r}_0^j(t)| \leq \epsilon_0^j, \quad \forall\, j \in \mathcal{M}, \tag{4.2.7a}$$

$$\mathcal{H}^{1} : \textit{attacked}, \qquad \text{if} \qquad |\mathrm{r}_0^j(t)| > \epsilon_0^j, \quad \exists\, j \in \mathcal{M}, \tag{4.2.7b}$$

or by

$$\mathrm{r}_0^{\top}\boldsymbol{\Sigma}_{\mathrm{r}_0}^{-1}\mathrm{r}_0 \underset{\mathcal{H}^{1}}{\overset{\mathcal{H}^{0}}{\lessgtr}} \texttt{threshold}, \tag{4.2.8}$$

with $\boldsymbol{\Sigma}_{\mathrm{r}_0}$ being the covariance of the residual r_0 having a zero-mean Gaussian distribution in stochastic settings where a discretized version of (4.2.5) as a Kalman filter

is used, together with χ^2 (chi-squared) tests, for attack detection [82].

As shown in [82] and here in Section 3.2.4, χ^2 detectors, Kalman filters, and Luenberger-type observers/monitors fail in detecting the stealthy attacks that were defined in Section 4.2.1 provided the stealthiness condition (4.1.13) holds, causing a false validation of the null hypothesis (4.2.7a). Here, based on the results in Theorem 3.2.4 we evaluate the effect of switching connections in inter-UAVs' communication on the violation of (4.1.13) and thus on the validation of the null hypothesis (4.2.7a). This procedure will be presented in Algorithm 3 in Section 4.3.

Decentralized detection scheme. In the decentralized detection scheme, a set of UAVs, equipped with on-board (local) monitors, leverage the information exchange with their neighboring UAVs to locally detect the stealthy attacks on their neighbors. Upon attack detection, a local monitor triggers an inter-UAV communication switch and informs other local monitors as part of a contingency plan (see Algorithm 3).

Note that in the networked UAVs with a connected communication graph $\mathcal{G}_{\sigma(t)}$, any UAV has access to the states of itself as well as the position states of the set of its immediate neighbors $\mathcal{N}_\sigma^{i(1)}$ (cf. control protocol (4.1.8)). Accordingly, we define, for the i-th UAV in the network, a set of local measurements, indexed by set $\mathcal{M}^i$, as follows:

$$\mathcal{M}^i = \mathcal{N}_\sigma^{i(1)} \cup \{i\}, \qquad \sigma(t) = 1 \in \mathcal{Q}, \tag{4.2.9a}$$

$$\mathbf{y}_i = \mathbf{C}_i \mathbf{x}, \text{ and } \mathrm{y}_i = C_i \mathrm{x}, \qquad \mathbf{C}_i = C_i \otimes I_2, \tag{4.2.9b}$$

$$C_i = \mathrm{diag}\left(\mathrm{C}_{\mathrm{p},i}, \mathfrak{e}_i^\top\right), \qquad \mathrm{C}_{\mathrm{p},i} = \mathrm{col}\left(\mathfrak{e}_j^\top\right)_{j \in \mathcal{M}^i}. \tag{4.2.9c}$$

where $\mathbf{x}$ and x are the system states in (4.1.9) and (4.1.12), respectively. Different from the networked measurements in (4.1.11), the local measurements in (4.2.9) are not transmitted through compromised network channels to the control center for

monitoring. Instead, they are locally available for each UAV and thus are not subject to alterations by sensory attacks.

Given the local measurements (4.2.9) and dynamics (4.1.12), we define the local attack detector $\Sigma_{\mathcal{O}}^{i}$ for the i-th UAV as follows:

$$\Sigma_{\mathcal{O}}^{i}: \begin{cases} \dot{\hat{\mathrm{x}}}_i = A_{\sigma(t)}\hat{\mathrm{x}}_i + B_{\sigma(t)}^{\boldsymbol{F}}\mathrm{x}^{\boldsymbol{*}} + H^i(\mathrm{y}_i - \hat{\mathrm{y}}_i), & \sigma(t) \in \mathcal{Q}, \\ \hat{\mathrm{y}}_i = C_i\hat{\mathrm{x}}, & \hat{\mathrm{x}}(0) = \mathbf{0}, \\ \mathrm{r}_i = \mathrm{y}_i - \hat{\mathrm{y}}_i, & \text{local residual}, \end{cases} \tag{4.2.10}$$

where $\hat{\mathrm{x}}_i$ is the local estimation of x in (4.1.12), and H^i is an observer gain such that $(A_{\sigma(t)} - H^iC_i)$ is stable in all modes. Therefore, in the absence of attacks, $\lim_{t\to\infty}\mathrm{r}_i = \mathbf{0}$, and similar to the central monitor's, the j-th component of local residuals, r_i^j's, hold an upper bound (threshold) as follows:

$$|\mathrm{r}_i^j(t)| \leq \bar{k}_{i,j}e^{-\bar{\lambda}_{i,j}t}\bar{\omega} + \epsilon_i =: \epsilon_i^j, \tag{4.2.11}$$

where $\bar{k}_{i,j}$ and $\bar{\lambda}_{i,j}$ are positive constants such that $|C_i^j e^{(A_1 - H^iC_i^j)t}| \leq \bar{k}_{i,j}e^{-\bar{\lambda}_{i,j}t}$, $\bar{\omega}$ is an upper bound such that $|\mathrm{e}_i(0)| = |\mathrm{x}(0) - \hat{\mathrm{x}}_i(0)| = |\mathrm{x}(0)| \leq \bar{\omega}$, and $\epsilon_i \in \mathbb{R}_{>0}$ is a sufficiently small constant to account for measurement noises.

Now, the hypothesis testing problem in (4.1.14) can be revisited and quantified using local residuals as follows:

$$\mathcal{H}^{\mathbf{0}}: \text{attack-free}, \qquad \text{if} \quad |\mathrm{r}_i^j(t)| \leq \epsilon_i^j, \ \ \forall\, j \in \mathcal{M}^i, \ \ \forall\, i \in \mathcal{D}, \tag{4.2.12a}$$

$$\mathcal{H}^{\mathbf{1}}: \text{attacked}, \qquad \text{if} \quad |\mathrm{r}_i^j(t)| > \epsilon_i^j, \ \ \exists\, j \in \mathcal{M}^i, \ \ \exists\, i \in \mathcal{D}, \tag{4.2.12b}$$

where $\mathcal{D}$ is the set of all the UAVs equipped with a local detector as in (4.2.5).

Note that a successful attack detection using the hypothesis testing (4.2.12)

does depend on the sensitivity of the local residuals, r_i's, to the stealthy attacks. In this regard, the following results characterize the capability of local detectors in detecting stealthy attacks.

Proposition 4.2.2. *Consider dynamics* (4.1.12) *and let the i-th UAV be equipped with the local attack detector $\Sigma_{\mathcal{O}}^i$ in* (4.2.5) *and local measurements* (4.2.9)*. Then, stealthy ZDA and covert attacks are detectable in $\Sigma_{\mathcal{O}}^i$'s residual r_i if the set of compromised UAVs satisfies $\mathcal{A} \subseteq \mathcal{N}_\sigma^{i(1)},\ \sigma(t) = 1 \in \mathcal{Q}$.*

Proof. See Appendix B.2.

Note that the i-th UAV's local monitor, $\Sigma_{\mathcal{O}}^i,\ i \in \mathcal{D}$ secures the networked UAVs against the stealthy attacks on its neighbors' set $\mathcal{N}_\sigma^{i(1)}$. Therefore, the problem of interest is to determine a set $\mathcal{D} \subseteq \mathcal{V}$ of local detectors $\Sigma_{\mathcal{O}}^i$'s, $i \in \mathcal{D}$ such that they cover the entire set $\mathcal{V}$ of UAVs.

Proposition 4.2.3. *Consider the networked UAVs with the dynamics in* (4.1.12) *subject to stealthy attacks on a set of compromised UAVs $\mathcal{A} \subseteq \mathcal{V}$ and let the set*

$$\mathcal{D} := \{i \in \mathcal{V} \mid \bigcup_{i \in \mathcal{D}} \mathcal{N}_\sigma^{i(1)} = \mathcal{V},\ \sigma(t) = 1 \in \mathcal{Q}\}, \tag{4.2.13}$$

represent the set of UAVs equipped with local attack detectors $\Sigma_{\mathcal{O}}^i$'s in (4.2.5)*. Then, stealthy ZDA and covert attacks undetectable in $\Sigma_{\mathcal{O}}^i$'s residual $\mathrm{r}_i,\ \forall\, i \in \mathcal{D}$, is impossible, securing the entire network set $\mathcal{V}$ of UAVs against stealthy attacks.*

Proof. See Appendix B.3.

Algorithm 2 Attack detection by the i-th local monitor, $i \in \mathcal{D}$

Input: $\Sigma_{\mathcal{O}}^{i},\ i \in \mathcal{D}$ in (4.2.5) and (4.2.13), $\mathbf{y}_i$ in (4.2.9), ϵ_i^j in (4.2.11)

procedure LOCAL HYPOTHESIS TESTING (4.2.12)
  while $\mathcal{H}^{\circ}$ in (4.2.12a) **do**
    Compute local residual r_i as in (4.2.5)
    Compute corresponding thresholds ϵ_i^j as in (4.2.11)
    if $|\mathrm{r}_i^j| > \epsilon_i^j$ **then**
      Reject the null hypothesis $\mathcal{H}^{\circ}$ in (4.2.12a) ▷ Stealthy attack is locally detected.
      cooperate with other local detectors in $\mathcal{D}$ to run a contingency plan for the entire network.
    end if
  end while
end procedure

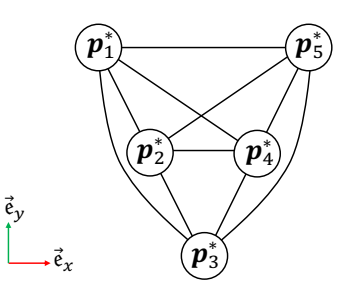
Algorithm 3 Topology switching for centralized attack detection

Inputs: local observer: $\Sigma_{\mathcal{O}}^{i},\ i \in \mathcal{D}$ in (4.2.5) and (4.2.13), $\mathbf{y}_i$ in (4.2.9), ϵ_i^j in (4.2.11);
centralized observer: $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ in (4.2.5), $\mathbf{y}$ in (4.1.11), ϵ_0^j in (4.2.6)

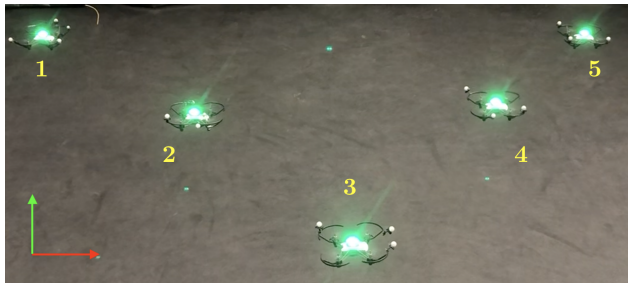
procedure CENTRAL HYPOTHESIS TESTING (4.2.7)
  Run Algorithm 2
  if $\mathcal{H}^1$ in (4.2.12b) **then**
    Switch to a new comm. mode $\sigma(t) \in \mathcal{Q} \setminus 1$ ▷ Stealthy attack has been detected locally.
  end if
  while $\mathcal{H}^0$ in (4.2.7a) **do**
    Compute central residual r_0 as in (4.2.5)
    Compute corresponding thresholds ϵ_0^j as in (4.2.6)
    if $|\mathrm{r}_0^j| > \epsilon_0^j$ **then**
      Reject the null hypothesis $\mathcal{H}^0$ in (4.2.7b) ▷ Stealthy attack is detected globally at the control center.
    end if
  end while
end procedure

It is worth mentioning that a trivial solution for (4.2.13) is $\mathcal{D} = \mathcal{V}$ that is all of the UAVs are equipped with a local detector, although this set can be optimally selected.

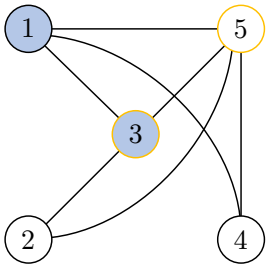
Given Propositions 4.2.2 and 4.2.3, one can verify that the networked UAVs can be secured against stealthy attacks using a set of local monitors, given by (4.2.5), that locally detect stealthy attacks, addressing problem (4.2.12). A procedure for this local hypothesis testing will be presented in Algorithm 2 in Section 4.3.



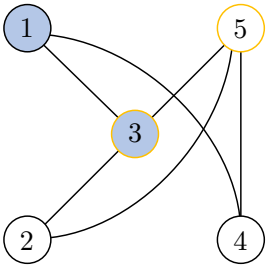
(a) Formation References



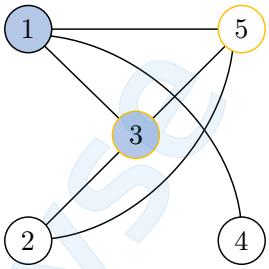
(b) V-shape Formation



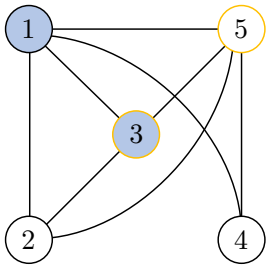
(c) mode 1



(d) mode 2



(e) mode 3



(f) mode 4

Figure 4.2: Multi-UAV's formation and communication topology. (a) Formation references specifying a V-shape in the x–y plane. (b) V-shape formation of UAVs. (c)-(f) Inter-UAV's communication graph $\mathcal{G}_{\sigma(t)}$ with four modes $\sigma(t) = \{1, 2, 3, 4\} =: \mathcal{Q}$. UAVs initially communicate in mode $\sigma(t) = 1$ and may switch to other modes $\sigma(t) = \{2, 3, 4\}$ if activated by a local detector. Blue nodes indicate the UAVs equipped with a local monitor and orange nodes specify the UAVs monitored by the ground control center.

4.3 Experimental Results

We conducted a set of experiments that served two purposes. First, the evaluation of the stealthiness of intrusions/deception attacks, described in Section 4.2.1, on the wireless communication network of a team of quadrotor UAVs in real-time practical settings. Second, the performance evaluation of the detection schemes is presented in Section 4.2.2.

4.3.1 Experimental Setup

Our experimental setup consists of a team of five homogeneous quadrotors (Tello Drones[5]), shown in Fig. 4.2b, flying in a 6 m $\times$ 4 m $\times$ 3 m flight area that is equipped with a VICON[6] motion capture system with 10 cameras. The VICON system provides the ground truth position and orientation of each UAV at 50 Hz for a central PC running Ubuntu 20.04 with ROS Noetic.

In our experiments, we use the VICON system's ground truth data available in the central PC to compute the high-level formation control commands that are sent to each UAV at 50 Hz and to run the central and local monitors, presented in Section 4.2.2. The central PC transmits the high-level formation control commands to the UAVs through different Wi-Fi channels and the UAVs' on-board attitude controllers use the received control commands to stabilize and steer the UAVs to their desired pose (see Fig. 4.1b). This connection setup allows us to replicate the peer-to-peer communication of the UAVs and also to implement stealthily intrusions on the Wi-Fi channels in a controlled setting.

4.3.2 Results

We conducted several experiments that serve as proof of concept of the real-world applicability of the proposed attack detection methods in multi-UAV cooperation settings. In these tests, the UAVs are tasked with achieving a V-shape formation. The special configuration of the V-shape formation and a picture of its real-world implementation are shown in Figs. 4.2a and 4.2b, respectively.

A video of our experiments is available at `https://www.youtube.com/watch?v=lVT_muezKLU`.

[5]https://www.ryzerobotics.com/tello.
[6]https://www.vicon.com.

Additionally, our framework is open-source and available at `https://github.com/SASLabStevens/TelloSwarm`.

In our experiments, the UAVs, indexed by $\mathcal{V} = \{1, 2, 3, 4, 5\}$, coordinate using the control protocol (3.1.3b), initially in communication mode, $\sigma(t) = 1$, shown in Fig. 4.2c, to achieve the V-shape formation. The UAVs also have consensus on their yaw angle $\psi_i = \psi^* = 0,\ \forall\, i \in \mathcal{V}$ as well as their hovering altitude. We select the position of UAVs 3 and 5 as the network-level monitored states at the ground control center that is $\mathcal{M}_{\rm p} = \{3, 5\}$ and $\mathcal{M}_{\rm v} = \emptyset$ in (4.1.11). These measurements are used in the realization of the central monitor (attack detector $\Sigma_{\mathcal{O}}^{\mathcal{M}}$) in (4.2.5) and its residuals in (4.2.6). We also let $\mathcal{D} = \{1, 3\}$ in (4.2.13), that is UAVs 1 and 3, which have the respective set of neighbors $\mathcal{N}^1_{\sigma(t)=1} = \{3, 4, 5\}$ and $\mathcal{N}^3_{\sigma(t)=1} = \{1, 2, 5\}$, and the local measurements $\mathbf{y}_i (\text{or } \mathrm{y}_i),\ i \in \mathcal{D}$ in (4.2.9), are selected as the host UAVs for local monitors (attack detectors $\Sigma_{\mathcal{O}}^i$'s) in (4.2.5). Accordingly, the condition (4.2.13) holds which in turn guarantees the local monitors of UAVs 1 and 3 are sufficient to locally detect the stealthy attacks on the entire network of UAVs in a decentralized manner. Also, as described earlier in Sections 4.1.2, the UAVs follow a decoupled dynamics in the x and y directions, and therefore we implement central and local monitors independently for the x- and y-direction dynamics based on the discretized models of (4.1.12), (4.2.5), and (4.2.5) with the sampling time $T_{\rm s} = 0.02$ sec.

In the following, we present the results of attack detection through the centralized detection scheme with central (global) hypothesis testing (4.2.7) and the central monitor (4.2.5) as well as through the decentralized detection scheme with local hypothesis testing (4.2.12) and the local monitors of UAVs 1 and 3. The procedure of the local hypothesis testing is presented in Algorithm 2 and that of the central hypothesis testing is presented in Algorithm 3.

Stealthy zero-dynamics attack. We conducted two experiments evaluating

the effectiveness of the central and local monitors in the detection of stealthy zero-dynamics attacks (ZDA). In the first experiment, UAVs 1, 4, and 5 are compromised such that their control channels are subject to the discretized version of ZDA signals in (4.2.2) as $\mathbf{u}_{\mathcal{A}} = \mathrm{col}\,(\boldsymbol{u}_{\mathrm{a}_i})_{i\in\mathcal{A}}$, $\mathcal{A} = \{1, 4, 5\}$ with $\boldsymbol{u}_{\mathrm{a}_i} = [\mathrm{u}^x_{\mathrm{a}_i}(0)e^{\lambda^x_{\mathrm{o}}(kT_{\mathrm{s}})} \;\; \mathrm{u}^y_{\mathrm{a}_i}(0)e^{\lambda^y_{\mathrm{o}}(kT_{\mathrm{s}})}\;]^\top$, $\lambda^x_{\mathrm{o}} = \lambda^y_{\mathrm{o}} = 0.5,\ k \in \mathbb{Z}_{\geq 0}$,

$$\mathrm{u}^x_{\mathrm{a}}(0) = [\mathrm{u}^x_{\mathrm{a}_1}(0)\ \mathrm{u}^x_{\mathrm{a}_4}(0)\ \mathrm{u}^x_{\mathrm{a}_5}(0)]^\top = [-2.34\mathrm{x}^{\mathrm{a}}_4(0)\ 10.24\mathrm{x}^{\mathrm{a}}_4(0)\ -2.34\mathrm{x}^{\mathrm{a}}_4(0)]^\top,$$

$\mathrm{u}^y_{\mathrm{a}}(0) = -0.7\mathrm{u}^x_{\mathrm{a}}(0),\ \mathrm{x}^{\mathrm{a}}_4(0) = 0.0086$, and the starting time $t = (kT_{\mathrm{s}}) = 0,\ k = 0$. The network-level measurements (4.1.11) with $\mathcal{M}_{\mathrm{p}} = \{3, 5\}$ and $\mathcal{M}_{\mathrm{v}} = \emptyset$, on the other hand, are not subject to sensory attacks that is $\mathbf{u}_{\mathcal{S}} = \mathbf{0}$.

In the first experiment, as shown in Fig. 4.3a, all the UAVs start from some initial positions and coordinate to achieve the desired formation while the stealthy ZDA steers UAV 4 away from its desired configuration that meets (4.1.6). This effect has been illustrated in Fig. 4.3b showing the relative positions of the UAVs as well as their desired values in the y direction over time. It is necessary to note that UAV 4 hits the safety net enclosing the indoor flight area at $t \approx 9.8$ sec.

In terms of attack detection, Figs. 4.4a and 4.4b show the residuals of local monitors $\Sigma^i_{\mathcal{O}}$'s, $i \in \{1, 3\}$ in (4.2.5) for UAVs 1 and 3, respectively. Also, Fig. 4.4c shows the residuals of the central monitor $\Sigma^{\mathcal{M}}_{\mathcal{O}}$ in (4.2.5) available in the control center. One can verify that the local monitor of UAV 1, $\Sigma^1_{\mathcal{O}}$, running Algorithm 2, has detected the stealthy ZDA in a timely manner ($t = 3.22$ sec) that is before UAV 4 collides with the safety net of the flight area at $t \approx 9.8$ sec. However, the ZDA remains stealthy in the residuals of the UAV 3's local monitor, $\Sigma^3_{\mathcal{O}}$, and those of the central monitor running Algorithm 3, regardless of the switch in the inter-UAV's communication topology from mode 1 to mode 4 (see Fig. 4.2) that is triggered by

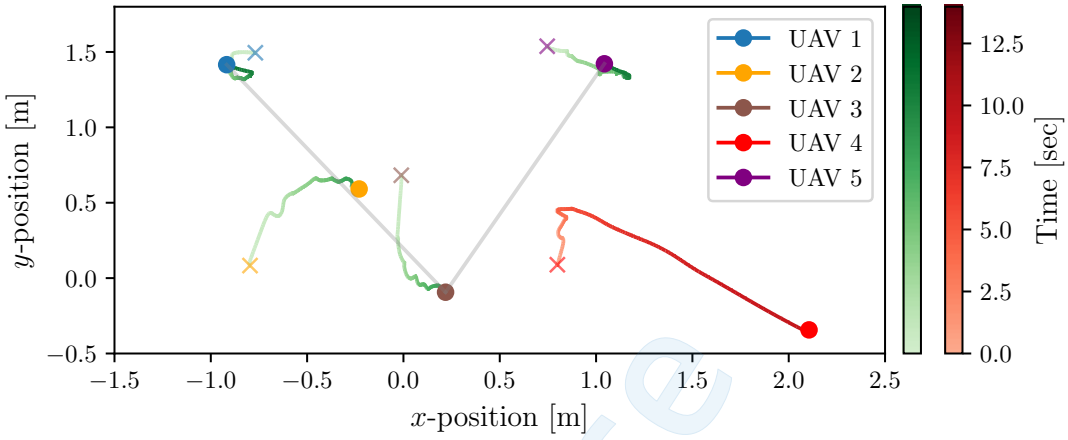
the local monitor $\Sigma_{\mathcal{O}}^1$ at $t = 3.22$ sec. This is due to the fact that switching from mode 1 to mode 4 does not meet the necessary conditions required for a topology switching to render stealthy attacks detectable for the central monitor in (4.2.5). The details of such conditions have been studied in Theorem 3.2.4.

In the second experiment, with the results shown in Fig. 4.5, the UAVs are under the same ZDA as in the first experiment expect $\mathrm{x}_4^{\mathrm{a}}(0) = 0.012$. The local monitor $\Sigma_{\mathcal{O}}^1$ successfully detects the ZDA at $t = 5.08$ sec (see Fig. 4.5b) and then triggers a switch in the UAVs' communication from mode 1 to mode 3 (cf. Fig. 4.2) that as opposed to Experiment 1, this topology switching results in detection of stealthy ZDA by the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ at $t = 5.6$ sec (see Fig. 4.5c).

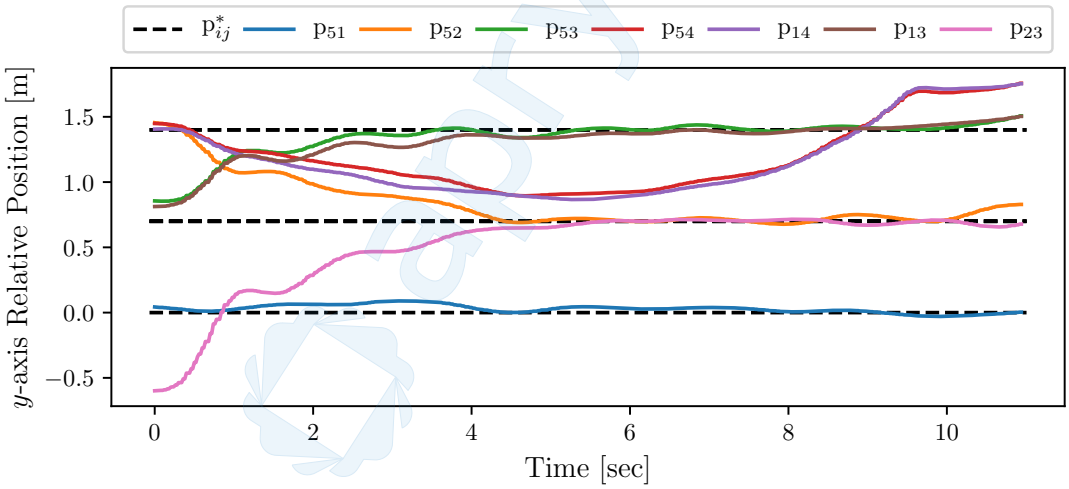
It is necessary to note that any topology switching in the inter-UAV communications results in a discrepancy between the actual dynamics of networked UAVs and its nominal counterpart that is used by the attacker to design stealthy attacks. Yet, the model discrepancy in Experiment 1 did not interfere with the stealthiness of ZDA in the central monitor's residuals while it renders ZDA detectable in the central monitor's residuals in Experiment 2. These results indicate that not only zero-dynamics attacks (ZDA) can be implemented in real-time on networked UAVs with partial measurements, but they also can remain stealthy regardless of switches in the inter-UAVs' communication topology. Theoretical results to detect stealthy ZDA through topology switching in networked systems with full-state measurements and with partial measurements can be found, respectively, in [78] and Section 3.2.

Covert attack. Similar to the ZDA case, we evaluated the detection of covert attacks on networked UAVs subject to topology switching by using the local monitors $\Sigma_{\mathcal{O}}^1$ and $\Sigma_{\mathcal{O}}^3$, and the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$. In this experiment, a covert attack, $\mathbf{u}_{a_i},\ i \in \mathcal{A} = \{2\}$, in the form of a ramp signal with a slope of 3 Deg , as the roll and pitch angles' perturbation, and the starting time of $t_{\mathrm{a}} = 5$ sec is injected through the

control channel of UAV 2. The covert attack's effect on the UAVs' formation is shown in Fig. 4.6a. As illustrated, all of the UAVs have deviated from their desired formation configuration that meets (4.1.6). The effect of this deviation/perturbation on the measurements $\mathbf{y}$ in (4.1.11) is simultaneously canceled out by implementing the discretized version of the sensory attack $\mathbf{u}_{\mathcal{S}}$ given in (4.2.3). Fig. 4.6b illustrates how the alteration of actual measurement $\mathbf{y}$ of the monitored UAV 3 using the sensory attack $\mathbf{u}_{\mathcal{S}}$ gives rise to a false state estimation by the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$, rendering the injected attack covert in the central residuals. The local monitors, however, are not subject to such alterations and thus are capable of detecting the covert attack in a timely manner as shown in Fig. 4.6c for the local monitor $\Sigma_{\mathcal{O}}^{3}$ of UAV 3. We note that the local monitor $\Sigma_{\mathcal{O}}^{3}$ triggers a topology switch from mode 1 to mode 2 (cf. Fig. 4.2) at $t = 6.4$ sec to make the covert attack detectable in the residuals of the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$. However, the attack remains stealthy in the central residuals, shown in Fig. 4.6d, regardless of topology switching. The results, consistent with those in the ZDA case, show the outperformance of the decentralized detection scheme (Algorithm 2) over the centralized detection scheme (Algorithm 3). It is worth mentioning that one can leverage a larger number of switching communication links on which the centralized monitor relies to improve the performance of the centralized detection scheme. However, this solution raises other challenges such as switching-induced unobservability as well as communication overhead. In the case of the decentralized detection scheme that relies on the network model of UAVs, scalability is a concern for larger teams of UAVs, for which clustering-based solutions such as the one in Section 3.2 can be applied.

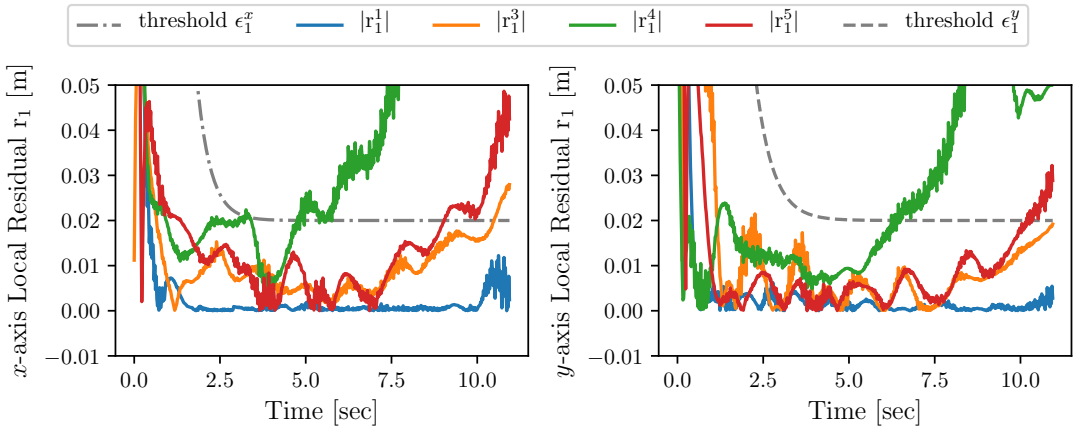


(a) UAVs' position trajectories in the $x-y$ plane.

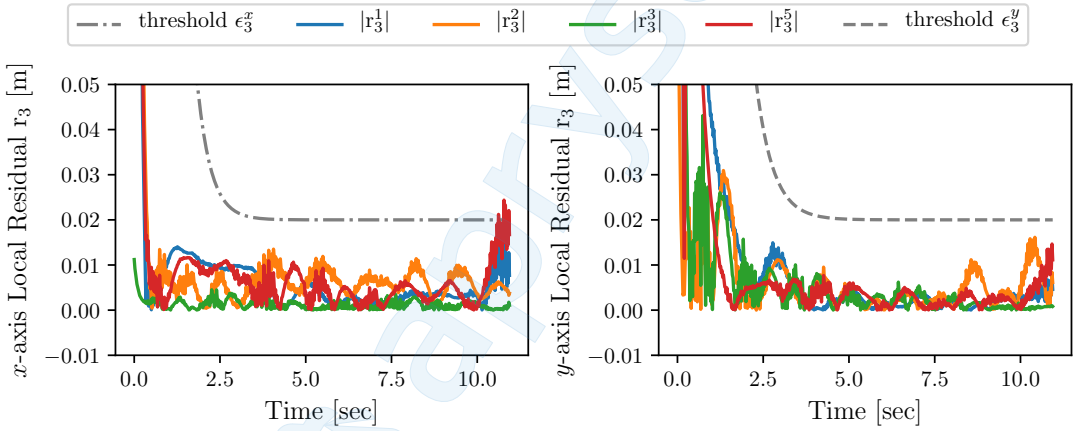


(b) Coordination of UAVs in the y direction.

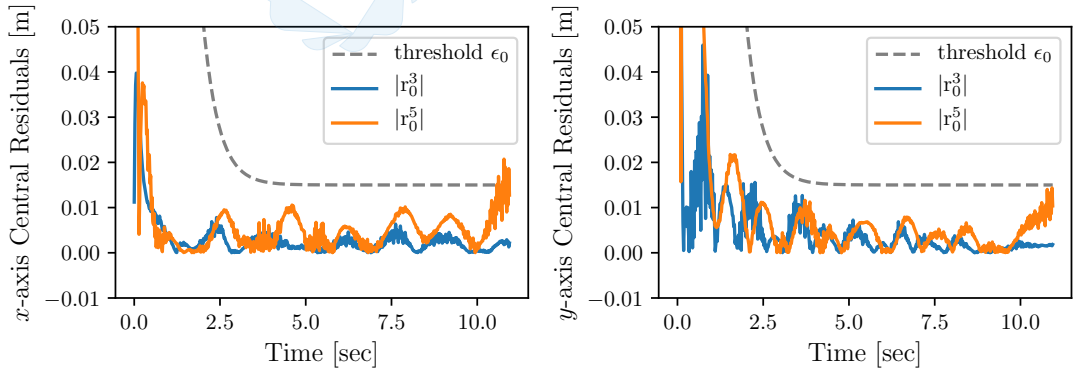
Figure 4.3: Experiment 1: ZDA on UAVs $1, 4, 5$ and topology switching from mode 1 to 4. (a) UAVs' position trajectories in the $x-y$ plane with the colorbars quantifying the timespan. The $\times$ markers and the colored circles show, respectively, the UAVs' initial position and final position during the experiment. Finally, the gray lines visualize the V-shape formation achieved by the final position of the UAVs. (b) The relative positions of UAVs in the y direction, corresponding to the inter-UAV communication links in mode $\sigma(t) = 1$, shown in Fig. 4.2c. Also, the dashed lines, labeled by $\mathrm{p}^*_{ij},\ i, j \in \mathcal{V}$, denote the desired relative positions based on the formation references in Fig. 4.2a.



(a) Residuals of local monitor $\Sigma_{\mathcal{O}}^1$ run on UAV 1. The stealthy ZDA is locally detected at $t = 3.22$ sec.



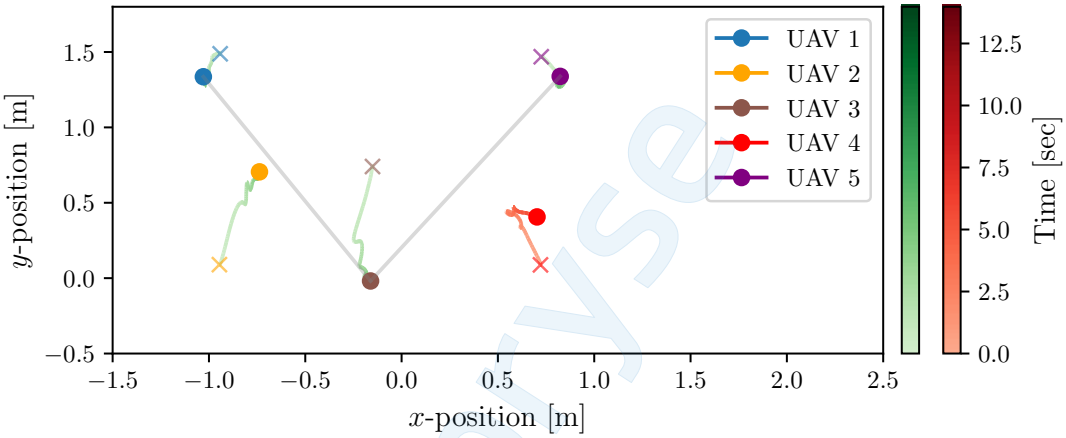
(b) Residuals of local monitor $\Sigma_{\mathcal{O}}^3$ run on UAV 3.



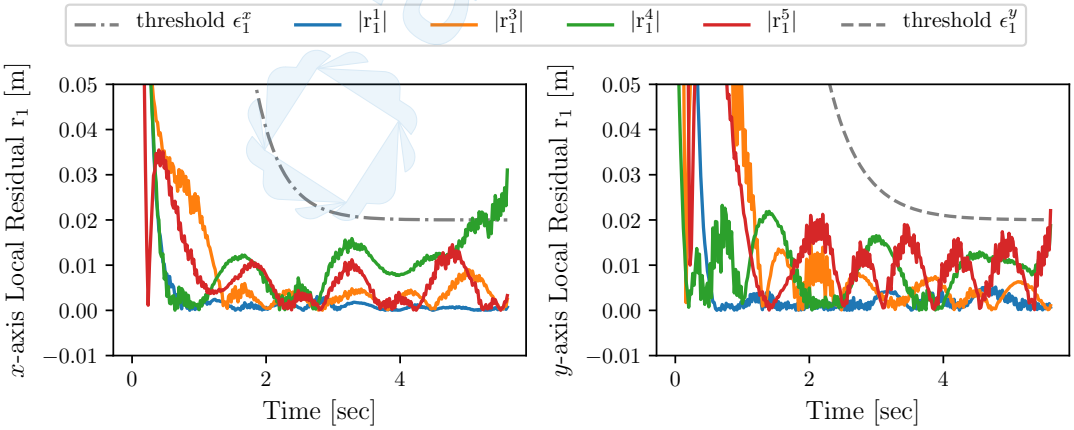
(c) Residuals of central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ run on the control center.

Figure 4.4: Experiment 1: ZDA on UAVs $1, 4, 5$ and topology switching from mode 1 to 4, which is triggered by local monitor $\Sigma_{\mathcal{O}}^1$ at $t = 3.22$ sec. [Continued on next page]

Figure 4.4: [cont'd]: (a)-(b) The notation $\mathrm{r}_1^i,\ i \in \{1,3,4,5\}$ ($\mathrm{r}_3^i,\ i \in \{1,2,3,5\}$), denotes the residual of position estimation for the UAV 1's (3's) neighbors obtained by its local monitor $\Sigma_{\mathcal{O}}^1$ ($\Sigma_{\mathcal{O}}^3$) in the x and y directions with the respective thresholds ϵ_1^x (ϵ_3^x) and ϵ_1^y (ϵ_3^y) as given in (4.2.11). (c) The notation $\mathrm{r}_0^i,\ i \in \{3,5\}$, denotes the residual of position estimation for UAVs 3 and 5 by the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ in the x and y directions with the threshold ϵ_0 as given in (4.2.6).

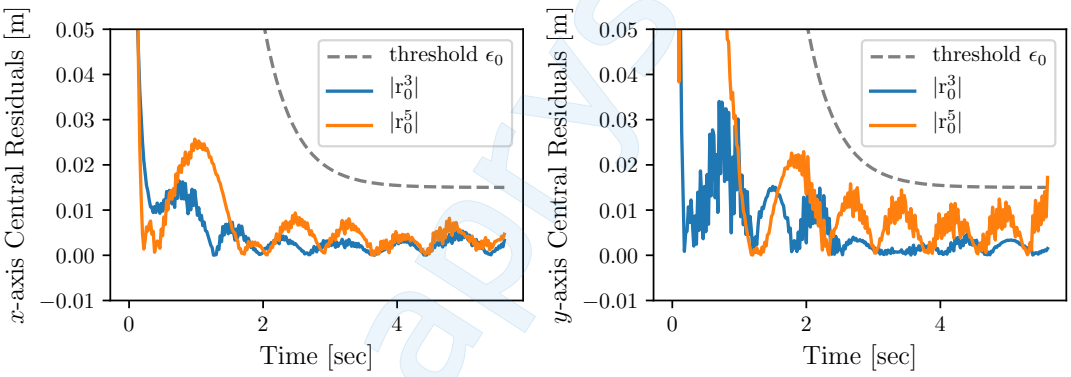


(a) UAVs' position trajectories in the $x-y$ plane. The central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ detects the ZDA at $t = 5.6$ sec and ends the experiment.



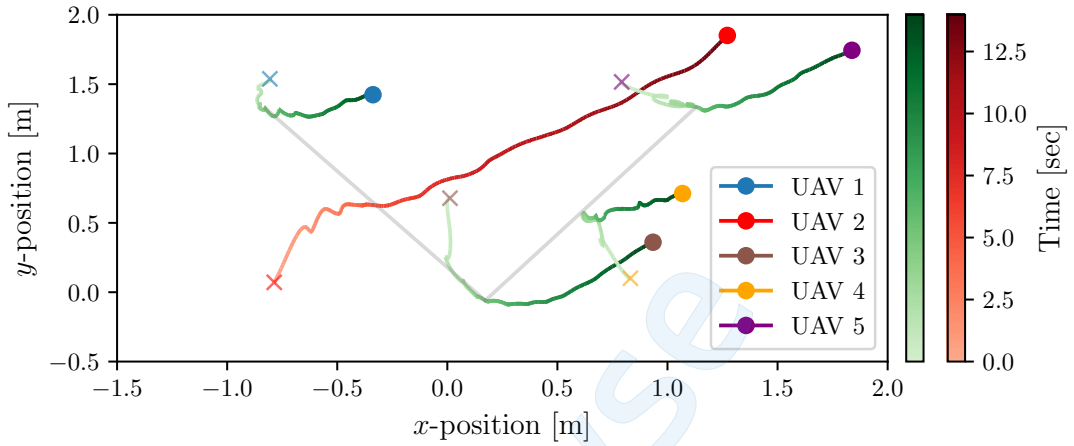
(b) Residuals of local monitor $\Sigma_{\mathcal{O}}^1$ run on UAV 1. The stealthy ZDA is locally detected at $t = 5.08$ sec.

Figure 4.5: Experiment 2: ZDA on UAVs $1, 4, 5$ and topology switching from mode 1 to 3, which is triggered by local monitor $\Sigma_{\mathcal{O}}^1$ at $t = 5.08$ sec. (a) UAVs' position trajectories in the $x-y$ plane with the same annotations as in Fig. 4.3a. (b) The residuals of local monitor $\Sigma_{\mathcal{O}}^1$ with the same annotations as in Figs. 4.4a and 4.4c, respectively. [Continued on next page]

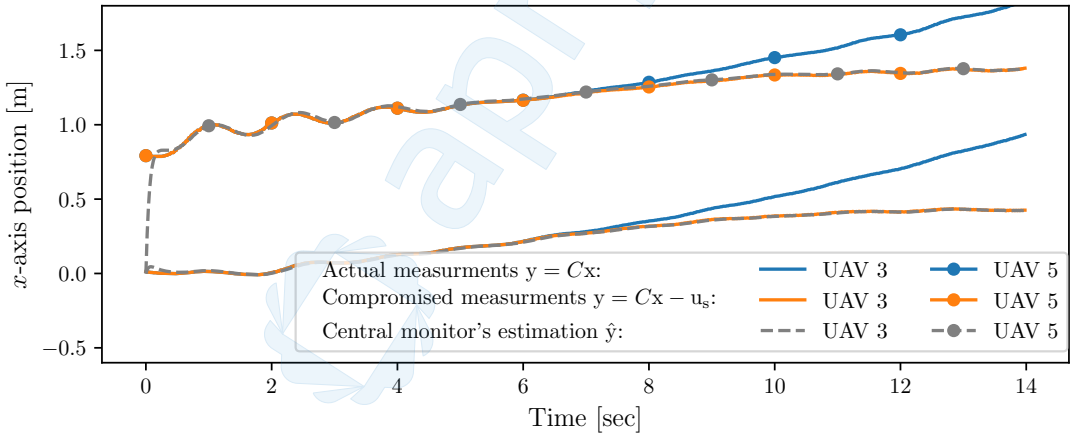


(c) Residuals of central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ run on the control center. The stealthy ZDA is detected at $t = 5.6$ sec using Algorithm 3.

Figure 4.5: [cont'd]: (c) The residuals of the central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ with the same annotations as in Figs. 4.4a and 4.4c, respectively.

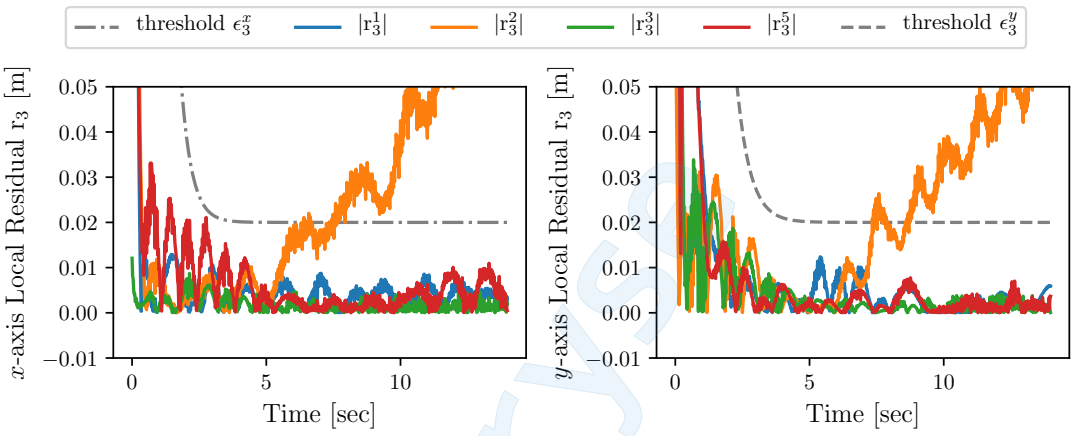


(a) UAVs' position trajectories in the $x-y$ plane.

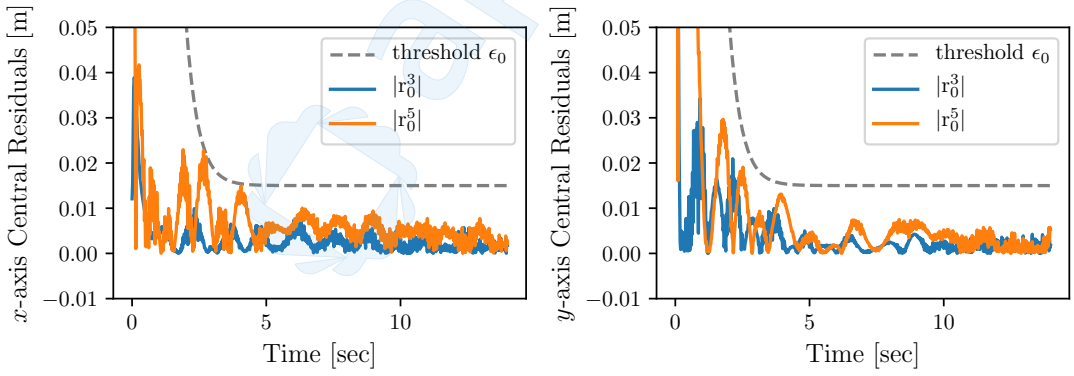


(b) Realization of stealthiness condition (4.1.13) using $\mathbf{u}_{\mathcal{S}}$ in (4.2.3) with the starting time $t_{\mathrm{a}} = 5$ sec.

Figure 4.6: Experiment 3: covert attack on UAV 2 and topology switching from mode 1 to 2, which is triggered by local monitor $\Sigma_{\mathcal{O}}^{1}$ at $t = 6.4$ sec. (a) UAVs' position trajectories in the $x-y$ plane with the same annotations as in Fig. 4.3a, except the gray lines that visualize the V-shape formation achieved by the UAVs at $t_{\mathrm{a}} = 5$ sec, the starting time of the covert attack. (b) The effect of measurement alteration using sensory attack $\mathbf{u}_{\mathcal{S}}$ starting at $t_{\mathrm{a}} = 5$ sec. [Continued on next page]



(c) Residuals of local monitor $\Sigma_{\mathcal{O}}^3$ run on UAV 1. The stealthy ZDA is detected at $t = 6.4$ sec using Algorithm 2.



(d) Residuals of central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ run on the control center.

Figure 4.6: [cont'd]: (c)-(d) The residuals of local monitor $\Sigma_{\mathcal{O}}^3$ and central monitor $\Sigma_{\mathcal{O}}^{\mathcal{M}}$ with the same annotations as in Figs. 4.4a and 4.4c, respectively.

Chapter 5

Distributed Deception-Attack Detection for Resilient Cooperation of Multi-Robot Systems with Intermittent Communication

This chapter[1] extends the results of previous chapters by considering both deception attacks and Denial-of-Service (DoS) attacks as well as arbitrarily time-varying (switching) communication networks. We consider time-varying communication networks subject to intermittent connections, which may be caused by DoS attacks, rendering the information exchange unreliable. We will show if the arbitrarily time-varying (switching) communication network maintains its connectivity only in an integral sense, uniformly in time, the following results are guaranteed:

In Section 5.2, we characterize and provide explicit bounds for the network resilience to both intermittent and permanent disconnections. The former is relevant to *DoS* attacks, and the latter is relevant to *deception* attacks. We also provide explicit bounds for uniformly exponentially fast convergence of the multi-agent systems in the presence of a class of *DoS* attacks as well as for their bounded-input-bounded-output (BIBO) stability in the presence of a class of *deception* attacks. Compared to the previous results [144, 37, 135], the network resilience is quantified explicitly based on algebraic connectivity in an integral sense, and the connectivity and stability analyses for both types of attacks are in the continuous-time domain.

In Section 5.3, we characterize the system vulnerability to a class of stealthy deception attacks, based on zero dynamics of the switched systems, and provide explicit worst-case bounds on the number of malicious agents subject to deception

[1]This chapter is adapted from a publication by the author of this dissertation. © 2024 IEEE. Reprinted, with permission, from Bahrami, M., & Jafarnejadsani, H. (2024, August). Distributed Detection of Adversarial Attacks for Resilient Cooperation of Multi-Robot Systems with Intermittent Communication. Provisionally Accepted at IEEE Transactions on Control of Network Systems.

attacks that can be detected in a given network. Compared to the previous results [97, 37], we show some of these well-known bounds can be improved, provided some extra information on the local dynamics is available only in an integral sense. We then present a distributed and reconfigurable framework with theoretical guarantees for the distributed detection of malicious agents introducing deception attacks. Compared to centralized frameworks [78], our framework relies solely on locally available information in an integral sense, making it well-suited for mobile agent applications subject to intermittent connectivity.

Additionally, in Section 5.4, we present an algorithmic framework for detaching from the detected set of malicious agents and for achieving resilient coordination and cooperation.

5.1 Problem Formulation

5.1.1 System Dynamics

Consider the multi-agent (robot) system in (3.1.4a), and let $\mathbf{y}_\sigma^i$ denote the state measurements available for the i-th mobile agent consisting of the (relative) position states of a set of neighboring agents $\{i\}\cup\mathcal{N}_\sigma^{i(1)} \subseteq \mathcal{I}_i \subseteq \mathcal{V}$ (where $\mathcal{I}_i$ will be determined later, and is different than (3.1.4b), (3.2.1), and (3.1.5)) and the velocity state $\boldsymbol{v}_i$. Then, we have

$$\Sigma_{\sigma(t)}: \begin{bmatrix} \dot{\widetilde{\boldsymbol{p}}} \\ \dot{\boldsymbol{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I \\ -\alpha \mathsf{L}_\sigma & -\gamma I \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{p}} \\ \boldsymbol{v} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ I_{\mathcal{A}} \end{bmatrix} \mathbf{u}_{\mathcal{A}}$$

$$=: \mathbf{A}_\sigma \mathbf{x} + \mathbf{B}_{\mathcal{A}} \mathbf{u}_{\mathcal{A}}, \qquad \mathbf{x}_0 = \mathbf{x}(t_0), \tag{5.1.1}$$

$$\mathbf{y}_\sigma^i = \mathrm{col}(\widetilde{\boldsymbol{p}}_{j\in\mathcal{I}_i}, \boldsymbol{v}_i) =: \mathbf{C}_\sigma^i \mathbf{x}. \tag{5.1.2}$$

It is necessary to note that the nature of arbitrary switching modes $\sigma(t) \in \mathcal{Q}$, induced by the unreliability of network $\mathcal{G}_{\sigma(t)}$, renders *a priori* unknown knowledge of system matrix $\mathbf{A}_\sigma$ in (5.1.1). This imposes stability and observability challenges in distributed settings which will be addressed in Sections 5.2 and 5.3.

5.1.2 Communication Topology

The switching[2] communication network, $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$, $\sigma(t) \in \mathcal{Q}$, with a piece-wise constant and right-continuous switching signal $\sigma(t) : \mathbb{R}_{\geq 0} \to \mathcal{Q}$, is considered to the represent time-varying communication links, intermittent communication, and lossy datalinks between mobile agents. We let the subgraphs $\mathcal{G}^{i'}_{\sigma(t)} = (\mathcal{V}^{i'}_\sigma, \mathcal{E}^{i'}_\sigma)$, $\bar{\mathcal{G}}^{i}_{\sigma(t)} = (\mathcal{V}^{i'}_\sigma, \bar{\mathcal{E}}^{i}_\sigma)$, and $\mathcal{G}^{i''}_{\sigma(t)} = (\mathcal{V}^{i''}_\sigma, \mathcal{E}^{i''}_\sigma)$, denote, respectively, the 1-hop proximity communication network, the 1-hop induced communication network, and the 2-hop proximity communication network of the i-th mobile agent with its k-hop neighbors, $k \in \{1, 2\}$, for which the vertex and edge sets are defined as

$$\mathcal{V}^{i'}_\sigma := \{i\} \cup \mathcal{N}^{i(1)}_\sigma, \qquad \bar{\mathcal{E}}^{i}_\sigma := \{i\} \times \mathcal{N}^{i(1)}_\sigma \subseteq \mathcal{E}_\sigma, \tag{5.1.3a}$$

$$\bar{\mathcal{E}}^{i}_\sigma := \left(\mathcal{V}^{i'}_\sigma \times \mathcal{V}^{i'}_\sigma\right) \cap \mathcal{E}_\sigma, \tag{5.1.3b}$$

$$\mathcal{V}^{i''}_\sigma := \mathcal{V}^{i'}_\sigma \cup \mathcal{N}^{i(2)}_\sigma, \qquad \mathcal{E}^{i''}_\sigma := \mathcal{E}^{i'}_\sigma \cup \left(\left(\mathcal{N}^{i(1)}_\sigma \times \mathcal{N}^{i(2)}_\sigma\right) \cap \mathcal{E}_\sigma\right). \tag{5.1.3c}$$

Having defined the k-hop neighbors, the Laplacian matrix L_σ can be partitioned according to the incoming flow of information to the agent $i \in \mathcal{V}$. Let $i \in \mathcal{V}$ be the first agent and accordingly $\mathcal{V}^{i'}_\sigma$ come first, $\mathcal{N}^{i(2)}_\sigma$ come second, and $\mathcal{V} \setminus \{\mathcal{V}^{i'}_\sigma \cup \mathcal{N}^{i(2)}_\sigma\}$

[2] interchangeably time-varying

come last, then L_σ can be rewritten as

$$\mathsf{L}_\sigma = \left[\begin{array}{c|c|c} \mathsf{L}_\sigma^{(11)} & \mathsf{L}_\sigma^{(12)} & \mathbf{0} \\ \hline \mathsf{L}_\sigma^{(21)} & \mathsf{L}_\sigma^{(22)} & \mathsf{L}_\sigma^{(23)} \\ \hline \mathbf{0} & \mathsf{L}_\sigma^{(32)} & \mathsf{L}_\sigma^{(33)} \end{array}\right], \mathsf{L}_\sigma'' = \left[\begin{array}{c|c} \mathsf{L}_\sigma^{(11)} & \mathsf{L}_\sigma^{(12)} \\ \hline \mathsf{L}_\sigma^{(21)} & \mathsf{L}_\sigma^{(\prime\prime\setminus\prime)} \end{array}\right], \tag{5.1.4a}$$

$$\mathsf{L}_\sigma^{(11)} = \mathsf{L}_\sigma' + \widetilde{\mathsf{L}}_\sigma, \qquad \left[\widetilde{\mathsf{L}}_\sigma \quad \mathsf{L}_\sigma^{(12)}\right]\mathbf{1} = \mathbf{0}, \tag{5.1.4b}$$

$$\mathsf{L}_\sigma^{(22)} = \mathsf{L}_\sigma^{(\prime\prime\setminus\prime)} + \widetilde{\widetilde{\mathsf{L}}}_\sigma, \qquad \left[\widetilde{\widetilde{\mathsf{L}}}_\sigma \quad \mathsf{L}_\sigma^{(23)}\right]\mathbf{1} = \mathbf{0}, \tag{5.1.4c}$$

in which L_σ' is the Laplacian matrix of the 1-hop proximity graph $\mathcal{G}_{\sigma(t)}^{i'} = (\mathcal{V}_\sigma^{i'}, \mathcal{E}_\sigma^{i'})$. $\widetilde{\mathsf{L}}_\sigma$ encodes the edge set $\bar{\mathcal{E}}_\sigma^i \setminus \mathcal{E}_\sigma^{i'}$, that is the set of existing edges between the 1-hop neighbors with one another, and the set of existing edges between the 1-hop neighbors and the 2-hop neighbors. $\mathsf{L}_\sigma^{(\prime\prime\setminus\prime)}$ is the Laplacian matrix encoding the edge set $\mathcal{E}_\sigma^{i''} \setminus \mathcal{E}_\sigma^{i'}$ that is the connections of the 2-hop neighbors with the 1-hop neighbors, and $\widetilde{\widetilde{\mathsf{L}}}_\sigma$ encodes the existing edges between the 2-hop neighbors with one another, and those existing between the 2-hop neighbors and the rest, i.e. $\mathcal{V} \setminus \{\mathcal{V}_\sigma^{i'} \cup \mathcal{N}_\sigma^{i(2)}\}$. Finally, L_σ'' is the Laplacian matrix associated with the 2-hop proximity graph $\mathcal{G}_{\sigma(t)}^{i''} = (\mathcal{V}_\sigma^{i''}, \mathcal{E}_\sigma^{i''})$.

5.1.3 Adversary Model

We consider two classes of adversarial attacks, namely *deception* attacks and *denial-of-service* (DoS) attacks.

Deception attacks. In this model, a set of malicious agents $\mathcal{A} \subset \mathcal{V}$, as described in Section 5.1.1, inject some undesirable data $0 \neq \boldsymbol{u}_i^{\mathrm{a}}(t) \in \mathcal{L}_{pe}$, $\forall\, i \in \mathcal{A}$, $\forall\, t \in [t_i^{\mathrm{a}}, \infty)$, where $t_i^{\mathrm{a}} \in \mathbb{R}_{\geq 0}$ is the activation time instant in (3.1.3). Among the well-studied deception attacks including data injection attack [36, 67], zero-dynamics attacks (ZDA) [97, 78, 5], covert attack [42, 4], replay attack [110], and Byzantine attacks [67], our analysis covers the first two models. Similar to [67, 37], the worst-case

upper bounds on the number of malicious agents in the network are parameterized as follows:

Definition 5.1.1. (F-local and F-total adversary sets). The unknown adversary set $\mathcal{A} \subset \mathcal{V}$ is termed F-total if $|\mathcal{A}| \leq F$, where $F \in \mathbb{Z}_{\geq 0}$, that is there exist at most F malicious agents in the network with $0 \neq \boldsymbol{u}_i^{\mathrm{a}}(t) \in \mathcal{L}_{pe}$ in (3.1.3). The set $\mathcal{A} \subset \mathcal{V}$ is termed F-local if $\forall\, i \in \mathcal{V} \setminus \mathcal{A},\ |\mathcal{A} \cap \boldsymbol{\mathcal{N}}^{i(1)}| \leq F$, where $F \in \mathbb{Z}_{\geq 0}$ and the aggregated set of 1-hop neighbors $\boldsymbol{\mathcal{N}}^{i(1)} = \{j \in \mathcal{V} \mid (i,j) \in \boldsymbol{\mathcal{E}}_T^{\mu}\}$ with the edge set $\boldsymbol{\mathcal{E}}_T^{\mu}$ (to be specified) defined uniformly over the time interval $[t,\, t+T)$, $\forall\, t \in \mathbb{R}_{\geq 0}$, $\exists\, T \in \mathbb{R}_{>0}$. i.e., each cooperative agent has no more than F malicious agents with $0 \neq \boldsymbol{u}_i^{\mathrm{a}} \in \mathcal{L}_{pe}$ in (3.1.3) among its aggregated set of 1-hop neighbors defined uniformly in time.

An explicit upper bound on F, and the explicit definition of the edge set $\boldsymbol{\mathcal{E}}_T^{\mu}$ will be given in Section 5.3.

Remark 5.1.1. *The F-local model presented herein is a relaxation of the model in [67, 146] that required the upper bound inequality holds point-wise in time. (cf. the discrete-time version in [37, Sec. 4.4] and [113]).*

Denial-of-Service (DoS) attack. We consider a time-constrained (distributed) DoS attack on the communication network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ that causes the intermittent unavailability of (state) information exchange, either partially or fully [2, 34, 74]. We take into account such DoS attacks by the inclusion of some modes $\sigma(t) \in \mathcal{Q}^{\mathrm{a}} \subset \mathcal{Q}$ for the network $\mathcal{G}_{\sigma(t)}$ where an unknown subset of edges $\mathcal{E}_{\sigma}(t)$ are nullified. Accordingly, $\mathcal{G}_{\sigma(t)\in\mathcal{Q}^{\mathrm{a}}}$ is at most disconnected as a consequence of nullified (blocked) edge links, that is

$$\begin{aligned} \exists\, \mathcal{E}_{\sigma(t)}^{\mathrm{a}} = \big\{(i,j) \in \mathcal{E}_{\sigma(t)} \mid (i,j) = \emptyset,\ i,j \in \mathcal{V}\big\} \qquad \text{s.t.} \\ \lambda_2(\mathsf{L}_{\sigma(t)}) = 0, \quad \forall\, t \in \boldsymbol{\mathcal{T}}_{\mathrm{a}} = \{\mathcal{T}_k^{\mathrm{a}}\}_{k\in\mathbb{Z}_{\geq 0}}^{\mathbf{n}}, \end{aligned} \tag{5.1.5}$$

where $\lambda_2(\mathsf{L}_{\sigma(t)})$ is the algebraic connectivity at t, $\boldsymbol{\mathcal{T}}_{\mathrm{a}} = \{\mathcal{T}_k^{\mathrm{a}}\}_{k\in\mathbb{Z}_{\geq 0}}^{\mathbf{n}}$ with $k \leq \mathbf{n} \in \mathbb{Z}_{\geq 0}$, denotes a finite sequence of $\mathbf{n}$ DoS attacks having bounded but not necessarily contiguous time intervals $\mathcal{T}_k^{\mathrm{a}}$'s, where $\mathcal{T}_k^{\mathrm{a}} = [t_k,\, t_k + T_k^{\mathrm{a}})$ with $T_k^{\mathrm{a}} \in \mathbb{R}_{>0}$ and $\sigma(t_k) \in \mathcal{Q}^{\mathrm{a}}$.

We will show that if the set $\boldsymbol{\mathcal{T}}_{\mathrm{a}}$ in (5.1.5) is sufficiently small the cooperation objective (5.1.8) is achievable.

Remark 5.1.2. *We note that the time-constrained DoS attack model* (5.1.5) *is consistent with the assumptions and discussions in [34, 74, 2]. It is also worth mentioning that blocking the communication channels (i.e., nullifying part of the edge set $\mathcal{E}_{\sigma(t)}$) can take place independently for each communication link in the communication settings with multiple transmission channels [74] (e.g., P2P communication over IEEE 801.11s networks) or centrally for a large group (or all) of transmission channels in the single-channel architectures [34, 2].*

5.1.4 Problem Statement

Consider the multi-agent system $\Sigma_{\sigma(t)}$ in (5.1.1) with an unreliable communication network, $\mathcal{G}_{\sigma(t)}$, $\sigma(t) \in \mathcal{Q}$, subject to the DoS attack in (5.1.5) as well as deception attacks that are injected by a set of *malicious* agents $\mathcal{A} \subset \mathcal{V}$. The problems of interest are distributed detection of the set of *malicious* agents $\mathcal{A}$, and the resilient cooperation of the remaining cooperative agents $\mathcal{V} \setminus \mathcal{A}$.

Distributed attack detection. We cast the attack detection problem as a form of distributed hypothesis testing problem where each mobile agent Σ_i in (3.1.1) locally verifies either the null hypothesis $\mathcal{H}^{\mathrm{o}}$: a*ttack*-f*ree*, if $\mathcal{N}_\sigma^{i(1)} \cap \mathcal{A} = \emptyset$ or the alternative hypothesis $\mathcal{H}^{\mathrm{1}}$: a*ttacked*, if $\mathcal{N}_\sigma^{i(1)} \cap \mathcal{A} \neq \emptyset$. For this purpose, we equip each Σ_i in (3.1.1) with a reconfigurable local attack detector module of the form

$\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}$: $\mathbf{r}^i_\sigma(t) = O^i_\sigma(\mathbf{y}^i_\sigma)$, in which $O^i_\sigma(\cdot)$ is a stable linear filter (e.g., a Luenberger-type observer) in each mode $\sigma \in \mathcal{Q}$, whose explicit expression will be given in Section 5.3.3. Also, its inputs and outputs are, resp., $\mathbf{y}^i_\sigma$ in (5.1.2) and the residual $\mathbf{r}^i_\sigma(t) = \mathbf{y}^i_\sigma - \hat{\mathbf{y}}^i_\sigma$, where $\hat{\mathbf{y}}^i_\sigma$ is the estimation of $\mathbf{y}^i_\sigma$. Given the switching nature of $\Sigma_{\sigma(t)}$ in (5.1.1) with possibly unknown modes of $\mathcal{G}_{\sigma(t)}$ subject to the DoS attack in (5.1.5), it is necessary to note that the realization and reconfiguration of $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}$, rely on (a minimum amount of) local information that is available intermittently not point-wise in time. This contains the set of 2-hop information available for each agent $i \in \mathcal{V}$, defined as

$$\Phi^i_{\sigma(t)} = \Big\{\mathcal{G}^{i''}_{\sigma(t)} = (\mathcal{V}^{i''}_\sigma, \mathcal{E}^{i''}_\sigma),\ \boldsymbol{p}_j(t),\ \forall\, j \in \mathcal{V}^{i''}_\sigma, \boldsymbol{v}_i(t)\Big\}, \tag{5.1.6}$$

where the topological knowledge $\mathcal{G}^{i''}_{\sigma(t)}$, defined by (5.1.3c), can be either obtained via information exchange with only the 1-hop neighbors $\mathcal{N}^{i(1)}_\sigma$, $i \in \mathcal{V}$ upon network availability or be pre-programmed as in autonomous monitoring scenarios [144, 78]. We remark that we do not explicitly address the case of F-total *Byzantine* agents that transmit inconsistent information to their neighbors, and refer to [67].

The attack detector module $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}$ allows for quantification and verification of the simple null and alternative hypotheses using the local residuals, $\mathbf{r}^i_\sigma(t)$'s, as follows:

$$\mathcal{H}^0 : \text{attack-free}, \quad \text{if} \quad \forall\, j \in \mathcal{N}^{i(1)}_\sigma,\ \forall\, i \in \mathcal{V} \setminus \mathcal{A},\ \ |\mathbf{r}^{i,j}_\sigma(t)| \le \epsilon^{i,j}_\sigma,\ \forall\, t \in \mathbb{R}_{\ge 0}, \tag{5.1.7a}$$

$$\mathcal{H}^1 : \text{attacked}, \quad \text{if} \quad \exists\, j \in \mathcal{N}^{i(1)}_\sigma,\ \exists\, i \in \mathcal{V} \setminus \mathcal{A},\ \ |\mathbf{r}^{i,j}_\sigma(t)| > \epsilon^{i,j}_\sigma,\ \exists\, t \in \mathbb{R}_{\ge 0}, \tag{5.1.7b}$$

where $\mathbf{r}^{i,j}_\sigma(t)$ is the j-th component of the residual signal of the local attack detector $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_i}$, and $\epsilon^{i,j}_\sigma$'s are the corresponding (dynamic) thresholds that will be defined later in Section 5.3.3.

Remark 5.1.3. *The detection scheme herein is also known as change detection and sequential hypothesis testing in stochastic settings [16, 138] and has been used for observer-based attack detection in distributed settings [42]. For any given class of dynamical systems subject to deception attacks, the choices of thresholds and the norm of the residuals are of significant importance to the trade-off between false alarms, namely false positive alarms and false negative alarms that give rise to a class of stealthy deception attacks (see [131, 147], and the references therein for a review).*

Resilient cooperative control. Resilient cooperation refers to detaching from the detected set of malicious agents $\mathcal{A} \subset \mathcal{V}$ and convergence of the remaining cooperative mobile agents, $\mathcal{V} \setminus \mathcal{A}$, to a modified version of the cooperation objective in (3.1.2), which is defined as:

$$\lim_{t\to\infty} \left|\boldsymbol{p}_i(t) - \boldsymbol{p}_j(t) - \boldsymbol{p}_{ij}^{\star}\right| = \mathbf{0}, \qquad \forall\, i,j \in \mathcal{V}\setminus\mathcal{A}, \tag{5.1.8a}$$

$$\lim_{t\to\infty} |\boldsymbol{v}_i(t)| = \mathbf{0}, \qquad \forall\, i \in \mathcal{V}\setminus\mathcal{A}, \tag{5.1.8b}$$

for which the cooperative agents communicate over the induced network $\bar{\mathcal{G}}_{\sigma(t)} = (\bar{\mathcal{V}}, \bar{\mathcal{E}}_{\sigma(t)})$ defined[3] as

$$\bar{\mathcal{V}} := \mathcal{V}\setminus\mathcal{A}, \qquad \bar{\mathcal{E}}_{\sigma(t)} := \mathcal{E}_{\sigma(t)} \cap (\bar{\mathcal{V}}\times\bar{\mathcal{V}}). \tag{5.1.9}$$

Then, the problem of interest is to investigate under what conditions the resilient cooperation (5.1.8) over $\bar{\mathcal{G}}_{\sigma(t)}$ is achievable.

[3]We note that the communication network of the cooperative agents $\bar{\mathcal{V}}$ does not necessarily need to be an induced subgraph of $\mathcal{G}_{\sigma(t)}$ for which the communication links admit $\bar{\mathcal{E}}_{\sigma(t)} = \mathcal{E}_{\sigma(t)} \cap (\bar{\mathcal{V}}\times\bar{\mathcal{V}})$. It is possible to have designed and pre-programmed other communication typologies as part of a contingency plan upon attack detection. This, however, is a context-dependent problem and is outside the scope of this dissertation.

5.2 Network Resilience and Stability analysis

In this section, we investigate the network resilience to intermittent and permanent disconnections, as well as the stability and convergence of the multi-agent system in (5.1.1) with the unreliable communication network $\mathcal{G}_{\sigma(t)}$. In what follows, we present some assumptions on the communication network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ of the system in (5.1.1)-(5.1.2), allowing for the analysis of network resilience and system stability.

Assumption 5.2.1. *It is assumed that there exists a finite number of switches in any finite time interval. This allows us to rule out the Zeno phenomenon. Formally, there exists a finite sequence* $\{t_k\}_{k=0}^{\mathbf{m}} = t_0, \ldots, t_{\mathbf{m}}$, *where* $\mathbf{m} \in \mathbb{Z}_{\geq 0}$ *and* $\mathbf{m} > \mathbf{n}$ *in* (5.1.5), *that forms the set of* $\mathbf{m}$ *time instants in the ascending order of occurrence during any given time interval* $[t_0,\, t_0{+}T)$, *where* $t_0 \in \mathbb{R}_{\geq 0}$, *and* $T \in \mathbb{R}_{>0}$ *are defined such that* $T > (t_{\mathbf{m}} - t_0) \geq 0$. *Accordingly, the* $\mathbf{m}{+}1$ *(possibly unknown) modes* $\sigma(t_0), \sigma(t_1), \ldots, \sigma(t_{\mathbf{m}})$ *(*$\{\sigma(t_k) \in \mathcal{Q}' \subseteq \mathcal{Q},\ k \in \{0, \ldots, \mathbf{m}\}\}$*) denote the respective active modes of* $\Sigma_{\sigma(t)}$ *in* (5.1.1) *during the interval* $[t_0,\, t_0 + T)$.

Remark 5.2.1. *The switches, in Assumption 5.2.1, may include proactive (pre-programmed) and reactive topology switching, random link dropouts, and adversarial link dropouts. It is also worth mentioning that, the assumption of finitely many switching modes does not generally pose practical challenges since it is compliant with the nature of establishing communication links in the case of topology switching and random link dropouts. It is also known that in the case of adversarial disruption (e.g., DoS attacks), the attacker's capability is limited in terms of the frequency and duration of occurrence [34].*

Definition 5.2.1. ((μ, T)-PE connected communication network). A communication network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ is called (μ, T)-PE connected with some $T \in \mathbb{R}_{>0}$

and $\mu \in (0,\, N]$ if its associated Laplacian matrix $\mathsf{L}_{\sigma(t)}$ satisfies a (μ, T)-Persistence of Excitation (PE) condition of the form

$$\frac{1}{T}\int_{t}^{t+T} Q\mathsf{L}_{\sigma(\tau)}Q^{\top}\,\mathrm{d}\tau \geq \mu I_{N-1}, \quad \forall\, t \in \mathbb{R}_{\geq 0}, \tag{5.2.1}$$

where the matrix $Q \in \mathbb{R}^{(N-1)\times N}$ is defined such that

$$Q\mathbf{1}_N = \mathbf{0},\ QQ^{\top} = I_{N-1},\ Q^{\top}Q = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\top}. \tag{5.2.2}$$

Assumption 5.2.2. *The communication network $\mathcal{G}_{\sigma(t)}$ of the system $\Sigma_{\sigma(t)}$ in* (5.1.1) *is assumed to be (μ, T)-PE connected.*

Remark 5.2.2. *The (μ, T)-PE connectivity has appeared in the literature in different forms [32, 3]. It relaxes the strict point-wise in-time connectivity to the case of connectivity only in the integral sense of* (5.2.1)*, whereby positive algebraic connectivity in the integral sense that $\lambda_2(\frac{1}{T}\int_{t-T}^{t}\mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau) > \mu$ holds $\forall\, t \geq T$ and $\exists\, \mu, T \in \mathbb{R}_{>0}$ as in* (5.2.1)*. This relaxation allows for modeling a class of networks including periodically switching networks, [78], periodic with intermittent communications [113, 144], and jointly connected networks [108], as well as for quantifying resilience to the deception and DoS attacks defined in Section 5.1.3. See Proposition 5.2.6. Finally, the matrix $Q \in \mathbb{R}^{(N-1)\times N}$ in* (5.2.2) *can be recursively obtained as follows [32, rmk. 2]:*

$$Q_k = \begin{bmatrix} \sqrt{\frac{k-1}{k}} & -\frac{1}{\sqrt{k(k-1)}}\mathbf{1}_{k-1}^{\top} \\ \mathbf{0} & Q_k \end{bmatrix}, \tag{5.2.3a}$$

$$Q_k^{\top}Q_k = \begin{bmatrix} \frac{k-1}{k} & -\frac{1}{k}\mathbf{1}_{k-1}^{\top} \\ -\frac{1}{k}\mathbf{1}_{k-1} & Q_{k-1}^{\top}Q_{k-1} + \frac{1}{k(k-1)}\mathbf{1}_{k-1}\mathbf{1}_{k-1}^{\top} \end{bmatrix}, \tag{5.2.3b}$$

where $k \in \{2, \dots, N\}$ *with the initial matrix* $Q_2 = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$ *and the final matrix* $Q_N =: Q \in \mathbb{R}^{(N-1)\times N}$.

We next show the equivalence of the (μ, T)-PE condition presented herein and that in [3] for ensuring a positive *algebraic connectivity* in the integral sense. The equivalent conditions will later be used in the stability and robustness analyses, particularly in Theorem 5.2.5 and Lemma 5.3.2.

Lemma 5.2.3. ***(Equivalence of*** (μ, T)***-PE conditions for Network Connectivity).*** *Consider a* (μ, T)*-connected network* $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ *with the associated Laplacian matrix* $\mathsf{L}_{\sigma(t)}$. *The following statements are equivalent:*

1. *The condition* (5.2.1) *holds.*

2. *There exist* μ_m, μ_M, $T \in \mathbb{R}_{>0}$ *such that* $\forall\, t \in \mathbb{R}_{\geq 0}$,

$$\mu_m I_N \leq \frac{1}{T} \int_t^{t+T} \left(\mathsf{L}_{\sigma(\tau)} + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} \right) \mathrm{d}\tau \leq \mu_M I_N. \tag{5.2.4}$$

3. *There exist* $\delta, T \in \mathbb{R}_{>0}$ *such that the set of edges*

$$\boldsymbol{\mathcal{E}}_T^\mu = \big\{ (i,j) \in \mathcal{E}_{\sigma(t)} \mid \frac{1}{T} \int_t^{t+T} a_{ij}^{\sigma(\tau)} \, \mathrm{d}\tau \geq \delta, \\ \forall\, t \in \mathbb{R}_{\geq 0}, \quad i, j \in \mathcal{V},\, i \neq j \big\}, \tag{5.2.5}$$

forms a connected graph in the integral sense, denoted by $\boldsymbol{\mathcal{G}}_T^\mu = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^\mu)$, *where* $\frac{1}{T} \int_t^{t+T} a_{ij}^{\sigma(\tau)} \, \mathrm{d}\tau$*'s in* (5.2.5) *form the entries of the corresponding weighted adjacency matrix* $\mathbf{A}$ *and Laplacian matrix* $\mathbf{L}$ *that are defined as follows:*

$$\mathbf{A} = \frac{1}{T} \int_t^{t+T} \mathsf{A}_{\sigma(\tau)} \, \mathrm{d}\tau, \quad \mathbf{L} = \frac{1}{T} \int_t^{t+T} \mathsf{L}_{\sigma(\tau)} \, \mathrm{d}\tau. \tag{5.2.6}$$

Proof. See Appendix C.2.

We next show the relation between the (μ, T)-PE connectivity and the bounds on the vertex connectivity and robustness of graphs. The connectivity-related bounds provide a measure of the network resilience to intermittent and permanent disconnections. The former is associated with resilience to DoS attack in (5.1.5) and the latter is required for resilience to malicious agents though disconnecting from them (see (5.1.9)).

Definition 5.2.2. ((r, T)-robust network). A time-varying network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ is called (r, T)-robust[4] with some $T \in \mathbb{R}_{>0}$ and $r \in \mathbb{Z}_{\geq 0} \setminus \{0\}$ if the resultant static network $\boldsymbol{\mathcal{G}}_T^{\mu} = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^{\mu})$ with $\boldsymbol{\mathcal{E}}_T^{\mu}$ in (5.2.5), obtained under the (μ, T)-PE condition (5.2.1), is r-robust, where $r \leq r(\boldsymbol{\mathcal{G}}_T^{\mu})$.

Definition 5.2.3. ($(\boldsymbol{\kappa}, T)$-vertex-connected network). A time-varying network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ is called $(\boldsymbol{\kappa}, T)$-vertex-connected with some $T \in \mathbb{R}_{>0}$ and $\boldsymbol{\kappa} \in \mathbb{Z}_{\geq 0} \setminus \{0\}$ if the resultant static network $\boldsymbol{\mathcal{G}}_T^{\mu} = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^{\mu})$ with $\boldsymbol{\mathcal{E}}_T^{\mu}$ in (5.2.5), obtained under the (μ, T)-PE condition (5.2.1), is $\boldsymbol{\kappa}$-vertex-connected, where $\boldsymbol{\kappa} \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu})$.

Proposition 5.2.4. *Let $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ be a (μ, T)-PE connected network under Assumptions 5.2.1 and 5.2.2. Then, $\mathcal{G}_{\sigma(t)}$ is at least $(\lceil \frac{\mu}{2} \rceil, T)$-vertex-connected and $(\lceil \frac{\mu}{2} \rceil, T)$-robust, and the following inequalities hold for the resultant network $\boldsymbol{\mathcal{G}}_T^{\mu} = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^{\mu})$ with $\boldsymbol{\mathcal{E}}_T^{\mu}$ in (5.2.5).*

$$\left\lceil \frac{\hat{\mu}}{2} \right\rceil \leq r(\boldsymbol{\mathcal{G}}_T^{\mu}) \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu}) \leq |\mathcal{V}| - 1, \quad \hat{\mu} := \lambda_2(\mathbf{L}) \geq \mu, \tag{5.2.7}$$

[4]The (r, T)-robust network herein is robustness in an integral sense as a relaxation of the r-robust static network (cf. the discrete-time version in [144, Def. 2.2]). The (r, T)-robust in Definition 5.2.2 should not be confused by the notation of (r, s)-robustness, for some $r \in \mathbb{Z}_{\geq 0}$ and $1 \leq s \leq |\mathcal{V}|$, that is a strict generalization of r-robustness defined for a static graph [67].

where the robustness $r(\boldsymbol{\mathcal{G}}_T^{\mu})$, vertex connectivity $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu})$ are defined based on the adjacency and Laplacian matrices in (5.2.6)*. Additionally, if $\boldsymbol{\mathcal{G}}_T^{\mu}$ is a noncomplete, we have $\lambda_2(\mathbf{L}) \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu})$ ensuring that $\mathcal{G}_{\sigma(t)}$ is at least $(\lceil \mu \rceil, T)$-vertex-connected.*

Proof. See Appendix C.3.

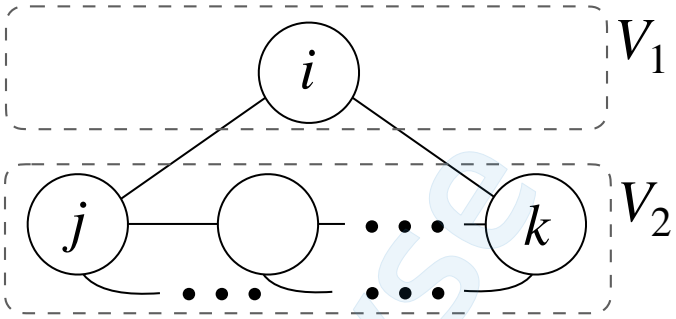


Figure 5.1: An example that illustrates how intermittent communication can drastically change the graph/network's algebraic connectivity $\lambda_2(\cdot)$ and thus its robustness. Let graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ such that $|\mathcal{V}| = N + 1$, with $\mathcal{V} = V_1 \cup V_2$ and $|V_2| = N$, where $N \geq 3$, and that the subgraph $\bar{\mathcal{G}}_{\sigma(t)} = (\mathcal{V} \setminus V_1, \bar{\mathcal{E}}_{\sigma(t)})$ induced by removing the set V_1 and its incident edges is a complete graph $\mathcal{K}_{|V_2|} = \bar{\mathcal{G}}_{\sigma(t)}$. Note that the singleton $i \in V_1$ can be connected to any pair of disjoint nodes $j \neq k \in V_2$, and thus $\mathcal{S} = \{j, k\} \subset \mathcal{V}$ and the bidirectional edge set $\mathcal{E}_{\text{cut}} = \{(i, j), (i, k)\}$ make, respectively, the minimum vertex cutset and edge cutset of $\mathcal{G}_{\sigma(t)}$. Accordingly, one can verify that $\lambda_2(\mathcal{G}_{\sigma(t)}) \leq \boldsymbol{\kappa}(\mathcal{G}_{\sigma(t)}) = \boldsymbol{e}(\mathcal{G}_{\sigma(t)}) = \boldsymbol{\delta}_{\min}(\mathcal{G}_{\sigma(t)}) = 2$, where $\boldsymbol{e}(\cdot)$ and $\boldsymbol{\delta}_{\min}(\cdot)$ are, resp., the edge connectivity and minimum node-degree. Also, if $\exists\, t \in \mathbb{R}_{\geq 0}$ s.t $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E} \setminus \mathcal{E}_{\text{cut}})$ because of an intermittent connection of the edges $\mathcal{E}_{\text{cut}}$, we have graph disconnection with $\lambda_2(\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E} \setminus \mathcal{E}_{\text{cut}})) = 0$. Yet, the induced subgraph $\mathcal{K}_{|V_2|}$ holds even a higher algebraic connectivity since $\lambda_2(\mathcal{K}_{|V_2|}) = |V_2| = N$, and $\boldsymbol{\kappa}(\mathcal{K}_{|V_2|}) = \boldsymbol{e}(\mathcal{K}_{|V_2|}) = \boldsymbol{\delta}_{\min}(\mathcal{K}_{|V_2|}) = N - 1$. This example has been constructed based on the discussions in [48, Ch. 13.5].

Theorem 5.2.5. ***(Network resilience to node and edge disconnections).*** *Let a (μ, T)-PE connected network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ be (r, T)-robust (resp. $(\boldsymbol{\kappa}, T)$-vertex-connected) under Assumptions 5.2.1 and 5.2.2. Let $\mathcal{A} \subset \mathcal{V}$ be a $(r-1)$-local (resp. $(\boldsymbol{\kappa} - 1)$-total) adversary set. Then, the induced subgraph $\bar{\mathcal{G}}_{\sigma(t)} = (\bar{\mathcal{V}} = \mathcal{V} \setminus \mathcal{A}, \bar{\mathcal{E}}_{\sigma(t)})$ in* (5.1.9) *admits the $(\bar{\mu}, \bar{T})$-PE connectivity condition in* (5.2.1)*, for some $\bar{\mu}, \bar{T} \in \mathbb{R}_{>0}$, where $\bar{T} \leq T$ and $\mu \leq \bar{\mu} + |\mathcal{A}|$.*

Proof. See Appendix C.4.

In other words, Theorem 5.2.5 together with (5.2.7) implies that if the network $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ holds algebraic connectivity $\lambda_2(\mathbf{L}) \geq \mu \geq 2F + \epsilon$ with $F \in \mathbb{Z}_{\geq 0}$ and $\epsilon \in \mathbb{R}_{>0}$ in the integral sense of (5.2.1), it will be at least $(F+1, T)$-vertex-connected (resp. $(F+1, T)$-robust) which in turn ensures the resulting network from the removal of an F-total (resp. F-local) adversary set, i.e. (5.1.9), still maintains its connectivity in the sense of (5.2.1) to a lower degree, allowing to achieve (5.1.8).

Remark 5.2.3. *We remark that we use the parameter μ in* (5.2.1) *and* (5.2.7) *as a rather conservative proxy metric of the resilience of time-varying networks to disconnections of sorts (i.e. permanent and/or intermittent). Moreover, $\hat{\mu}$ in* (5.2.7) *can be interpreted as the quality of service (QoS) [32] associated with communication. It is also noteworthy that the lower bound $\left\lceil \frac{1}{2}\lambda_2(\mathbf{L}) \right\rceil \leq r(\boldsymbol{\mathcal{G}}_T^{\mu})$ in* (5.2.7) *is tight as shown in [120, lemma 1], [116, Thm. 2] for fixed graphs. On the other hand, the gap between $r(\boldsymbol{\mathcal{G}}_T^{\mu})$ and $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu})$ can be arbitrarily large ([67, 120]) depending on a priori unknown intermediary typologies $\mathcal{G}_\sigma$'s, $\sigma \in \mathcal{Q}'$ that form a network $\boldsymbol{\mathcal{G}}_T^{\mu}$. We refer to Fig. 5.1 as an illustrative example that demonstrates how an intermittent connection of edges in a class of graphs can affect the bounds in* (5.2.7). *Moreover, in the special case of complete graphs over N nodes, denoted by $\mathcal{K}_N$, we have $\lceil \lambda_2(\mathcal{K}_3)/2 \rceil = \boldsymbol{\delta}_{\min}(\mathcal{K}_3) = 2$ for $N = 3$ nodes, that shows the bound in* (5.2.7) *is tight. If the exclusion of complete graphs can be guaranteed the lower bound to $\hat{\mu} \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu})$ can be used for node connectivity. (cf. [48, Cor. 13.5.2]).*

We now provide a convergence bound for the consensus/formation equilibrium in (5.1.8). Associated with (5.1.8), we define an output (coordinate) vector $\mathcal{Y} \in \mathbb{R}^{2N-1}$

as

$$\mathcal{Y} = \begin{bmatrix} \zeta \\ \boldsymbol{v} \end{bmatrix} := \begin{bmatrix} Q & \mathbf{0}_{(N-1)\times N} \\ \mathbf{0}_N & I_N \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{p}} \\ \boldsymbol{v} \end{bmatrix} = \mathbf{C}_{eq}\mathbf{x}, \tag{5.2.8}$$

where $\widetilde{\boldsymbol{p}}$, $\boldsymbol{v}$ and Q are given in (5.1.1) and (5.2.2), respectively. It then follows that $\zeta = Q\widetilde{\boldsymbol{p}} = \mathbf{0}_{N-1}$ and $\boldsymbol{v} = \mathbf{0}_N$ imply $\widetilde{\boldsymbol{p}}_i - \widetilde{\boldsymbol{p}}_j = 0,\ \forall\, i, j \in \mathcal{V}$ and $\boldsymbol{v}_i = 0, \forall\, i \in \mathcal{V}$ (that is $\mathcal{Y}(t) \to \mathbf{0} \equiv$ (5.1.8)).

Proposition 5.2.6. ***(Stability under*** (μ, T)***-PE connectivity).*** *Consider the system in* (5.1.1) *and let Assumptions 5.2.1 and 5.2.2 hold. Also, let* $\sum_{k\in\mathbb{Z}_{\geq 0}}^{\mathbf{n}} T_k^{\mathrm{a}} < T$*, where* T *is the period in* (5.2.1)*, hold uniformly in time for the DoS attack in* (5.1.5)*. Then, there exists a sufficiently large control gain* γ *for* (3.1.3) *that ensures, for each* $\|\mathbf{x}(t_0)\| \leq \infty$ *and for every* $\mathbf{u}_{\mathcal{A}} \in \mathcal{L}_{pe}$ *with* $\sup_{t_0\leq t\leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| < \infty$*, the system* (5.1.1) *with the output* $\mathcal{Y}(t)$ *in* (5.2.8) *is finite-gain* $\mathcal{L}_p$ *stable with the following upper bound:*

$$\|\mathcal{Y}(t)\|_{\mathcal{L}_p} \leq \kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\| + \kappa_{\mathbf{u}} \|(\mathbf{u}_{\mathcal{A}})_{T_d}\|_{\mathcal{L}_p}, \quad \forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}, \tag{5.2.9a}$$

$$\kappa_{\mathbf{x}} = \|C\| \sqrt{\frac{\max\left\{\lambda_\chi^{-1}, \beta\right\}}{\min\left\{\frac{\gamma}{\alpha N}, \beta\right\}}} \|C^{-1}\|, \qquad \kappa_{\mathbf{u}} = \|C\| \frac{\max\left\{\lambda_\chi^{-1}, \beta\right\}}{\lambda_{\mathbf{x}} \min\left\{\frac{\gamma}{2\alpha N}, \frac{\beta}{2}\right\}}, \tag{5.2.9b}$$

$$0 < \lambda_{\mathbf{x}} < \lambda_\chi = \eta e^{-2\eta T}, \qquad C = \begin{bmatrix} \frac{1}{\gamma} I_{N-1} & -\frac{1}{\gamma} Q \\ \mathbf{0}_{N\times(N-1)} & I_N \end{bmatrix}, \tag{5.2.9c}$$

where $\eta = -\frac{1}{2T} \ln(1 - \frac{(\alpha/\gamma)\mu T}{1+(\alpha/\gamma)^2 N^2 T^2})$ *and* $\beta \in \mathbb{R}_{>0}$*. Additionally, if* $\mathbf{u}_{\mathcal{A}} = \mathbf{0}$ *(or* $\mathcal{A} = \emptyset$*), the system's state trajectories uniformly exponentially converge to the equilibrium*

(5.1.8), *(provided the formation configuration is feasible) with the bound*

$$\left|\boldsymbol{p}_i(t) - \boldsymbol{p}_j(t) - \boldsymbol{p}_{ij}^{\star}\right| \leq \sqrt{2}\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \left\|\mathbf{x}(t_0)\right\|, \tag{5.2.10a}$$

$$\left|\boldsymbol{v}_i(t)\right| \leq \kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \left\|\mathbf{x}(t_0)\right\|, \quad \forall\, i, j \in \mathcal{V} \tag{5.2.10b}$$

for all $t \geq t_0 \in \mathbb{R}_{\geq 0}$*, with* $\kappa_{\mathbf{x}}$, $\lambda_{\mathbf{x}} \in \mathbb{R}_{>0}$ *as given in* (5.2.9).

Proof. See Appendix C.5.

We remark that the choice of $\gamma = \alpha N$, for $\alpha \geq 1$ yields a convergence rate λ_χ in (5.2.9) that depends only on μ, T, with the maximum occurring at $\mu = N, T = 1$, associated with complete network connectivity, see (5.2.1) and Remark 5.2.2. This, however, is not the only valid choice.

We also note to [30, 31] provide insightful results on the stability of multi-agent systems using passivity-based control and contraction theory.

5.3 Observer Design and Attack Detection

Here, we consider the design of observers serving the reconfigurable local attack detector module $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i}_{\sigma}{}''}$ in Section 5.1.4. The observer design for $\Sigma_{\sigma(t)}$ in (5.1.1) is subject to two constraints. First, *a priori* full knowledge of $\mathbf{A}_\sigma$ may not be available for each mobile agent due to random communication link dropouts or switching links. Second, local state information $\mathbf{y}^i_\sigma$ in (5.1.2), which is available for each mobile agent, is subject to change since the respective k-hop neighbors change in an *a priori* unknown time-varying network. Consequently, ensuring the uniform observability of $(\mathbf{A}_\sigma, \mathbf{C}^i_\sigma)$, $\forall\, t \in \mathbb{R}_{\geq 0}$, $\forall\, i \in \mathcal{V}$ may not be tractable or feasible.

In what follows, we, first, characterize network-level conditions under which almost any set of adversarial inputs $\mathbf{u}_{\mathcal{A}}$ is observable at the measurements of the

cooperative agents that are $\mathbf{y}_\sigma^i$'s, for $i \in \mathcal{V} \setminus \mathcal{A}$. Second, we propose a class of local observers for $\Sigma^{\mathcal{O}}_{\mathcal{V}_\sigma^{i''}}$ that are realizable using (5.1.6), enabling distributed attack detection through (5.1.7).

5.3.1 Detectability of Adversarial Inputs

We note that the switched system $\Sigma_{\sigma(t)}$ in (5.1.1)-(5.1.2) represents a family of linear time-invariant (LTI) systems, each of which is associated with one mode $\sigma(t) \in \mathcal{Q}$. Therefore, similar to [78, 97], the results herein are derived based on the concepts of output-zeroing and *state and input observability* of the (switched) LTI systems.

Consider a generic solution $\mathbf{x}(t; \mathbf{x}(t_0), \mathbf{u}_\mathcal{A}(t))$ to $\Sigma_{\sigma(t)}$ in (5.1.1) under Assumptions 5.2.1 and 5.2.2. Then, the concatenation of the measurements $\mathbf{y}_\sigma^i$'s, given in (5.1.2), of the set of cooperative agents, $\mathcal{V} \setminus \mathcal{A} = \{i_1, \ldots, i_{|\mathcal{V}\setminus\mathcal{A}|}\}$, is defined as

$$\begin{aligned} \mathbf{y}_\sigma^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}(t_0), \mathbf{u}_\mathcal{A}(t)) = \mathrm{col}(\mathbf{y}_\sigma^{i_1}, \ldots, \mathbf{y}_\sigma^{i_{|\mathcal{V}\setminus\mathcal{A}|}}) = \\ \mathrm{col}(\mathbf{C}_\sigma^{i_1}, \ldots, \mathbf{C}_\sigma^{i_{|\mathcal{V}\setminus\mathcal{A}|}}) \mathbf{x}(t; \mathbf{x}(t_0), \mathbf{u}_\mathcal{A}(t)) = \\ \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}}\, \mathbf{x}(t; \mathbf{x}(t_0), \mathbf{u}_\mathcal{A}(t)). \end{aligned} \tag{5.3.1}$$

It is necessary to note that the entirety of measurement $\mathbf{y}_\sigma^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}(t_0), \mathbf{u}_\mathcal{A}(t))$ is not available for any agent $i \in \mathcal{V}$. We use this collective set of the measurements of cooperative agents $\mathcal{V} \setminus \mathcal{A}$ and a generic set of adversarial inputs $\mathbf{u}_\mathcal{A}$ introduced by the set of malicious agents $\mathcal{A}$, in an input observability context for attack detection analyses.

Definition 5.3.1. (Stealthy and Indistinguishable Attacks). For $\Sigma_{\sigma(t)}$ in (5.1.1) under Assumptions 5.2.1 and 5.2.2, any generic set of inputs $\mathbf{u}_\mathcal{A}(t) \in \mathcal{L}_{pe}$ injected by

a set of malicious agents $\mathcal{A}$ is stealthy for the remaining cooperative agents $\mathcal{V} \setminus \mathcal{A}$ if

$$\begin{aligned} &\exists\, \mathbf{x}(t_0), \mathbf{x}'(t_0) \in \mathbb{R}^{2N} \ \text{ s.t. } \ \forall\, t \in [t_0,\ t_0 + T), \\ &\mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}(t_0), \mathbf{u}_{\mathcal{A}}(t)) = \mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}'(t_0), \mathbf{0}), \end{aligned} \tag{5.3.2}$$

with $\mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t;\cdot,\cdot)$ as in (5.3.1), $t_0 \in \mathbb{R}_{\geq 0}$, and $T \in \mathbb{R}_{>0}$ as in (5.2.1). Likewise, for any given two generic sets $\mathbf{u}_{\mathcal{A}_1}(t) \in \mathcal{L}_{pe}$ and $\mathbf{u}_{\mathcal{A}_2}(t) \in \mathcal{L}_{pe}$ injected, resp., by a nonempty set of malicious agents $\mathcal{A}_1 \in \mathcal{A}$ and some $\mathcal{A}_2 \in \mathcal{A}$, where $\mathcal{A}_1 \neq \mathcal{A}_2$, $\mathbf{u}_{\mathcal{A}_1}$ is indistinguishable from $\mathbf{u}_{\mathcal{A}_2}$ for the cooperative agents $\mathcal{V} \setminus \mathcal{A}$ if

$$\begin{aligned} &\exists\, \mathbf{x}(t_0), \mathbf{x}'(t_0) \in \mathbb{R}^{2N} \ \text{ s.t. } \ \forall\, t \in [t_0,\ t_0 + T), \\ &\mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}(t_0), \mathbf{u}_{\mathcal{A}_1}(t)) = \mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}'(t_0), \mathbf{u}_{\mathcal{A}_2}(t)). \end{aligned} \tag{5.3.3}$$

Lemma 5.3.1. ***(Characterization of Stealthy and Indistinguishable Attacks).*** *Consider $\Sigma_{\sigma(t)}$ in (5.1.1) and let Assumptions 5.2.1 and 5.2.2 hold. Also, let $\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(t)$ and $\mathbf{B}_{\mathcal{A}_2}\mathbf{u}_{\mathcal{A}_2}(t)$ be two generic sets of adversarial $\mathcal{L}_{pe}$-norm bounded inputs injected by a nonempty set of malicious agents $\mathcal{A}_1 \in \mathcal{A}$ and some $\mathcal{A}_2 \in \mathcal{A}$, where $\mathcal{A}_1 \neq \mathcal{A}_2$. Then, $\mathbf{u}_{\mathcal{A}_1}(t)$ and $\mathbf{u}_{\mathcal{A}_2}(t)$ are indistinguishable for the cooperative agents $\mathcal{V} \setminus \mathcal{A}$ during $t \in [t_0,\, t_0 + T)$, if and only if $\exists\, \mathrm{x}(t_0) \in \mathbb{R}^{2N}$ such that*

$$\begin{aligned} &\mathbf{C}_{\sigma(t_k)}^{\mathcal{V}\setminus\mathcal{A}} e^{\mathbf{A}_{\sigma(t_k)}(t-t_k)} \mathrm{x}(t_k) = \\ &\mathbf{C}_{\sigma(t_k)}^{\mathcal{V}\setminus\mathcal{A}} \int_{t_k}^{t} e^{\mathbf{A}_{\sigma(t_k)}(t-\tau)} \big(\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(\tau) - \mathbf{B}_{\mathcal{A}_2}\mathbf{u}_{\mathcal{A}_2}(\tau)\big)\, \mathrm{d}\tau, \quad t \in\ [t_k,\, t_{k+1}), \end{aligned} \tag{5.3.4}$$

where $0 \leq k \leq \mathbf{m}$ and $t_{\mathbf{m}+1} =: t_0 + T$, with $\mathbf{m}$ and T as in Assumptions 5.2.1 and 5.2.2, and $\mathrm{x}(t_0) = (\mathbf{x}'(t_0) - \mathbf{x}(t_0))$ when $k = 0$, and for $k \neq 0$, we have $\mathrm{x}(t_k) =$

$\mathbf{x}'(t_k) - \mathbf{x}(t_k)$, *where*

$$\mathrm{x}(t_k) = \prod_{i=k}^{1} e^{\mathbf{A}_{\sigma(t_{i-1})}(t_i - t_{i-1})}\mathrm{x}(t_0) + \sum_{i=1}^{k}\prod_{j=k}^{i+1} e^{\mathbf{A}_{\sigma(t_{j-1})}(t_j - t_{j-1})} \int_{t_{i-1}}^{t_i} e^{\mathbf{A}_{\sigma(t_{i-1})}(t_i - \tau)}\big(\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(\tau) - \mathbf{B}_{\mathcal{A}_2}\mathbf{u}_{\mathcal{A}_2}(\tau)\big)\,\mathrm{d}\tau\big). \tag{5.3.5}$$

Additionally, if (5.3.4)-(5.3.5) *hold for* $\mathbf{u}_{\mathcal{A}_2}(t) = \mathbf{0}$, $\forall\, t \in [t_0,\, t_0 + T)$, *then* $\mathbf{u}_{\mathcal{A}_1}(t)$ *is stealthy.*

Proof. See Appendix C.6.

It is necessary to note that the realization of (5.3.4) requires a priori knowledge of the system which is not available for any agent. Moreover, based on the concepts of *state and input observability* [15, Thm. 2], [152, Ch. 3.11], and the invariant zeros of the switched LTI systems [78], the realizations of (5.3.4) in each mode coincide with the existence of the set of vector-valued adversarial input $\mathbf{u}_{\mathcal{A}}$ unobservable at the vector-valued output $\mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}$ (see Definition 2.3.1). For LTI systems, it is well-known that such a set of inputs, referred to as zero-dynamics attacks (see [78, 97] and Section 5.1.3), is not generic, and is characterized using the output-zeroing directions of the system. Particularly, for $\Sigma_{\sigma(t)}$ *in each mode* $\sigma \in \mathcal{Q}$, it follows from [152, Ch. 3.11] that the output-zeroing directions are induced by the rank deficiencies of the matrix pencil $\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma)$ for some $\lambda_{\mathrm{o}} \in \mathbb{C}$, where

$$\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma) = \begin{bmatrix} \lambda_{\mathrm{o}} I - \mathbf{A}_{\sigma} & -\mathbf{B}_{\mathcal{A}} \\ \mathbf{C}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}} & \mathbf{0} \end{bmatrix}. \tag{5.3.6}$$

We next present conditions under which the intersection of the output-zeroing subspaces of $\Sigma_{\sigma(t)}$ in (5.1.1) make an empty set, ensuring almost no deception attacks, defined in Section 5.1.3, can be stealthy in the sense of (5.3.2).

Lemma 5.3.2. *Consider* $\Sigma_{\sigma(t)}$ *in* (5.1.1)-(5.1.2) *with* (5.1.6) *during an interval* $[t_0,\, t_0+T)$ *defined under Assumptions 5.2.1 and 5.2.2. Let* $\Sigma_{\sigma(t)}$ *be subject to any generic set of adversarial inputs* $\mathbf{u}_{\mathcal{A}}(t) \in \mathcal{L}_{pe}$ *injected by an* F*-total (resp.* F*-local) set* $\mathcal{A}$ *of malicious agents such that* $0 \leq F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^{\mu}) - 1$ *(resp.* $0 \leq F \leq r(\boldsymbol{\mathcal{G}}_T^{\mu}) - 1$*). Then, the following statements are equivalent.*

1. *There exists no generic set of inputs* $\mathbf{u}_{\mathcal{A}}(t) \in \mathcal{L}_{\infty e}$ *stealthy in the sense of* (5.3.2).
2. *For almost all* $\lambda_{\mathrm{o}} \in \mathbb{C}$, $\cap_{\sigma\in\mathcal{Q}'} \ker\left(\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma)\right) = \emptyset$, *where* $\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma)$ *is given by* (5.3.6).

Proof. See Appendix C.7.

In other words, Lemma 5.3.2 states that $(F+1, T)$-vertex connectivity (resp. $(F+1, T)$-robustness), where $F \in \mathbb{Z}_{\geq 0}$, ensures almost no F-total (resp. F-local) set of malicious agents with the deception attacks defined in Section 5.1.3 is stealthy in the sense of (5.3.2) for the cooperative agents in (5.1.1) with (5.1.6). (cf. [97, 37] where $(F+1)$-vertex connectivity and $(2F+1)$-robustness are required point-wise in time.)

We next investigate the level of local observability for each agent given the locally available information $\Phi^i_{\sigma(t)}$ in (5.1.6) and measurements $\mathbf{y}^i_\sigma$ in (5.1.2), as opposed to ensuring the global observability of the pair $(\mathbf{A}_\sigma, \mathbf{C}^i_\sigma)$ associated with (5.1.1)-(5.1.2) that might not be tractable.

5.3.2 Local Dynamics and Observability Analysis

Consider the set of 2-hop information available for each agent $i \in \mathcal{V}$ as defined by $\Phi^i_{\sigma(t)}$ in (5.1.6), and local measurements $\mathbf{y}^i_\sigma$ in (5.1.2). Let $\mathcal{I}_i = \mathcal{V}^{i''}_\sigma$ ($\mathcal{I}$ in short) in (5.1.2) and $\mathcal{R}_i = \mathcal{V} \setminus \mathcal{V}^{i''}_\sigma$ and assume $\mathcal{I}_i$ and $\mathcal{R}_i$ ($\mathcal{R}$ in short) are sorted in the

ascending order of agents' indices. Then, by using (5.1.4), the $\Sigma_{\sigma(t)}$'s dynamics in (5.1.1) with $\mathbf{y}_\sigma^i$ in (5.1.2) can be partitioned as

$$\Sigma_{\mathcal{V}_\sigma^{i''}} : \begin{cases} \dot{\mathbf{x}}_\mathcal{I} = \mathbf{A}_\sigma^\mathcal{I}\mathbf{x}_\mathcal{I} + \rho(\mathbf{x}_\mathcal{I}, \mathbf{x}_\mathcal{R}) + \mathbf{B}_{\mathcal{A}''}\mathbf{u}_{\mathcal{A}''}, \\ \mathbf{y}_\sigma^i = \mathbf{C}_\sigma^\mathcal{I}\mathbf{x}_\mathcal{I} \end{cases} \tag{5.3.7}$$

$$\Sigma_{\mathcal{R}_i} : \ \dot{\mathbf{x}}_\mathcal{R} = \mathbf{A}_\sigma^\mathcal{R}\mathbf{x}_\mathcal{R} + \mathbf{A}_\sigma^{\mathcal{R},\mathcal{I}}\mathbf{x}_\mathcal{I} + \mathbf{B}_{\mathcal{A}^\mathrm{r}}\mathbf{u}_{\mathcal{A}^\mathrm{r}}, \tag{5.3.8}$$

where $\mathbf{x}_\bullet \coloneqq \mathrm{col}\left(\widetilde{\boldsymbol{p}}_\bullet, \boldsymbol{v}_\bullet\right)$, $\bullet \in \{\mathcal{I}, \mathcal{R}\}$ denotes the position and velocity states of the agents in each set, and the system matrices are defined as

$$\mathbf{A}_\sigma^\mathcal{I} = \begin{bmatrix} \mathbf{0}_{|\mathcal{I}|\times|\mathcal{I}|} & I_{|\mathcal{I}|} \\ -\alpha\mathsf{L}_\sigma'' & -\gamma I_{|\mathcal{I}|} \end{bmatrix}, \ \mathbf{B}_{\mathcal{A}''} = \begin{bmatrix} \mathbf{0}_{|\mathcal{I}|\times|\mathcal{A}''|} \\ \hline I_{\mathcal{A}''} \end{bmatrix}, \tag{5.3.9a}$$

$$\mathbf{C}_\sigma^\mathcal{I} = \mathrm{diag}(I_{|\mathcal{I}|}, \mathfrak{e}_{|\mathcal{I}|}^{1\,\top}), \tag{5.3.9b}$$

$$\rho(\mathbf{x}_\mathcal{I}, \mathbf{x}_\mathcal{R}) = \widetilde{\widetilde{\mathbf{A}}}_\sigma^\mathcal{I}\mathbf{x}_\mathcal{I} + \mathbf{A}_\sigma^{\mathcal{I},\mathcal{R}}\mathbf{x}_\mathcal{R} = \begin{bmatrix} \mathbf{0}_{|\mathcal{I}|\times 1} \\ \hline \\ \mathbf{0}_{|\mathcal{V}_\sigma^{i'}|\times 1} \\ \underline{\rho} \end{bmatrix}, \ \underline{\rho} = -\alpha\big(\widetilde{\widetilde{\mathsf{L}}}_\sigma \widetilde{\boldsymbol{p}}_{\mathcal{N}_\sigma^{i(2)}} + \mathsf{L}_\sigma^{(23)}\widetilde{\boldsymbol{p}}_\mathcal{R}\big), \tag{5.3.9c}$$

$$\widetilde{\widetilde{\mathbf{A}}}_\sigma^\mathcal{I} = \left[\begin{array}{cc|c} \multicolumn{2}{c|}{\mathbf{0}_{|\mathcal{I}|\times|\mathcal{I}|}} & \mathbf{0}_{|\mathcal{I}|} \\ \hline \mathbf{0} & \mathbf{0} & \\ \mathbf{0} & -\alpha\widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathbf{0}_{|\mathcal{I}|} \end{array}\right], \ \mathbf{A}_\sigma^{\mathcal{I},\mathcal{R}} = \left[\begin{array}{c|c} \mathbf{0}_{|\mathcal{I}|\times|\mathcal{R}|} & \mathbf{0}_{|\mathcal{I}|\times|\mathcal{R}|} \\ \hline \mathbf{0}_{|\mathcal{V}_\sigma^{i'}|\times|\mathcal{R}|} & \mathbf{0} \\ -\alpha\mathsf{L}_\sigma^{(23)} & \end{array}\right], \tag{5.3.9d}$$

$$\mathbf{A}_\sigma^\mathcal{R} = \begin{bmatrix} \mathbf{0} & I_{|\mathcal{R}|} \\ -\alpha\mathsf{L}_\sigma^{(33)} & -\gamma I_{|\mathcal{R}|}, \end{bmatrix}, \ \mathbf{A}_\sigma^{\mathcal{R},\mathcal{I}} = \left[\begin{array}{cc|c} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & -\alpha\mathsf{L}_\sigma^{(32)} & \mathbf{0} \end{array}\right], \tag{5.3.9e}$$

$$\mathcal{A}'' = \mathcal{A} \cap \mathcal{V}_\sigma^{i''}, \qquad \mathcal{A}^\mathrm{r} = \mathcal{A} \setminus \mathcal{A}''. \tag{5.3.9f}$$

where $\mathbf{u}_{\mathcal{A}''} = \mathrm{col}\left(\boldsymbol{u}_i^\mathrm{a}\right)_{i\in\mathcal{A}''} \in \mathbb{R}^{|\mathcal{A}''|}$, $I_{\mathcal{A}''} = \left[\mathfrak{e}_{|\mathcal{I}|}^{i_1}\ \mathfrak{e}_{|\mathcal{I}|}^{i_2}\ \ldots\ \mathfrak{e}_{|\mathcal{I}|}^{i_{|\mathcal{A}''|}}\right] \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{A}''|}$.

We note that $\Sigma_{\mathcal{V}_i''}$ with known $\mathbf{A}_\sigma^{\mathcal{I}}$ and $\mathbf{C}_\sigma^{\mathcal{I}}$, is the dynamics available for each agent $i \in \mathcal{V}$ given $\Phi^i_{\sigma(t)}$, $\forall\, t \in \mathbb{R}_{\geq 0}$, and the dynamics $\Sigma_{\mathcal{R}_i}$ and the possibly existing coupling term $\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})$ are unknown to the agent $i \in \mathcal{V}$. Moreover, for every agent $i \in \mathcal{V}$, $\mathbf{y}_\sigma^i$ in (5.3.7) and (5.1.2) are the same set of measurements obtained by reordering $\mathbf{C}_\sigma^{\mathcal{I}}$ and $\mathbf{C}_\sigma^i$.

The following results address the effect of $\Sigma_{\mathcal{R}_i}$ on $\Sigma_{\mathcal{V}_\sigma^{i\prime\prime}}$ as well as the observability of $\Sigma_{\mathcal{V}_\sigma^{i\prime\prime}}$, which will be used later in local observer design.

Proposition 5.3.3. *Consider* (5.3.7) *and* (5.3.8) *under Assumptions 5.2.1 and 5.2.2. The coupling term* $\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})$ *in* $\Sigma_{\mathcal{V}_\sigma^{i\prime\prime}}$ *of agent* $i \in \mathcal{V}$*, holds the bound* $\|\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\| \leq \alpha\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\| + \alpha\kappa_{\mathbf{u}} \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\|$, $\forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}$*, where* α *is given in* (3.1.3)*, and* $\kappa_{\mathbf{x}}$*,* $\kappa_{\mathbf{u}}$*, and* $\lambda_{\mathbf{x}}$ *are given in* (5.2.9)*. Additionally, if* $\mathcal{A} = \emptyset$*, the coupling term exponentially converges to* $\mathbf{0}$ *with* $\|\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\| \leq \alpha\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\|$, $\forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}$.

Proof. See Appendix C.8.

Proposition 5.3.4. *Consider the 2-hop dynamics* $\Sigma_{\mathcal{V}_\sigma^{i\prime\prime}}$ *in* (5.3.7) *for each agent* $i \in \mathcal{V} \setminus \mathcal{A}$ *communicating over* $\mathcal{G}_{\sigma(t)}$ *under Assumptions 5.2.1 and 5.2.2. Then, the following statements hold.*

1. *the pair* $(\mathbf{A}_\sigma^{\mathcal{I}}, \mathbf{C}_\sigma^{\mathcal{I}})$ *in* $\Sigma_{\mathcal{V}_\sigma^{i\prime\prime}}$, $\forall\, i \in \mathcal{V}$*, is observable in each mode* $\sigma \in \mathcal{Q}$*.*

2. *There exists no generic set of inputs* $\mathbf{u}_{\mathcal{A}}(t) \in \mathcal{L}_{pe}$ *stealthy in the sense of* (5.3.2)*, where* $\mathbf{y}_\sigma^i$*'s are given in* (5.3.7)*, provided the set* $\mathcal{A}$ *of malicious agents is* F*-total (resp.* F*-local), with* $0 \leq F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) - 1$ *(resp.* $0 \leq F \leq r(\boldsymbol{\mathcal{G}}_T^\mu) - 1$*).*

Proof. See Appendix C.9.

Having quantified the conditions on the attack stealthiness and local observability, we next propose the reconfigurable local attack detector module that relies

only on the time-varying local information $\Phi^i_{\sigma(t)}$ in (5.1.6) and that performs the distributed hypothesis testing in (5.1.7).

5.3.3 Reconfigurable Attack Detector (local observer)

For each agent $i \in \mathcal{V}$ with the 2-hop dynamics $\Sigma_{\mathcal{V}_i''}$ in (5.3.7) and local information $\Phi^i_{\sigma(t)}$ in (5.1.6), the local attack detector $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}$ is proposed as follows

$$\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}: \begin{cases} \dot{\hat{\mathbf{x}}}_{\mathcal{I}} = \mathbf{A}^{\mathcal{I}}_\sigma \hat{\mathbf{x}}_{\mathcal{I}} + \mathbf{H}^{\mathcal{I}}_\sigma (\mathbf{y}^i_\sigma - \hat{\mathbf{y}}^i_\sigma) \\ \hat{\mathbf{y}}^i_\sigma = \mathbf{C}^{\mathcal{I}}_\sigma \hat{\mathbf{x}}_{\mathcal{I}} \\ \mathbf{r}^i_\sigma = \mathbf{y}^i_\sigma - \hat{\mathbf{y}}^i_\sigma \end{cases}, \tag{5.3.10a}$$

$$\hat{\mathbf{x}}_{\mathcal{I}}(t_k) = \begin{cases} \mathbb{I}_{\mathcal{I}_i} \mathbf{x}_{\mathcal{I}}(t_k), \ \text{ if } \ \mathcal{V}^{i''}_{\sigma(t_k)} \neq \mathcal{V}^{i''}_{\sigma(t_{k-1})} \text{ OR } k = 0, \\ \hat{\mathbf{x}}_{\mathcal{I}}(t_k), \quad \text{ if } \ \mathcal{V}^{i''}_{\sigma(t_k)} = \mathcal{V}^{i''}_{\sigma(t_{k-1})}, \end{cases} \tag{5.3.10b}$$

where $\hat{\mathbf{x}}_{\mathcal{I}}$ is the estimation of $\mathbf{x}_{\mathcal{I}}$ in (5.3.7), and the initial conditions $\hat{\mathbf{x}}_{\mathcal{I}}(t_k)$ are updated at $\{t_k\}^{\mathbf{m}}_{k=0}$, $\mathbf{m} \in \mathbb{Z}_{\geq 0}$ corresponding to the modes $\sigma(t_k)$'s $\in \mathcal{Q}$ (see Assumption 5.2.1), and $\mathbb{I}_{\mathcal{I}_i} = \text{diag}(I_{|\mathcal{I}_i|}, \mathfrak{e}^1_{|\mathcal{I}_i|} \mathfrak{e}^{1\top}_{|\mathcal{I}_i|})$. $\mathbf{H}^{\mathcal{I}}_\sigma = \begin{bmatrix} \mathbf{0}_{|\mathcal{I}|\times|\mathcal{I}|} & \mathbf{0}_{|\mathcal{I}|} \\ H^{\mathcal{I}}_\sigma & h_\sigma \mathfrak{e}^1_{|\mathcal{I}|} \end{bmatrix}$ is the observer's gain matrix with a scalar $h_\sigma \in \mathbb{R}_{>0}$ and a symmetric positive definite matrix $H^{\mathcal{I}}_\sigma \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}_{>0}$ such that $\bar{\mathbf{A}}^{\mathcal{I}}_\sigma = (\mathbf{A}^{\mathcal{I}}_\sigma - \mathbf{H}^{\mathcal{I}}_\sigma \mathbf{C}^{\mathcal{I}}_\sigma)$ is Hurwitz stable in every mode $\sigma \in \mathcal{Q}$. Note that the availability of $\Phi^i_{\sigma(t)}$ in (5.1.6) allows each agent to readily update $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_\sigma}$ upon a switch occurs between the communication modes.

Let estimation error $\mathbf{e}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}} - \hat{\mathbf{x}}_{\mathcal{I}}$, its dynamics are obtained from (5.3.7) and (5.3.10) as follows

$$\Sigma^{\tilde{\mathcal{O}}}_{\mathcal{V}^{i''}_\sigma}: \begin{cases} \dot{\mathbf{e}}_{\mathcal{I}} = \bar{\mathbf{A}}^{\mathcal{I}}_\sigma \mathbf{e}_{\mathcal{I}} + \rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}}) + \mathbf{B}_{\mathcal{A}''} \mathbf{u}_{\mathcal{A}''} \\ \mathbf{r}^i_\sigma = \mathbf{C}^{\mathcal{I}}_\sigma \mathbf{e}_{\mathcal{I}} \end{cases}, \tag{5.3.11}$$

in which $\mathbf{e}_{\mathcal{I}}(t_k) = \text{diag}(\mathbf{0}_{|\mathcal{I}_i|+1\times|\mathcal{I}_i|+1}, I_{|\mathcal{I}_i|-1})\mathbf{x}_{\mathcal{I}}(t_k)$ if $\mathcal{V}^{i''}_{\sigma(t_k)} \neq \mathcal{V}^{i''}_{\sigma(t_{k-1})}$ or $k = 0$, and $\mathbf{e}_{\mathcal{I}}(t_k) = \mathbf{x}_{\mathcal{I}}(t_k) - \hat{\mathbf{x}}_{\mathcal{I}}(t_k)$ otherwise, with $\{t_k\}_{k=0}^{\mathbf{m}}$, $\mathbf{m} \in \mathbb{Z}_{\geq 0}$.

The following result characterizes the dynamical response of the attack detectors (local observers) in the presence and absence of adversaries. The results can be used in threshold design for residual signals later in attack detection.

Theorem 5.3.5. *Consider $\Sigma_{\sigma(t)}$ in (5.1.1) with a $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}^{\mu}_{T})$-vertex-connected (resp. $r(\boldsymbol{\mathcal{G}}^{\mu}_{T})$-robust) communication network as defined in (5.2.7) under Assumptions 5.2.1 and 5.2.2. Let (5.1.1) be subject to an F-total, where $F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}^{\mu}_{T}) - 1$, (resp. F-local, where $F \leq r(\boldsymbol{\mathcal{G}}^{\mu}_{T}) - 1$) adversary set with input $\mathbf{u}_{\mathcal{A}} \in \mathcal{L}_{pe}$. Let each mobile agent $i \in \mathcal{V}$ be equipped with a reconfigurable local attack detector $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_{\sigma}}$ given by (5.3.10) and local information (5.1.6). Then, for each $\|\mathbf{e}_{\mathcal{I}}(t_k)\| < w_{\mathcal{I}}$, with $w_{\mathcal{I}} \in \mathbb{R}_{>0}$, (5.3.11) is finite-gain $\mathcal{L}_p$ stable and the residuals $\mathbf{r}^i_{\sigma}(t)$'s hold the bound*

$$\begin{aligned}\left|\mathbf{r}^{i,j}_{\sigma}(t)\right| \leq \kappa^{\mathcal{I}}_{\mathbf{e}} w_{\mathcal{I}} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)} + \left(\tfrac{\kappa^{\mathcal{I}}_{\mathbf{r}}}{\lambda^{\mathcal{I}}_{\mathbf{e}}} \|\mathbf{x}(t_0)\| e^{-\lambda_{\mathbf{x}}(t_k-t_0)}\right)\left(1 - e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}\right) + \\ \left(\tfrac{1+\kappa^{\mathcal{I}}_{\mathbf{r}}}{\lambda^{\mathcal{I}}_{\mathbf{e}}}\right) \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| \left(1 - e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}\right), \quad \forall\, t \in [t_k,\ t_{k+1}), \end{aligned} \tag{5.3.12}$$

where $\mathbf{r}^{i,j}_{\sigma}(t)$ is the j-th component of $\mathbf{r}^i_{\sigma}(t)$ and denotes the position estimation of the two-hop neighbors, corresponding to the j-th row of $\mathbf{C}^i_{\sigma}$, in each mode $\sigma(t_k) \in \mathcal{Q}$, $\forall\, t \in [t_k\ t_{k+1})$, $k \in \mathbb{Z}_{\geq 0}$. Also, $\kappa^{\mathcal{I}}_{\mathbf{r}} = \alpha\kappa_{\mathbf{x}}\kappa^{\mathcal{I}}_{\mathbf{e}}$, with the known constants[5] $\kappa^{\mathcal{I}}_{\mathbf{e}}$, $\lambda^{\mathcal{I}}_{\mathbf{e}} \in \mathbb{R}_{>0}$, such that $\|e^{\bar{\mathbf{A}}^{\mathcal{I}}_{\sigma(t_k)}(t-t_k)}\| \leq \kappa^{\mathcal{I}}_{\mathbf{e}} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}$, and $\kappa_{\mathbf{x}}$ and $\lambda_{\mathbf{x}}$ are given in Proposition 5.2.6. Additionally, if $\mathcal{A} = \emptyset$, each $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_{\sigma}}$ is exponentially stable with $\mathbf{e}_{\mathcal{I}}(t) \to \mathbf{0}$, and the residuals in (5.3.12) hold the following bound

$$\left|\mathbf{r}^{i,j}_{\sigma}(t)\right| \leq \kappa^{\mathcal{I}}_{\mathbf{e}} w_{\mathcal{I}} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)} + \left(\tfrac{\kappa^{\mathcal{I}}_{\mathbf{r}}}{\lambda^{\mathcal{I}}_{\mathbf{e}}} \|\mathbf{x}(t_0)\| e^{-\lambda_{\mathbf{x}}(t_k-t_0)}\right)\left(1 - e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}\right) := \epsilon^{i,j}_{\sigma}, \tag{5.3.13}$$

[5]Recall that $\mathcal{I}$ is a shorthand for the set $\mathcal{I}_i = \mathcal{V}^{i''}_{\sigma}$ and thus the constants are mode-dependent for each cooperative agent $i \in \mathcal{V} \setminus \mathcal{A}$.

Algorithm 4 Resilient & Reconfigurable Cooperation

Input: $\Phi^{i}_{\sigma(t)}$ in (5.1.6), $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i}_{\sigma}{}''}$ in (5.3.10), and $\epsilon^{i,j}_{\sigma}$ in (5.3.13), $\forall\, i \in \mathcal{V} \setminus \mathcal{A}$

// Accept the null $\mathcal{H}^{\mathrm{o}}$ in (5.1.7a) and assume $\mathcal{A}'' = \emptyset$
procedure 1: DISTRIBUTED DETECTION & ISOLATION
// Use the most recent info $\Phi^{i}_{\sigma(t)}$ to (re)initialize $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i}_{\sigma}{}''}$
Compute the residual $\mathbf{r}^{i}_{\sigma}(t)$ and the corresponding thresholds $\epsilon^{i,j}_{\sigma}$
    for $j \in \mathcal{N}^{i(1)}_{\sigma}$ **do**
        if $|\mathbf{r}^{i,j}_{\sigma}(t)| > \epsilon^{i,j}_{\sigma}$ **then**
            Reject the null hypothesis $\mathcal{H}^{\mathrm{o}}$ in (5.1.7a) ▷ Detection: Malicious agent $j \in \mathcal{N}^{i(1)}_{\sigma}$ is detected by the cooperative agent i.
            $\mathcal{A}'' \leftarrow j \in \mathcal{N}^{i(1)}_{\sigma}$
            Set $a^{\sigma}_{ij} = 0,\ j \in \mathcal{A}'' \cap \mathcal{N}^{i(1)}_{\sigma}$ ▷ Stop communication with $\mathcal{A}''$
            Update $\Phi^{i}_{\sigma(t)}$
        end if
    end for
end procedure
procedure 2: RESILIENT COOPERATION DEFINED IN (5.1.8)
    Run $\boldsymbol{u}^{\mathrm{n}}_{i}(t)$ given in (3.1.3) with the information from $\mathcal{N}^{i(1)}_{\sigma} \setminus \mathcal{A}''$
end procedure

where $\epsilon^{i,j}_{\sigma}$ is a threshold that can be used in (5.1.7).

Proof. See Appendix C.10.

Theorem 5.3.5 shows that the local observer (5.3.10) with residual $\mathbf{r}^{i}_{\sigma}$ has bounded-input bounded-output (BIBO) stability for the worst-case number of malicious agents with deception attacks that are defined in Section 5.1.3, and that whose detectability is ensured by a certain degree of network connectivity that is quantified in Lemma 5.3.2.

5.4 Resilient Cooperation

Building upon the results in the previous sections, we present an algorithmic framework, summarized in Algorithm 4, as a solution to the resilient cooperation problem stated in Section 5.1.4. Algorithm 4 comprises two simultaneous procedures addressing the distributed detection and isolation of malicious agents by using (5.1.7) for

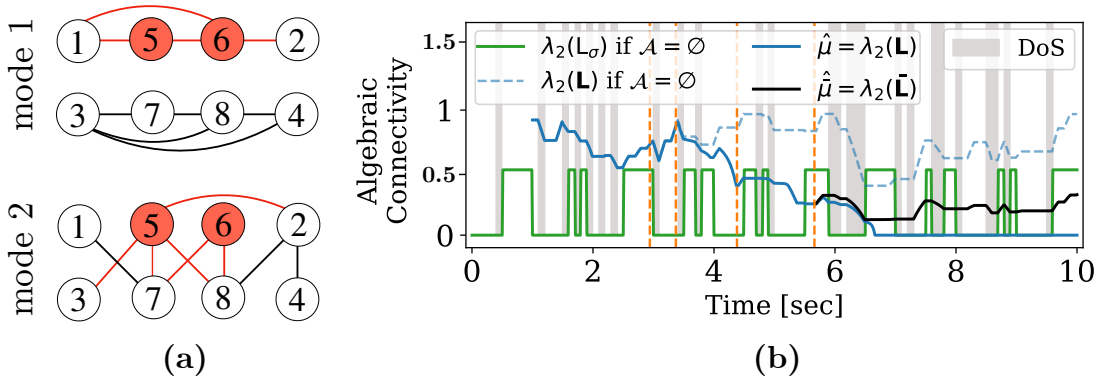


Figure 5.2: Communication network $\mathcal{G}_{\sigma(t)}$ in (a) and its algebraic connectivity in the integral sense of (5.2.1) in (b) for Section 5.5-Example 1. (a) The network switches between two modes every 0.5 sec whose union forms a static overlay network $\boldsymbol{\mathcal{G}}_T^\mu$ with $\lambda_2(\mathbf{L}) = 2.1049$ that is 3-robust [67, Fig. 4], ensuring $(3,1)$-robustness, and $(3,1)$-vertex-connectivity (see Section 5.2 and (5.2.7)). per Section 5.1.3, the network $\mathcal{G}_{\sigma(t)}$ is subject to a 2-total and 2-local set of malicious agents $\mathcal{A} = \{5,6\}$. It is also subject to a distributed DoS whose link dropouts follow a binomial distribution with 100 trials and a success probability of 0.3 during 10 sec. (b) The illustration of positive algebraic connectivity $\lambda_2(\cdot)$ in the integral sense (5.2.1) for for the network $\mathcal{G}_{\sigma(t)}$ and its induced network $\bar{\mathcal{G}}_{\sigma(t)}$ in (5.1.9) despite their intermittent connections (See also remark 5.2.2). The results in (b) are from resilient consensus in Fig. 5.3-(a) through Algorithm 4. The decrements in $\lambda_2(\cdot)$ during $t \in [0,\ 5.66]$ are due to the permanent link disconnections that occurred in the attack detection and isolation procedure, see Fig. 5.3-(a).

decision-making, and resilient cooperation. In what follows, we present the technical discussions of Algorithm 4.

Isolation of the set of the malicious agents $\mathcal{A} \subset \mathcal{V}$. Upon detection of neighboring malicious agents by each cooperative agent $i \in \mathcal{V} \setminus \mathcal{A}$, there follows the isolation (removal) of the detected malicious agents from the network (Lines 7-10 in Algorithm 4). Note that the results in Proposition 5.3.4, Lemma 5.3.2, and Theorem 5.3.5 allow each cooperative agent $i \in \mathcal{V} \setminus \mathcal{A}$ in a $(F+1, T)$-robust (resp. $(F+1, T)$-vertex-connected) network to perform the distributed hypothesis testing in (5.1.7) and detect a *candidate* set of malicious agents within its 1-hop neighbors, provided the *actual* set of malicious agents, $\mathcal{A}$, is at most F-local (resp. F-total).

Here, the distinction between a *candidate* set and the *actual* set of malicious agents is due to the possibility of *false* alarms in (5.1.7). (i.e., a *candidate* set is almost always a superset of the *actual* set for a sufficiently small threshold in (5.1.7)). The foregoing sets coincide if $\bar{\mathbf{A}}_\sigma^\mathcal{I} = (\mathbf{A}_\sigma^\mathcal{I} - \mathbf{H}_\sigma^\mathcal{I}\mathbf{C}_\sigma^\mathcal{I})$ in (5.3.10) features distinct eigenvalues, guaranteeing that each residual's component $\mathbf{r}_\sigma^{i,j}$ in (5.3.12) is most sensitive to only one of the input directions associated with the malicious agents within the 1-hop neighbors.

State-dependent switching We note the isolation process based on (5.1.7) (lines 7-10 in Algorithm 4) imposes a finite number of state-dependent switches that are not explicitly incorporated in the condition (5.2.1) for $\mathsf{L}_{\sigma(t)}$ with time-dependent switches $\sigma(t) : \mathbb{R}_{\geq 0} \to \mathcal{Q}$. On the other hand, the results of Theorem 5.2.5 for the bound on the network connectivity in the integral sense of (5.2.1) after node and edge removal holds independent of the type of switches. Therefore, upon a link removal between a cooperative and malicious agent(s), there exists a new Laplacian matrix $\mathsf{L}_{\sigma(t)}$ that holds the connectivity condition of the from (5.2.1) for the system in (5.1.1) starting from the new initial condition $\mathbf{x}(t_k) \in \mathbb{R}^{2|\mathcal{V}|}$ with t_k, $k \in \mathbb{Z}_{\geq 0}$, being the time instant of the newly active mode $\sigma(t_k) \in \mathcal{Q}$. Having the integral connectivity as in (5.2.1) independent of the states' initial condition, Proposition 5.2.6 can be applied. It is worth mentioning that the independence from the states' initial conditions for the (μ, T)-PE connectivity in (5.2.1) is a special case of having (5.2.1) parameterized of the form $\frac{1}{T}\int_t^{t+T} Q\mathsf{L}_{\sigma(\tau,\boldsymbol{\lambda})}Q^\top \mathrm{d}\tau \geq \mu I_{N-1}$, $\forall\, t \in \mathbb{R}_{\geq 0}$, that holds for each $\boldsymbol{\lambda} := (t_\mathrm{o}, \mathbf{x}_\mathrm{o}) \neq (t_0, \mathbf{x}(t_0))$ with the switching signal $\sigma(t, \mathbf{x}(t_k)) : \mathbb{R}_{\geq 0} \times \mathcal{X} \to \mathcal{Q}$, $\mathcal{X} \subset \mathbb{R}^{2|\mathcal{V}|}$ (see [73]).

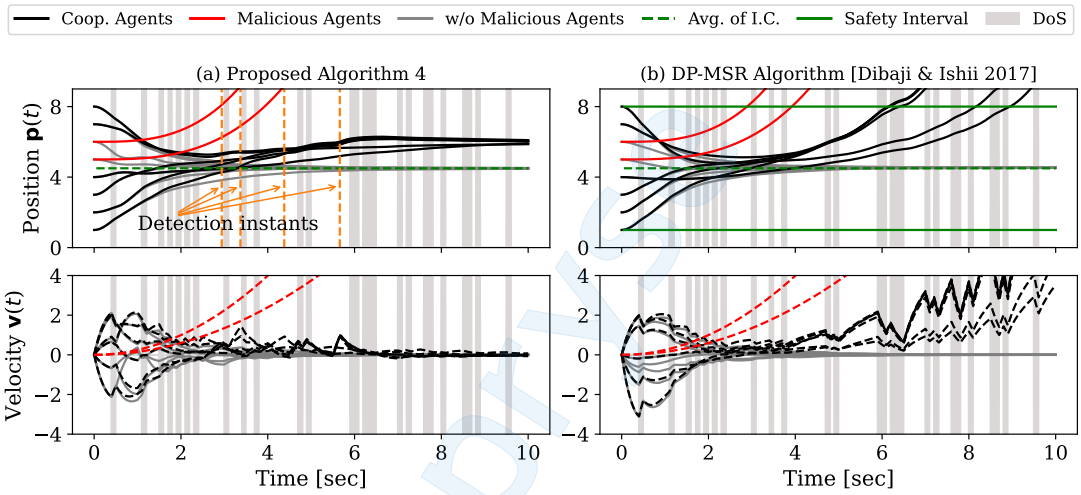
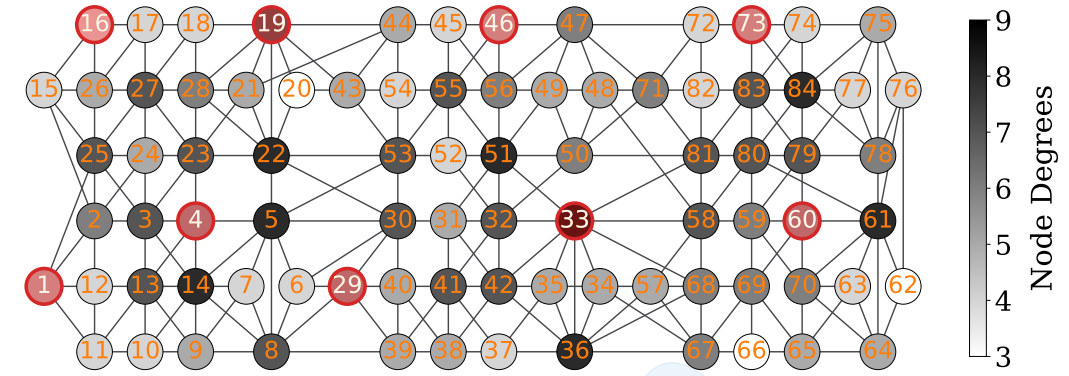
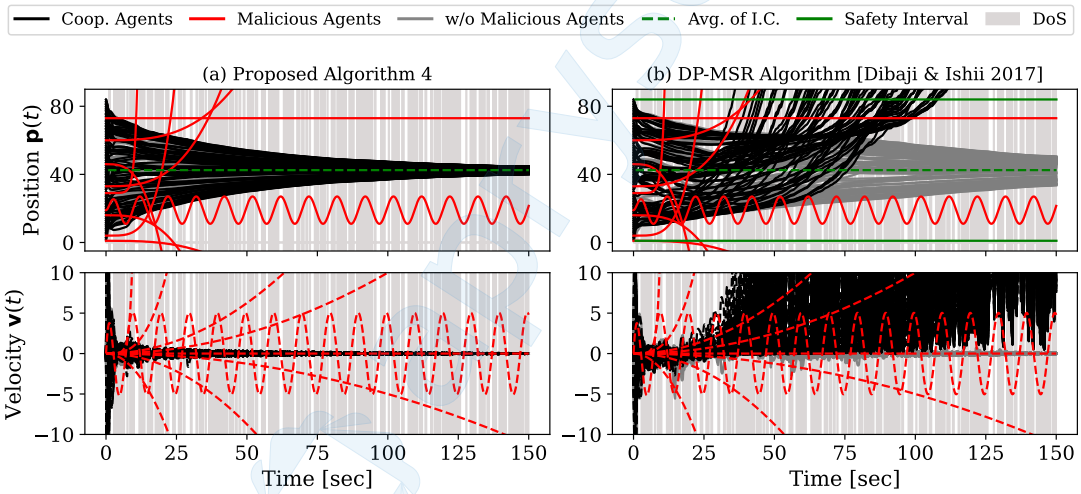


Figure 5.3: Example 1: Comparison of resilient consensus in an 8-agent network $\mathcal{G}_{\sigma(t)}$ that is, as shown in Fig. 5.2, $(3,1)$-robust and subject to DoS attacks and a 2-total and 2-local set of malicious agents $\mathcal{A} = \{5, 6\}$ with $\boldsymbol{u}_5(t) = 0.3t$ and $\boldsymbol{u}_6(t) = 0.5t$ in (3.1.3). (a) Resilient consensus using Algorithm 4 whose resilient to the 2-total/2-local set $\mathcal{A}$ in the $(3,1)$-robust network is guaranteed by Lemma 5.3.2 and Theorem 5.3.5. Also, the vertical orange dashed lines specify the time instants where cooperative agents detected and disconnected from their respective neighboring malicious agents (lines 7-10 of Algorithm 4 with $\epsilon_\sigma^{i,j} = 0.95$) using its local attack detector in (5.3.10). (b) Resilient consensus using the DP-MSR algorithm that for a 3-robust network has provable resilient consensus only in the presence of up to 1-local or 1-total malicious agents [37, 36], accounting for the failure of the approach in this case where $\mathcal{A}$ is 2-local and 2-total.



(a) Underlying Communication Network $\boldsymbol{\mathcal{G}}_T^{\mu}$



(b) Agents' State Trajectories During Consensus

Figure 5.4: Example 2: Resilient consensus in an 84-agent network $\mathcal{G}_{\sigma(t)}$ subject to deception and DoS attacks defined in Section 5.1.3. The deception attacks are introduced by a 1-local set of 9 malicious agents, $\mathcal{A} = \{1, 4, 16, 19, 29, 33, 46, 60, 73\}$, which are shown in red color. The distributed DoS attack (5.1.5) imposes link dropouts following a binomial distribution, with 600 trials and a success probability of 0.4 during 150 sec. (a) The static overlay network $\boldsymbol{\mathcal{G}}_T^{\mu}$ is 2-robust, constructed using the preferential-attachment model in [67, Thm. 5] based on the topology in [67, Fig. 6]. Despite intermittent connections, the network $\mathcal{G}_{\sigma(t)}$ is $(2,1)$-robust and $(3,1)$-vertex-connected (see Definitions 5.2.2 and 5.2.3, and Lemma 5.2.3). $(2,1)$-robustness, then, ensures resilience to any 1-local set $\mathcal{A}$ as it follows from Lemma 5.3.2 and Theorem 5.3.5. (b) Resilient consensus using Algorithm 4 over the intermittent network $\mathcal{G}_{\sigma(t)}$ in (a) and in the presence of the 1-local malicious set $\mathcal{A}$.

5.5 Simulation Results

We conduct two simulation studies to illustrate the theoretical results and compare them with the state-of-the-art [78, 37]. We also provide the code at `https://github.com/SASLabStevens/rescue`.

Example 1. We compare our proposed Algorithm 4 with the DP-MSR algorithm [36, 37]. Employing Algorithm 4, we achieve resilient consensus in an 8-agent network that is subject to switching topology and both deception and DoS attacks. See Figs. 5.2 and 5.3(a). In contrast, the DP-MSR algorithm fails to achieve resilient consensus for the same network (Fig. 5.3(b)), despite the advantage of operating over a network that is the union of the two modes shown in Fig. 5.2(a). The outperformance of Algorithm 4 is because of its observer-based nature that leverages local information Φ_σ^i in (5.1.6) to detect a larger set of malicious agents in a network with a specific degree of connectivity and r-robustness (see Lemma 5.3.2), a capability not shared by the DP-MSR algorithm. We note that the analysis of resilient consensus via the DP-MSR algorithm was originally developed for a discretized version of (5.1.1) in [37, 36] while our results are in the continuous-time domain. To have the results in a comparable time scale, we used the DP-MSR procedure with the small sample time $T_\mathrm{s} = 0.001$ and the gains $\gamma = 3$ and $\alpha = 1$ in the zero-order-hold discretization of (3.1.3). This set of parameters does not completely satisfy the sufficient condition in [37, eq. (9)], but does satisfy a relaxation thereof, similar to the discussion in a footnote in [37]. This enables an asymptotic resilience consensus in the case $\mathcal{A} = \emptyset$ (shown with the gray-colored state trajectories) and also in the cases of $(F=1)$-local and $(F=1)$-total adversary sets (not shown herein) over any 3-robust network.

Example 2. We evaluate the scalability of our framework on an 84-agent network subject to a DoS attack and 9 malicious agents that form a 1-local set, see

Fig. 5.4. Notably, each agent has at most 9 neighbors (less than 11% of total agents), resulting in a sparse graph. Despite sparsity, the graph has the required robustness properties. This observation underscores the significance of sparse graphs with strong robustness/connectivity properties (e.g., expander graphs), offering resilience without excessive communication overhead.

Scalability and computational complexity. We remark that the proposed Algorithm 4 improves the scalability of the system-theoretic frameworks relying on observers for attack detection [97, 98, 78]. Note that, for each agent, attack detection in Algorithm 4 (lines 4-10), requires only one observer with a 2-hop dynamics $\mathbf{A}_\sigma^{\mathcal{I}}$ in (5.3.10) with worst-case complexity $\mathcal{O}(|\mathcal{V}_\sigma^{i''}|^2)$ rather than the complete model $\mathbf{A}_\sigma$ in (5.1.1) with $\mathcal{O}(|\mathcal{V}|^2)$, ($|\mathcal{V}_\sigma^{i''}| \leq |\mathcal{V}|$, see (5.1.3)), which is the case in [78]. This local topological information, $\mathbf{A}_\sigma^{\mathcal{I}}$ can be pre-programmed [144] or transmitted as formalized in (5.1.6), in which case it may incur only a minimal communication overhead, given the often sparse communication topology of mobile robots due to their mobility (see Fig. 5.4). Moreover, the local information (5.1.6) allows for detecting a greater number of malicious agents in a given network, compared to the prior work [97] including the graph-theoretic MSR-like algorithms [36, 37, 113] (see Fig. 5.3), whose worst-case complexity is quadratic in time $\mathcal{O}(|n|^2)$ and linear in space $\mathcal{O}(n)$, w.r.t. the size of inclusive 1-hop neighbors [66], i.e. $n = |\mathcal{V}_\sigma^{i'}|$, see (5.1.3). Finally, given the switching nature of the local observer (5.3.10) with resetting initial conditions, an increased frequency of topology switching, potentially violating Assumption 5.2.1, would lead to significant performance degradation in attack detection as observer's residuals would persist in a transient convergence phase.

Chapter 6

Multi-Robot Coordination with Adversarial Perception

Learning-enabled visual perception is of significant importance to many robotic tasks, particularly in the forms of visual servoing [68], visuomotor policies [70], learned visual odometry (VO) [80, 22], foundation models for visual navigation [119] and object tracking [75], drone flocking [117, 148] and collaborative perception [154]. However, learned perception models are vulnerable to adversarial instances [49, 63] where a human-imperceptible level of noise on the input data (e.g., camera images) can significantly mislead the model's output (e.g., object misclassification and mislocalization that may be dynamically infeasible or unsafe [54, 143, 21, 58]).

This chapter extends the prior work on perception-based multi-robot coordination to the case of adversarial image attacks [49, 19, 143, 54] that incur misclassification and mislocalization in the learned perception module[1] of the robots. More specifically, we consider a network of robots that rely on an onboard sensor suite of IMU and RGB camera images for relative localization in a map and coordination with one another over a wireless communication network. Similar to [117, 44, 100], a custom-trained object detection model processes each camera frame to output 2D bounding boxes around objects (e.g., robots) within the field of view (FoV). Adversarial image attacks targeting this perception model can cause misclassification and mislocalization of the objects in the FoV. We formulate these adversarial misclassifications and mislocalizations as *spurious* measurements (false-positive detections) and

[1] In the context of statistical inference, this adversarial effect can be formulated, subject to certain conditions, as a covariate shift [10]. That is for a given learned model $y = \hat{f}_\theta(x)$ trained over a dataset $D = \{(x_i, y_i)\}_{i=1}^n$ with probability distribution $p_{\text{train}}(x_i)$, a covariate shift is induced by the perturbations δx on the input data at test-time such that $p_{\text{test}}(x_i + \delta x) \neq p_{\text{train}}(x_i)$. This causes type I (false positive) and type II errors (false negative) at inference.

sporadic measurements (intermittent measurements incurred by misclassifications), and propose a system-theoretic approach based on a variant Kalman filter to evaluate their effects on relative localization and multi-robot coordination.

We evaluate our proposed framework through experiments. Additionally, we present two multi-robot platforms equipped with open-source software used in our experiments. Our framework is lightweight and well-suited for real-time applications.

6.1 Related Work

Adversarial Perturbations on Vision Tasks. Adversarial instances in a single-task and static settings such as image classification are an active field of research [49, 19, 65]. The adversarial samples are designed by adding carefully designed noise or patches to the original image to mislead the model. The noise-based adversarial samples are designed by two metrics: Euclidean L_p-norms to bound the noise level [19] or human-level perceptual similarity measured by Learned Perceptual Image Patch Similarity (LPIPS) distance [65]. As most of these methods entail an iterative optimization process, some studies proposed the design of a universal (single) small perturbation, for all images in image classification, semantic segmentation, and object detection tasks [85, 137, 26]. The transferability of a designed adversary across different architectures was studied in [72, 145]. Alternatively, some adversarial attacks target the availability of object detection models by overloading the module, which causes a significant increase in the inference time [121, 23]. A few studies extended the previous results to the dynamical settings (e.g., object detection and following, [54, 143], pose estimation [21], and perception-based control [58]) where the system dynamics are of consideration in designing successful adversarial perturbations in real-time.

We further extend the prior results to the case of multi-robot coordination where adversarial perturbations on the visual measurements may lead to the instability of the entire system.

Object Tracking, Localization, and Data Association. Tracking objects (robots) using perceptual observations is a well-studied task [6, 115, 123, 44, 100, 150]. State-of-the-art approaches employ a *tracking-by-detection* paradigm, a sequential process of (object) detection to obtain measurements, and then data association to determine which measurement is associated with which *track* (e.g., relative position to the object of interest [117]), enabling object tracking via a tracking filter. Some of the widely used algorithms for solving the data association problem are GNN, JPDA, PHD, MHT [100, 44, 123, 60, 105], and more recently learning-enabled solutions such as SORT [13] and BYTE [150]. Alternative approaches follow a simultaneous detection and tracking paradigm, wherein the association and detection are *learned* jointly as one module [153].

However, the robustness of these methods for multi-robot coordination under adversarial perception conditions is not well understood.

Adversarial Robustness and Defences for Learned Perception Models. We only review the most relevant work here and refer to [95] for a comprehensive review of adversarial threat models and defense mechanisms for learning-enabled frameworks. A very common approach to adversarial robustness in test-time is adversarial training. Either L_p-norm bounded perturbations [19, 145, 26] or human-level perceptual similarity metrics that approximate the set of all imperceptible adversarial perturbations [65] are used to generate the adversarial samples. The former is faster while the latter results in a higher level of robustness. Moreover, adversarial training has inherently a larger sample complexity and can cause standard vs. adversarially robust generalization trade-offs in both static [118, 104] and dynamic settings [149].

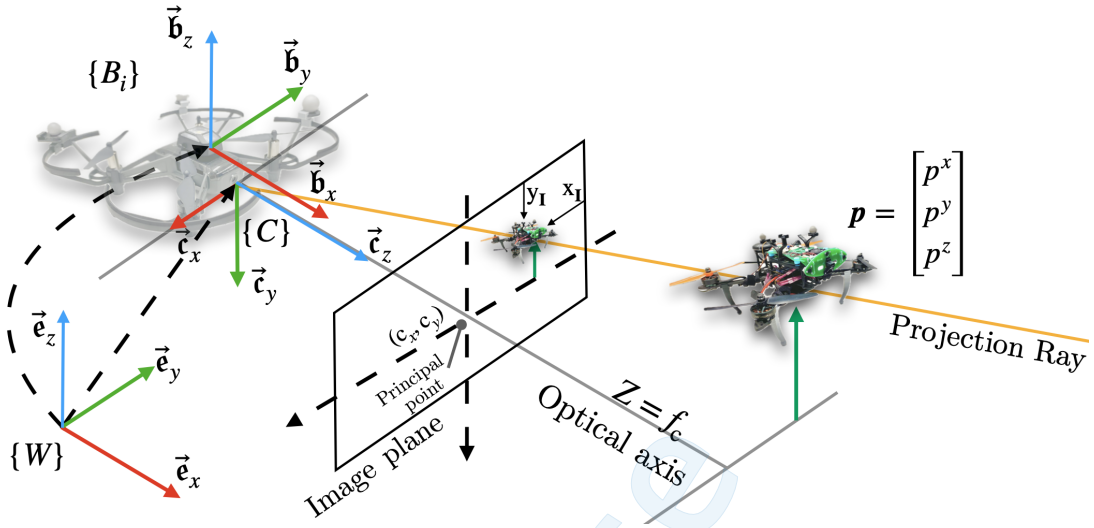


Figure 6.1: Illustration of reference frames and the *perspective* camera projection model. $\{\mathcal{W}\}$ is the common inertial (world) frame, and $\{\mathcal{B}_i\}$ is the body-fixed frame of the i-th agent (robot) on which a forward-pointing centered camera is attached with the coordinate frame $\{\mathcal{C}\}$. We let $\mathrm{R}_{\mathcal{WB}} =: \mathrm{R}$ and $\mathrm{R}_{\mathcal{BC}} =: \bar{\mathrm{R}}$ which yields $\mathrm{R}_{\mathcal{CW}} = \mathrm{R}_{\mathcal{CB}}\mathrm{R}_{\mathcal{BW}} = \bar{\mathrm{R}}^\top \mathrm{R}^\top$. Finally, without loss of generality, we assume that the body frame $\{\mathcal{B}_i\}$ and the camera frame $\{\mathcal{C}\}$ have no offset and differ only in orientation.

Evaluating the consistency between the outputs of two perception modules can be used to detect adversarial cases [61]. Alternatively, adversarial purification is used to purify the adversarial perturbation before running the task [89]. Finally, conformal prediction can be adopted to obtain a set of valid answers for any given adversarial sample in classification tasks [45, 133, 10].

6.2 Methodology

Notations. We refer to Fig. 6.1 for the notations of robots' poses, and the coordinate frames. In particular, $\boldsymbol{p}_{ij} = \boldsymbol{p}_i - \boldsymbol{p}_j$ denotes the relative position expressed in the global frame $\{\mathcal{W}\}$, while $\boldsymbol{p}^{\mathcal{C}}_{ij} = \mathrm{R}_{\mathcal{CW}}\boldsymbol{p}_{ij}$ denotes the relative position expressed in the camera frame $\{\mathcal{C}\}_i$ of the i-th agent (robot).

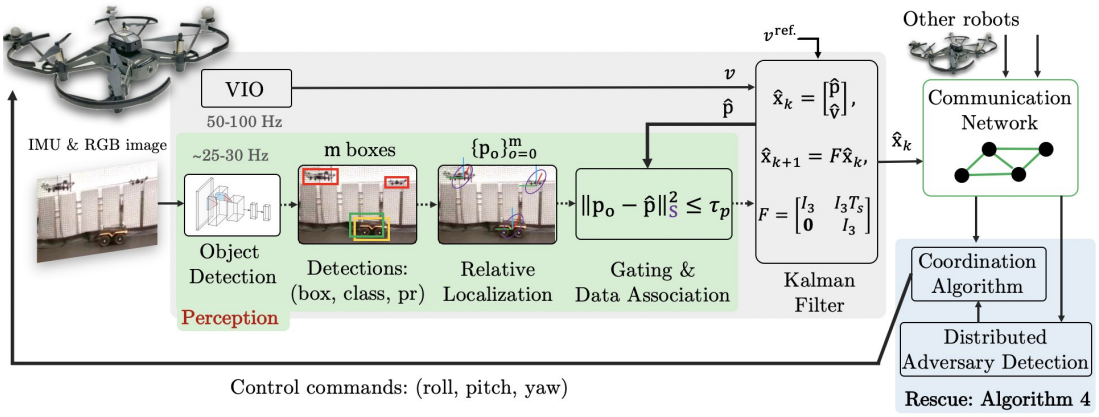


Figure 6.2: Overview of the perception-based multi-robot coordination. The contribution of this chapter is highlighted in the gray box, which encompasses the perception module shown in the green box. This module integrates (Visual-Inertial Odometry) VIO data and detected objects from the object detection module to provide state estimation for the ego-robot, along with capabilities for relative localization and object tracking. The blue box shows the consensus-based coordination algorithm and the adversary detection algorithm developed in Chapter 5. These two modules allow for resilient coordination in the presence of adversarial attacks on images or transmitted information over the communication network.

6.2.1 Objectives

We propose a framework to evaluate the resilience of multi-robot coordination with learned perception modalities against adversarial image attacks. We model the effects of a class of adversarial image attacks as producing *sporadic* (intermittent) and *spurious* (false) measurements in perception-based relative localization. Our proposed framework is shown in Fig. 6.2. It integrates the following modules: *Detection* in Section 6.2.2: This module uses a learned perception model to process onboard RGB camera images in real-time, detecting objects of interest (e.g., landmarks or neighboring robots) within the Field of View (FoV). The robot's localization depends on this module, which is vulnerable to adversarial attacks [49, 143]. *Vision-based Relative Localization* in Section 6.2.4: This module converts the 2D bounding-box detections from the perception module into relative positions of the robot to the objects of in-

terest, and assigns a level of uncertainty to these 3D relative positions. Here, we represent the effects of adversarial mislocalization as spurious localization measurements. *State Estimation* in Section 6.2.5: This module employs a variant of the Kalman filter to integrate the recovered 3D relative positions and VIO data, providing an accurate estimate of the robot's position relative to the object of interest. It addresses spurious and intermittent measurement data caused by adversarial image attacks on the perception module through gating and data association techniques. *Coordination and Control* in Section 6.2.6: This module provides consensus-based coordination using relative positions provided by the state estimation (Kalman filter) module.

In Section 6.3, we present two multi-robot platforms with open-source software projects developed for this study. Finally, experimental results in Section 6.4 evaluate the resilience of the proposed framework for the perception-based multi-robot coordination subject to adversarial image attacks.

6.2.2 Perception Model: Object Detection

We consider a multi-task learned perception model $\widehat{Y} = P(\mathbf{I})$ for object detection (e.g., YOLOv7 [136] or RT-DETR [151]). $P(\cdot)$ takes an RGB images $\mathbf{I}$ as input and outputs $m \geq 0$ detections of the form $\{Y\}_{i=0}^{m} = \{\text{box, class, pr}\}_{i=0}^{m}$, where the 4D vector box $= (\mathrm{x}_{\mathbf{I}}, \mathrm{y}_{\mathbf{I}}, \mathrm{w}_{\mathbf{I}}, \mathrm{h}_{\mathbf{I}})$ is a bounding box at image space, centered at $(\mathrm{x}_{\mathbf{I}}, \mathrm{y}_{\mathbf{I}})$ with the width $\mathrm{w}_{\mathbf{I}}$ and height $\mathrm{h}_{\mathbf{I}}$, around each detected object belonging to a class with a confidence probability pr. Here, we custom-trained the original YOLOv7 model with 80 classes to detect 82 classes that include drones and jackal-UGV in our experiments. See Section 6.4.1 for details. We also note we use YOLOv7 since it is fast (30 FPS), and it also has a better detection performance for small objects (e.g., small quadrotors) compared to its Transformer-based counterpart, RT-DETR [151].

6.2.3 Adversarial Image Attacks as Adversarial Measurements

In adversarial settings [19, 65, 21, 143], a human-imperceptible adversarial perturbation (noise) $\delta\mathbf{I}$ is designed and added to the original image frame $\mathbf{I}$ such that the error of perception model $P(\cdot)$ is maximized by some metrics. There are various methods to design the adversarial perturbation either $\delta\mathbf{I}$ online [143] or offline through a universal image attack [85, 137]. Despite the variety of methods for designing adversarial image perturbations, their effects on the perception model's output are categorically similar. Specifically, for object detection models, such adversarial attacks can cause *misclassification* [65, 10, 49], *mislocalization* [143, 54, 58, 21], and increased *latency* [121, 23]. Formally, for a perception (object detection) model $P(\cdot)$ and any two samples $S_1 = (\mathbf{I}_1, \{\widehat{Y}_1\}_{i=0}^{m})$ and $S_2 = (\mathbf{I}_2, \{\widehat{Y}_2\}_{i=0}^{m'})$, where $\mathbf{I}_2 = \mathbf{I}_1 + \delta\mathbf{I}_1$, we define

$$d(S_1, S_2) = \begin{cases} d_{\mathcal{I}}(\mathbf{I}_2, \mathbf{I}_1), & \text{if} \quad \text{class} = \text{class}', \\ \infty, & \text{otherwise}, \end{cases} \tag{6.2.1a}$$

$$P(\mathbf{I}) = \{\widehat{Y}_1\}_{i=0}^{m} = \{\text{box, class, pr}\}_{i=0}^{m}, \tag{6.2.1b}$$

$$P(\mathbf{I} + \delta\mathbf{I}) = \{\widehat{Y}_2\}_{i=0}^{m'} = \{\text{box}', \text{class}', \text{pr}'\}_{i=0}^{m'}. \tag{6.2.1c}$$

in which $d_{\mathcal{I}}(\cdot, \cdot)$ can be either an L_p distance with $p \in \{0, 1, 2, \infty\}$, as defined in [19], or a Learned Perceptual Image Patch Similarity (LPIPS) distance [65]. Additionally, overload (latency) attacks [23] cause $m' \gg m$ in (6.2.1).

In static settings, the fast-gradient sign method (FGSM) [63] to design adver-

sarial image attacks as follows[2]:

$$\begin{gathered}\mathbf{I}_2 = \mathbf{I}_1 + \epsilon \text{sign}(\nabla_{\mathbf{I}_1} \mathrm{J}(Y, \hat{Y})), \\ \text{s.t.} \\ d_{\mathcal{I}}(\mathbf{I}_2, \mathbf{I}_1) = \|\mathbf{I}_2 - \mathbf{I}_1\|_\infty = \|\delta \mathbf{I}\|_\infty \leq \eta, \\ \mathbf{I}_2 \in [0, 255],\end{gathered} \tag{6.2.2}$$

where ϵ is a hyper-parameter chosen to ensure that η remains sufficiently small, the sign operator $\text{sign}(\cdot)$ is applied element-wise to the gradient of the loss function $\nabla_{\mathbf{I}_1} \mathrm{J}(Y, \hat{Y})$. In the Fast Gradient Method (FGM), the gradient of the loss function is used directly without applying $\text{sign}(\cdot)$. In our case, the loss function of YOLOv7 [136] is defined by three terms as follows:

$$\mathrm{J}(Y, \widehat{Y}) = \overbrace{L_{\text{obj}} + L_{\text{class}}}^{\text{classification}} + \overbrace{L_{\text{box}}}^{\text{regression}}, \tag{6.2.3}$$

where $Y = \{\text{box, class, pr}\}$ denotes the target (i.e. true class labels with confidence probability $\text{pr} = 1$ and their respective box coordinates) and $\widehat{Y} = P(\mathbf{I})$ is the model's inference output. We refer to [136] for details on the terms of the loss function.

Example 6.2.1. (FGSM adversarial image attack on YOLO Object Detection). Fig. 6.3 demonstrates the effect of FGSM adversarial image attack, as defined in (6.2.2) with $\eta = 10/255$ on our custom-trained YOLOv7 object detection model. To calculate the adversarial noise, we used minimally perturbed ground-truth boxes and kept the class IDs unchanged to focus the adversarial attack's impact on the localization and objectness terms of the cost function (6.2.3). As shown, the adversarial noise resulted in a false positive by detecting a giraffe, a false negative by failing to

[2]We note that similar to [19], we normalize the 8-bit RGB values to the range [0, 1] when calculating the adversarial perturbation and then remap them back to the range [0, 255].

detect a drone, and a reduction in classification confidence for other detected objects.

In real-time dynamic settings, designing adversarial attacks and assessing their effects pose additional challenges. Adversarial attacks on perception data can have a longitudinal impact on the system's stability and dynamics [143, 54, 57]. Given that object detection outputs are used as measurements in a closed-loop control system (see Fig. 6.4), we propose that adversarial image attacks targeting classification integrity/ accuracy (i.e. $d(S_1, S_2) = \infty$ in (6.2.1)) cause the unavailability of measurements. In contrast, adversarial image attacks targeting localization integrity/accuracy (i.e. $d(S_1, S_2) \neq \infty$ in (6.2.1)) induce (bounded) perturbations in measurements, specifically affecting the localization of 2D bounding boxes in the image space. These perturbations translate into 3D localization errors in Euclidean space and affect state estimation, which will be modeled in Sections 6.2.4 and 6.2.5. Therefore, adversarial *misclassification* and *mislocalization* are modeled as *sporadic* (intermittent) and *spurious* measurements. This formulation facilitates resilience analysis that is agnostic to both the specific adversarial image attack model and the targeted learned perception (object detection) model.

Remark 6.2.1. ***(The Scope of Adversarial Image Attacks).*** *It is important to note that adversarial attacks causing norm-bounded disturbances on measurements have been explored previously for perception-based control [1, 35] and state estimation [149] in single-robot scenarios. In this dissertation, we extend this consideration to both spurious and sporadic measurements induced by adversarial image attacks in multi-robot coordination settings. Additionally, we note that we do not address the class of generative adversarial image attacks, where inauthentic (fake) images are generated to replace the original robot's camera image frames, resulting in perceptual data injection (alteration) attacks with maximum disruption capability. For fundamental*

limitations on the detectability of such attacks, we refer to [57, 58].

6.2.4 Relative Localization with Adversarial Perception Data (Mislocalization Effect)

Recall that the perception model $P(\mathbf{I}_k)$ provides detections as bounding boxes, $\text{box} = (\mathrm{x}_\mathbf{I}, \mathrm{y}_\mathbf{I}, \mathrm{w}_\mathbf{I}, \mathrm{h}_\mathbf{I})$, for the objects with position $\boldsymbol{p}_r \in \{\mathcal{W}\}$ visible at the RGB image $\mathbf{I}_k$ observed at a time instant $t_k \in \mathbb{R}_{\geq 0}$ by the i-th robot in $\boldsymbol{p}_i \in \{\mathcal{W}\}$. Following the pinhole camera model [50, 20], the mapping from the observed 3D point $\boldsymbol{p}_r \in \{\mathcal{W}\}$ onto the 2D image space is given by

$$\bar{x} := \frac{\mathrm{x}_\mathbf{I} - \mathrm{c}_x}{f_\mathrm{c}} = \frac{\mathrm{x}_\mathcal{C}}{\mathrm{z}_\mathcal{C}}, \qquad \bar{y} := \frac{\mathrm{y}_\mathbf{I} - \mathrm{c}_y}{f_\mathrm{c}} = \frac{\mathrm{y}_\mathcal{C}}{\mathrm{z}_\mathcal{C}}, \tag{6.2.4a}$$

$$\begin{bmatrix} \mathrm{z}_\mathcal{C} & \mathrm{y}_\mathcal{C} & \mathrm{z}_\mathcal{C} \end{bmatrix}^\top = -\mathbf{p}_i^\mathcal{C} = -\mathrm{R}_{\mathcal{C}\mathcal{W}}\mathbf{p}_i = -\mathrm{R}_{\mathcal{C}\mathcal{W}}(\boldsymbol{p}_i - \boldsymbol{p}_r), \tag{6.2.4b}$$

in which the camera intrinsics (i.e. the focal length f_c and the principal point $(\mathrm{c}_x, \mathrm{c}_y) = (W/2, H/2)$ in pixels) are known in a calibrated camera (see Fig. 6.1).

Next, we describe the robot's relative localization with respect to a known object of a known size (e.g., a landmark or another robot) detected by the object detection module.

Assumption 6.2.1. *We assume that the object of interest is in the field of view of all robots coordinating in a common inertial frame (the world frame). Additionally, the object is either sufficiently distant from the robots or small with uniform dimensions, ensuring that the orthographic projection assumption holds.*

Under Assumption 6.2.1, and for a planer object of known size (i.e. width W_Obj and height H_Obj) and given the detected bounding box $\text{box} = (\mathrm{x}_\mathbf{I}, \mathrm{y}_\mathbf{I}, \mathrm{w}_\mathbf{I}, \mathrm{h}_\mathbf{I})$ in

the image plane, one can readily estimate of the object's depth[3] as follows:

$$f_{\mathrm{c}} \frac{\mathrm{W}_{\mathrm{Obj}}}{\mathrm{w}_{\mathbf{I}}} \approx \mathrm{z}_{\mathcal{C}}, \tag{6.2.5}$$

where $\mathrm{z}_{\mathcal{C}}, \mathrm{w}_{\mathbf{I}}$, and f_{c} are given in (6.2.4).

Using the camera orientation $\mathrm{R}_{\mathcal{C}\mathcal{W}}$, which is available from the VIO pipeline, and (6.2.4)-(6.2.5), one can *approximately* recover the *nominal* relative position of the robot with respect to the center of the object of interest, denoted by $\mathbf{p}_i^{\mathrm{n}} \approx (\boldsymbol{p}_i - \boldsymbol{p}_o) \in \mathbb{R}^3$ and expressed in the common reference frame $\{\mathcal{W}\}$, as follows:

$$\mathbf{p}_i^{\mathrm{n}} \approx \mathrm{R}_{\mathcal{C}\mathcal{W}}^{\top} \left(f_{\mathrm{c}} \frac{\mathrm{W}_{\mathrm{Obj}}}{\mathrm{w}_{\mathbf{I}}} \right) \begin{bmatrix} \bar{x} \\ \bar{y} \\ 1 \end{bmatrix}, \tag{6.2.6}$$

Additionally, note that an adversarial image attack $\delta\mathbf{I}$ as in (6.2.1) that induces localization error can be modeled as an offset $\delta\mathrm{box} = (\delta\mathrm{x}_{\mathbf{I}}, \delta\mathrm{y}_{\mathbf{I}}, \delta\mathrm{w}_{\mathbf{I}}, \delta\mathrm{h}_{\mathbf{I}})$ in the detected box. As such, this offset affects the 3D localization in (6.2.6). Therefore, we modify (6.2.6) to incorporate the effect of localization error and define a relative localization uncertainty term for the recovered relative position as follows:

$$\mathbf{p}_i := \mathbf{p}_i^{\mathrm{n}} + \delta\mathbf{p}_i \approx \mathrm{R}_{\mathcal{C}\mathcal{W}}^{\top} \left(f_{\mathrm{c}} \frac{\mathrm{W}_{\mathrm{Obj}}}{\mathrm{w}_{\mathbf{I}} + \delta\mathrm{w}_{\mathbf{I}}} \right) \begin{bmatrix} \bar{x} + \delta\bar{x} \\ \bar{y} + \delta\bar{y} \\ 1 \end{bmatrix}, \tag{6.2.7a}$$

$$\mathbf{R}_i^{\mathrm{pos}} = ((1 - \mathrm{pr})\bar{\epsilon} + \underline{\epsilon})\, I_3, \tag{6.2.7b}$$

[3]For planar objects, under the orthographic projection assumption, the depth is approximately equal to the distance from the camera to the object along the z-direction of the camera frame. (see Fig. 6.1).

where the additive term $\delta\mathbf{p}_i$ represents the 3D localization error caused by the adversarial image attack, pr is the confidence probability of object detection module, and $\bar{\epsilon}, \underline{\epsilon}$ are small positive constants in the measurement covariance matrix $\mathbf{R}_i^{\text{pos}}$, modeling the relative localization uncertainty. We will later use the covariance term (6.2.7b) in a gating and data association problem in Section 6.2.5.

Finally, we remark that the assumption of known object size is common in prior work on relative localization using single-view monocular cameras [117]. Alternative approaches can be employed for depth estimation and relative localization when detection is available from multiple views [112, 88].

6.2.5 State Estimation with Intermittent Adversarial Perception Data (Misclassification Effect)

We use a variant of the Kalman filter with intermittent measurements [122, 140] to integrate Visual-Inertial Odometry (VIO) data with perception data from the object detection module. This integration compensates for the four-dimensional unobservable subspace[4] in the VIO pipeline [127], allowing us to estimate the positions of robots with respect to an object of interest within a map (e.g., a landmark in the map). Additionally, it is important to note that the adversarial image attacks (perturbations) on the perception module can cause spurious and sporadic (intermittent) measurement data (see Section 6.2.3), which do not follow the Gaussian noise distribution assumed in the standard (optimal) Kalman filter derivation. It is known that such measurement degeneracy can lead to instability in the optimal Kalman filter [11, 122, 84, 142]. We empirically evaluate such degeneracy induced by adversarial image attacks on the Kalman filter defined in what follows.

[4]The 4D unobservable subspace is induced by unknown initial conditions in 3D translational dynamics and the heading (yaw) angle of the robot in the inertial (world) frame.

Consider the robot's relative position to a stationary object of interest, denoted by $\mathbf{p}_i =: \mathbf{p}$ in[5] (6.2.7), the robot's velocity $\boldsymbol{v}$, and finally a common reference velocity, denoted by $\boldsymbol{v}^{\text{ref.}}$. We let the Kalman filter state $\hat{\mathbf{x}}_k = \text{col}\,(\widehat{\mathbf{p}}, \widehat{\mathbf{v}}) \in \mathbb{R}^6$ be the estimation of $\mathbf{p}$ and $\mathbf{v} = \boldsymbol{v} - \boldsymbol{v}^{\text{ref.}}$, with the covariance $\mathbf{P}_k$, and the update rules as follows:

$$\hat{\mathbf{x}}_{k|k-1} = F\hat{\mathbf{x}}_{k-1}, \qquad \mathbf{P}_{k|k-1} = F\mathbf{P}_{k-1}F^\top + \mathbf{Q}, \tag{6.2.8a}$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k|k-1} + \bar{\beta}_k \mathbf{K}_{\text{pos}}(y_{\text{pos}} - \text{C}_{\text{pos}}\hat{\mathbf{x}}_{k|k-1}) + \mathbf{K}_{\text{vel}}(y_{\text{vel}} - \text{C}_{\text{vel}}\hat{\mathbf{x}}_{k|k-1}), \tag{6.2.8b}$$

$$\mathbf{P}_k = \mathbf{P}_{k|k-1} - \bar{\beta}_k \mathbf{K}_{\text{pos}}\text{C}_{\text{pos}}\mathbf{P}_{k|k-1} - \mathbf{K}_{\text{vel}}\text{C}_{\text{vel}}\mathbf{P}_{k|k-1}, \tag{6.2.8c}$$

$$\mathbf{K}_\bullet = \mathbf{P}_{k|k-1}\text{C}_\bullet^\top \mathbf{S}_k^{-1},$$

$$\mathbf{S}_k = \left(\text{C}_\bullet \mathbf{P}_{k|k-1}\text{C}_\bullet^\top + \mathbf{R}_i^\bullet\right), \quad \bullet \in \{\text{pos}, \text{vel}\}, \tag{6.2.8d}$$

where $F = \begin{bmatrix} I_3 & T_s I_3 \\ \mathbf{0} & I_3 \end{bmatrix}$, $\mathbf{Q} = \begin{bmatrix} \sigma_{\text{pos}}^2 I_3 & \mathbf{0} \\ \mathbf{0} & \sigma_{\text{vel}}^2 I_3 \end{bmatrix}$, $\text{C}_{\text{pos}} = \begin{bmatrix} I_3 & \mathbf{0} \end{bmatrix}$, $\text{C}_{\text{vel}} = \begin{bmatrix} \mathbf{0} & I_3 \end{bmatrix}$, and $\bar{\beta}_k = (1 - \beta_k) \in \{0, 1\}$ is a binary random variable that quantifies the availability of relative position measurements $y_{\text{pos}} = \mathbf{p}$ obtained using the perception data as described in (6.2.7), while the velocity measurements $y_{\text{vel}} = \mathbf{v} = \boldsymbol{v} - \boldsymbol{v}^{\text{ref.}}$ are constantly available from the VIO module. In other words, $\beta_k = 1$ at $t_k \in \mathbb{R}_{\geq 0}$ corresponds to the case of missed measurements of (6.2.7) due to an adversarial image attacks. Therefore the rate of missed measurements (i.e. the distribution of β_k) is directly influenced by the rate of successful adversarial misclassification as well as by the magnitude of mislocalization errors in (6.2.7).

We note that the adversarially intermittent observation model in (6.2.8) is adopted from the formulation of Kalman filter with intermittent measurements transmitted over wireless networks [122, 11, 140]. Additionally, the fusion of VIO and

[5]For notational brevity and with a slight abuse of notation, we will drop the subscript i in this section.

perception data using a Kalman filter is similar to the approach in [41].

Gating and Data Association. Recall that the object detection model generates multiple bounding boxes and thus multiple relative position measurement candidates $\{\mathbf{p}\}_{o=0}^{m}$'s are available for the Kalman filter in (6.2.8) through relative localization (6.2.7) with the uncertainty quantified by $\mathbf{R}_i^{\text{pos}}$. To reduce the number of candidate measurements for the Kalman filter, we use the Mahalanobis distance [6, 100] to select an admissible subset of relative position measurements that are close to the tracked relative position. This is achieved through gating as follows:

$$V = \left\{ \mathbf{p}_o \mid (1-\beta_k)\,(\mathbf{p}_o - \widehat{\mathbf{p}})^\top \mathbf{S}_k^{-1} (\mathbf{p}_o - \widehat{\mathbf{p}}) \leq \tau_{\text{p}}^2 \right\}, \tag{6.2.9}$$

where β_k and innovation covariance $\mathbf{S}_k$ are given in (6.2.8), and is τ_{p} is the gating threshold. We then associate the relative position with the minimum Mahalanobis distance as the new measurement for the Kalman filter (see Fig. 6.2).

Remark 6.2.2. ***(Stability of Kalman Filter with Adversarial Measurements).*** *The stability of the Kalman filter in* (6.2.8) *is influenced by both the system dynamics and the characteristics of adversarial measurements. First, the second-order dynamics of the system, represented by the matrix F in* (6.2.8)*, feature defective eigenvalues on the unit circle. This poses challenges for the stability analysis of the Kalman filter with intermittent measurements [84, 140]. Additionally, since the relative position measurements in* (6.2.7) *of the double-integrator system are subject to adversarial perturbations, the conditions for designing undetectable attacks are satisfied [58, 64], [83, Thrm. 2], posing fundamental challenges (See also Remark 6.2.1). Moreover, the probability distribution of β_k, which reflects the success rate of adversarial image attacks on the relative localization measurements* (6.2.7)*, is unknown a priori. Previous studies have investigated the stability of the Kalman filter*

under the assumption that $\bar{\beta}_k = (1 - \beta_k)$ follows a Bernoulli random process [122] or the Gilbert-Elliott model [140]. Generally, there exists a critical threshold for the rate of missed measurements (i.e., the probability distribution of $\bar{\beta}_k = 0$) below which the estimation error covariance remains bounded with high probability, while it becomes unbounded above this threshold.

Remark 6.2.3. ***(Kalman Filter with Adversarial Training).*** *An alternative approach to* (6.2.8)*, originally proposed in [149], can be adapted to adversarially train a Kalman gain that allows for robustness to measurement perturbations $\delta\mathbf{p}_i$. The approach in [149], however, does not consider adversarially intermittent measurements, which are modeled by the binary variable $\bar{\beta}_k \in \{0, 1\}$ in* (6.2.8)*.*

6.2.6 Resilient Multi-Robot Coordination

Consider a multi-robot system consisting of $N \geq 3$ mobile robots (quadrotors) with states $\mathbf{x}_i = \mathrm{col}\,(\widetilde{\boldsymbol{p}}_i, \widetilde{\boldsymbol{v}}_i) \in \mathbb{R}^6$, with $\widetilde{\boldsymbol{p}}_i = \mathbf{p}_i - \boldsymbol{p}_i^{\star}$ and $\widetilde{\boldsymbol{v}}_i = \boldsymbol{v}_i - \boldsymbol{v}^{\mathrm{ref.}}$, $\forall\, i \in \mathcal{V} = \{1, \ldots, N\}$. Similar to Chapter 4, one can obtain a reduced-order model of quadrotor dynamics as follows:

$$\Sigma_i : \dot{\mathbf{x}}_i = \overbrace{\begin{bmatrix} \mathbf{0} & I_3 \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}^{A} \mathbf{x}_i + \overbrace{\begin{bmatrix} \mathbf{0}_3 \\ I_3 \end{bmatrix}}^{B} \begin{bmatrix} \boldsymbol{u}_i(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_i) \\ -g + \frac{f_i}{m} \end{bmatrix}, \tag{6.2.10}$$

$$\boldsymbol{u}_i(\mathbf{x}_i, \mathbf{x}_j, \boldsymbol{\theta}_i) = g \underbrace{\begin{bmatrix} \cos\phi_i & \sin\phi_i \\ \cos\phi_i & -\cos\phi_i \end{bmatrix}}_{R(\psi_i)} \begin{bmatrix} \Delta\theta_i^* \\ \Delta\phi_i^* \end{bmatrix}, \tag{6.2.11}$$

for which the control commands are obtained using feedback linearization as follows:

$$\begin{bmatrix} \Delta\theta_i^* \\ \Delta\phi_i^* \end{bmatrix} = \frac{1}{g} R^{-1}(\psi_i^*)\mathbf{u}_i^*, \tag{6.2.12}$$

with the coordination protocol

$$\begin{aligned} \mathbf{u}_i^* &= -\alpha \sum_{j\in\mathcal{V}} a_{ij}^{\sigma(t)}(\widetilde{\boldsymbol{p}}_i - \widetilde{\boldsymbol{p}}_j) - \gamma\widetilde{\boldsymbol{v}}_i + \dot{\boldsymbol{v}}^{\text{ref.}}, \\ &= -\alpha \sum_{j\in\mathcal{V}} a_{ij}^{\sigma(t)}((\mathbf{p}_i - \boldsymbol{p}_i^\star) - (\mathbf{p}_j - \boldsymbol{p}_j^\star)) - \gamma(\boldsymbol{v}_i - \boldsymbol{v}^{\text{ref.}}) + \dot{\boldsymbol{v}}^{\text{ref.}}, \end{aligned} \tag{6.2.13}$$

where we obtain the robot's relative position $\mathbf{p}_i$ and velocity $\boldsymbol{v}_i$ from the Kalman filter in (6.2.8) and the neighbors' position $\mathbf{p}_j$'s (or $\widetilde{\boldsymbol{p}}_j$) from the communication network (see Fig. 6.2). With a slight abuse of notation, the right-hand side of (6.2.13) refers to the 2D positions in the x-y plane of the common reference frame $\{\mathcal{W}\}$. The robots can coordinate at the same altitude through altitude consensus or other approaches [5].

Effect of adversarial image attacks. Recall (6.2.7) that models the 3D localization error caused by adversarial image attacks on the i-th robot. Then (6.2.13) can be represented as

$$\mathbf{u}_i^* = -\alpha \sum_{j\in\mathcal{V}} a_{ij}^{\sigma(t)}(\widetilde{\boldsymbol{p}}_i^{\text{n}} - \widetilde{\boldsymbol{p}}_j^{\text{n}}) - \gamma\widetilde{\boldsymbol{v}}_i + \dot{\boldsymbol{v}}^{\text{ref.}} \overbrace{-\alpha \sum_{j\in\mathcal{V}} a_{ij}^{\sigma(t)} \delta\mathbf{p}_i}^{\mathbf{u}_i^{\text{a}}}, \tag{6.2.14}$$

which implies adversarial image attacks on the i-th robot perception can be modeled as bounded attacks on the control channel of the i-th robot that will be propagated to the neighboring robots as well. In Chapter 5, we designed an observer-based monitoring framework that allows for detecting robots with compromised control channels

(Also, see Fig. 6.2). We refer to the MSR-like algorithms as alternative approaches to discarding compromised agents in a consensus-based coordination problem [37, 67].

6.3 Multi-Robot Platform Development

We have developed two multi-robot platforms for vision-based coordination with learned perception modules over wireless communication networks. The first platform includes a software package, `TelloSwarm+`, which we developed[6] for Tello-EDU[7] quadrotors with SDK 3. The second platform includes two custom-built quadrotors equipped with VOXL flight kit[8] manufactured with ModalAI. We are currently developing a software package[9] for the custom-built quadrotors (see Fig. 6.4).

6.3.1 Communication Network Architecture

A low-latency communication network is central to the safe operation of multi-robot systems with decentralized and/or distributed tasks. This communication encompasses information transmission both between the robots and between the robots and a workstation (PC). We refer to [47, 46] for a critical review of communication networks and their open problems in the context of multi-robot systems. Here, we employ the server-client model which has been demonstrated to be an efficient communication approach for multi-robot systems with collaborative tasks [132].

In what follows we elaborate on the communication network of Tello-EDU quadrotors in `TelloSwarm+` (see Fig. 6.5). It is important to note that Tello-EDUs are small quadrotors that can communicate over 2.4 GHz Wi-Fi but cannot run algorithms onboard. Therefore, an efficient communication network must be established

[6] `https://github.com/SASLabStevens/TelloSwarm`
[7] `https://www.ryzerobotics.com/tello-edu`
[8] `https://www.modalai.com/`
[9] `https://github.com/SASLabStevens/AutonomyStack`

to replicate peer-to-peer communication among the robots on a centralized PC, allowing distributed algorithms to be executed for each robot in a controlled setting. To this end, we use a server-client model to establish the communication network for `TelloSwarm+`.

Our network includes one UDP client that sends control commands from a workstation PC to the robots, and two UDP servers on two *threads* to receive the robot's onboard state information (e.g., IMU data) and the acknowledgment message for received control commands. Additionally, onboard video streams from each drone are available through separate *threads* running an OpenCV UDP stream module[10]. Overall, the network runs $(2 + N)$ *threads* for N Tello-EDU quadrotors.

We note that using *threads* as independent units of execution within a process with shared memory enables not only lightweight, low-latency communication between the robots and the PC but also efficient inter-robot message passing, which is essential for decentralized control and monitoring algorithms. It is also noteworthy that an alternative ROS implementation of this architecture, particularly for video streaming, encounters considerable latency due to the more computation-intensive nature of message passing between ROS nodes. For a detailed latency analysis of ROS, we refer to [62, 90].

6.4 Experimental Results

We conducted 15 experiments to evaluate the framework shown in Fig. 6.2, excluding the adversary detection component, using the developed `TelloSwarm+` platform[11]. The objective is to evaluate how adversarial image attacks targeting the learned perception module (object detection), with varying success rates, induce different

[10] `VideoCapture()`

[11] The open-source code is available at `https://github.com/SASLabStevens/TelloSwarm`.

levels of degeneracy in the relative localization, state estimation, and coordination of robots that rely on the compromised perception module (See Remark 6.2.2).

6.4.1 Custom-trained Object Detection Model

We fine-tuned a YOLOv7 model [136] to extend its detection capability from the 80 classes of the COCO-MS dataset to 82 classes, including our drones (the quadrotors in Fig. 6.4) and jackal-UGV[12].

Custom dataset. We collected 720 images of the quadrotors shown in Fig. 6.4 and jackal-UGV in the flight area of the Safe Autonomous Systems Lab at Stevens as well as some high-fidelity synthetic images of the parrot ar drone 2.0 (quadrotor) in the AirSim simulation environments. The images were split into 500 for training, 150 for validation, and 70 for testing. We then augmented the training dataset to a total of 1,500 images. Our dataset is available as open source[13]. Additionally, for the rest of the 80 classes of COCO-MS, we used a mini training set[14] (25K images $\approx$ 20% of the original COCO dataset 2017) that has been shown to have a strong performance correlation with the original dataset [114].

Training procedure on the custom dataset. We first pre-trained the YOLOv7-tiny model using its original weights on our custom dataset for 15 epochs, with a batch size of 32 and a learning rate of 0.001. Next, we froze the backbone (the first 28 layers) and fine-tuned the pre-trained model for 50 epochs, with a batch size of 32 and a learning rate of 0.0001. During both the training and experimental phases, we used an image size of 640 $\times$ 640. The accuracy of the custom-trained model is reported in Fig. 6.6.

Adversarial Image Attacks. As discussed in Section 6.2.3, adversarial image

[12] `https://clearpathrobotics.com/jackal-small-unmanned-ground-vehicle/`
[13] `https://universe.roboflow.com/saslab/saslab-multirobot`.
[14] `https://github.com/giddyyupp/coco-minitrain`

attacks, regardless of their design method, can cause categorically similar adversarial effects that are misclassification [65, 10, 49], mislocalization [143, 54, 58, 21], and increased latency [121, 23] in learned perception models (e.g., object detection [143, 54] and pose estimation [23]). Therefore, we manually generate adversarial effects of varying severity to evaluate our framework shown in Fig. 6.2. This approach allows for resilience analysis of the proposed framework, independent of the specific adversarial image attack model and the targeted learned perception (object detection) model.

6.4.2 Perception-based Multi-Robot Coordination

Fig. 6.7 shows an overview of our experimental setup. In the experiments, two Tello-EDU quadrotors were equipped with VIO and a custom-trained YOLOv7 model and communicated over a wireless network as detailed in Fig. 6.5. To achieve higher quality, we use pose data from a Vicon motion capture system to simulate the VIO data. Each quadrotor then runs the framework outlined in Fig. 6.2 and detailed in Section 6.2 on a separate *thread* for 1,000 iterations, with each iteration taking an average of 35 milliseconds[15] on a workstation PC running Ubuntu 20.04 LTS.

In the experiments, the jackal-UGV is the point of interest $\boldsymbol{p}_r$ in the map. Each Tello-EDU quadrotor uses a custom-trained YOLOv7 object detection model to detect the jackal-UGV and then calculates its relative position to the detected jackal-UGV as detailed in Section 6.2.4. The quadrotors then coordinate using the control protocol defined in (6.2.13) with $\alpha = 0.72828$ and $\gamma = 1.09242$. In the x-direction of the common frame (see Fig. 6.7), the control protocol sets the common velocity

[15]The value, 35^{+74}_{-15} milliseconds per iteration, is reported under standard settings (i.e., no adversarial attack), associated with the experiment listed in the first row of Table 6.1). Adversarial attacks causing overload can increase this value to 41^{+100}_{-21} milliseconds per iteration, associated with the experiment listed in the second row of Table 6.2, or potentially higher.

Table 6.1: Adversarial Misclassification as Intermittent Measurements (False Negatives) - 11 Experiments

Adversary	**Performance Metrics**[1]				
$\beta_k \sim \text{Bin}(n,p)$	$\text{RMS}(\widetilde{\boldsymbol{p}}_{21}, \widehat{\widetilde{\boldsymbol{p}}}_{21})$	$\text{RMS}(\boldsymbol{p}^\star_{21}, \widehat{\boldsymbol{p}}_{21})$	$\text{RMS}(\boldsymbol{p}^\star_{21}, \boldsymbol{p}_{21})$	$\sup_{k\geq 1}\|\mathbf{P}_k\|_2$	$\sum_{k=1}^{1000}\|\mathbf{P}_k\|_2$
$n=0,\quad p=0$	0.16	0.06	0.18	0.09	41.46
$n=1000,\quad p=0.2$	0.29	0.08	0.30	1.05	56.92
$n=1000,\quad p=0.4$	0.15	0.09	0.17	0.40	75.20
$n=1000,\quad p=0.6$	0.13	0.09	0.11	1.40	124.88
$n=1000,\quad p=0.8$	0.10	0.10	0.09	1.40	231.77
$n=1000,\quad p=0.95$	1.99	0.62	1.49	11.88	1985.12
$n=200,\quad p=0.2$	0.07	0.06	0.05	0.60	87.87
$n=200,\quad p=0.4$	0.12	0.10	0.16	1.54	213.00
$n=200,\quad p=0.6$	0.38	0.11	0.42	2.54	294.12
$n=200,\quad p=0.8$	0.15	0.12	0.21	3.31	586.10
$n=200,\quad p=0.95$	0.55	0.32	0.55	12.86	3613.21

[1] Root mean square (RMS) was calculated for the 2D position in the x-y plane for $t \geq 10$ sec to exclude the effects of initial conditions.

reference $\boldsymbol{v}^{\text{ref.}} = 0$, and $\boldsymbol{p}^\star_{21} = \boldsymbol{p}^\star_2 - \boldsymbol{p}^\star_1 = -0.9$ meters. In the y-direction, the control protocol sets the common velocity reference $\boldsymbol{v}^{\text{ref.}} = 2\pi f \cos(\frac{2\pi}{500}k)$, where $f = 0.1$ and $k \in [0, 1000]$, and $\boldsymbol{p}^\star_{21} = \boldsymbol{p}^\star_{12} = 0$. We set the IoU and confidence thresholds of the object detection model to 0.45 and 0.15, respectively, at inference time. The Kalman filter in (6.2.8) is initialized with $\hat{\mathbf{x}}_{0|-1} = \mathbf{0}$, $\mathbf{P}_{0|-1} = \text{diag}\,(I_3, 0.05I_3)$, $T_\text{s} = t_k - t_{k-1} \geq 0.02$ in the state transition matrix F, $\sigma^2_{\text{pos}} = 0.05$, $\sigma^2_{\text{vel}} = 0.04$ in the covariance of the process noise $\mathbf{Q}$, and finally $\bar{\epsilon} = 0.4$, $\underline{\epsilon} = 0.01$ for $\mathbf{R}^{\text{pos}}_i$ in (6.2.7b) and $\mathbf{R}^{\text{vel}}_i = 0.078I_3$. We also set the gating threshold $\tau_\text{p} = 2.4476$ in (6.2.9).

Experiment Set I (Adversarial Misclassification as Sporadic Measurements). We conducted a set of 11 experiments, listed in Table 6.1, to evaluate the degenerative effect of adversarial misclassification (6.2.1), modeled as sporadic (intermittent) measurements, on the perception-based relative localization and state estimation in the framework shown in Fig. 6.2. The perception (YOLOv7 object detection) model of agent (quadrotor) 2, shown in Fig. 6.7, is subject to adversarial misclassification.

The success rate of adversarial misclassification (equiv. the failure rate of intermittent measurements) is quantified by the probability distribution of the binary variable $\beta_k \in \{0, 1\}$ in (6.2.8) and (6.2.9). We let β_k follow a binomial distribution, $\beta_k \sim \text{Bin}(n, p)$, with n trials and a success probability[16] of p. As detailed in Remark 6.2.2, the probability distribution of $\bar{\beta}_k = (1 - \beta_k)$, which reflects the rate of intermittent measurements, has a direct degenerative effect on the stability of Kalman filter in (6.2.8).

In the experiments listed in Table 6.1, the jackal-UGV (the reference point for coordination) was adversarially misclassified as an airplane by the compromised perception of agent (quadrotor) 2, which caused missed measurements, represented by $\beta_k = 1$ in (6.2.8) and (6.2.9). From the induced 2-norm of the state estimation covariance matrix $\mathbf{P}_k$ of the Kalman filter, reported in the last column of Table 6.1, one can conclude that as the rate of missed measurements increases (i.e. the probability of adversarial misclassification p in the Adversary column), the uncertainty in state estimation correspondingly increases. Additionally, for a given success probability p of adversarial misclassification, experiments with fewer trials ($n = 200$ compared to $n = 1000$, as listed in the Adversary column) have longer consecutive periods of misclassification, which causes a larger increase in state estimation uncertainty, as reported in the last column. This effect is also demonstrated in Figs. 6.8 and 6.9. Fig. 6.8 shows the evolution of the induced 2-norm of state estimation covariance matrix $\mathbf{P}_k$ in (6.2.8) over time over time for three cases of no adversarial attack, and adversarial attacks at two different rates, corresponding to the first, third, and seventh rows of Table 6.1. The induced norm of the state estimation covariance matrix serves as a metric for evaluating the peak-covariance stability of the Kalman filter

[16]The probability of success of a single trial, p, in the binomial distribution represents the probability of successful misclassification in (6.2.1) and equivalently represents the probability of single failed relative localization measurement (i.e. $\bar{\beta}_k = 1$) for the Kalman filter in (6.2.8).

with intermittent measurements [140]. Additionally, Fig. 6.9 shows the evolution of the trace of the state estimation covariance matrix $\mathbf{P}_k$, as another stability metric [122], together with timestamped image frames and perception data for the adversarial case of $\beta_k \sim \text{Bin}(n = 200, p = 0.4)$ listed in the seventh row of Table 6.1. Finally, Figs. 6.10 and 6.11 show the coordination trajectories of the quadrotors for the standard and adversarial cases, respectively, corresponding to the first and seventh rows of Table 6.2. Overall, the results suggest the higher rate of adversarial misclassification causes a larger level of degradation in the Kalman filter (6.2.8) that relies on the localization measurements (6.2.7). In our experiment, we observed that a higher rate of missed measurements caused the compromised agent (quadrotor 2) to stall (hover) for larger periods (see (6.2.13)), leading to a drift in the coordination. Also, the experiment listed in the fifth row of Table 6.1 resulted in a crash. However, it is also important to note that the proposed framework, shown in Fig. 6.2, significantly reduced the level of degradation and maintained the system's stability in the presence of adversarial misclassification that caused intermittent measurements. For instance, despite the presence of missed measurements and spurious measurements, Fig. 6.11a shows the successful state estimation of the second robot's relative position to the jackal-UGV, denoted by $\mathbf{p}_2$, that is used in the coordination protocol (6.2.13).

Experiment set II: Adversarial Mislocalization as Spurious Measurements. We conducted a set of 4 experiments, listed in Table 6.2, to evaluate the degenerative effect of adversarial mislocalization (6.2.1) that cause spurious measurements and adversarial overload [23], on the perception-based relative localization, state estimation (6.2.8) and gating (6.2.9) in the framework shown in Fig. 6.2. The perception (YOLOv7 object detection) model of agent (quadrotor) 2, shown in Fig. 6.7, is subject to adversarial mislocalization at different rates. In the experiments, the bounding

Table 6.2: Adversarial Mislocalization as Spurious Measurements (False Positives) - 4 Experiments

Adversary[1]	**Performance Metrics**[2]				
δbox	RMS($\widetilde{\boldsymbol{p}}_{21}, \widehat{\widetilde{\boldsymbol{p}}}_{21}$)	RMS($\boldsymbol{p}^{\star}_{21}, \widehat{\boldsymbol{p}}_{21}$)	RMS($\boldsymbol{p}^{\star}_{21}, \boldsymbol{p}_{21}$)	$\sup_{k\geq 1} \|\mathbf{P}_k\|_2$	$\sum_{k=1}^{1000} \|\mathbf{P}_k\|_2$
$b = 10,\ \ q = \pm 15\%$	1.07	0.12	1.07	0.07	40.40
$b = 10,\ \ q = \pm 30\%$	0.65	0.20	0.59	1.20	44.32
$b = 10,\ \ q = \pm 45\%$	1.12	0.25	0.99	1.25	46.08
$b = 10,\ \ q = \pm 75\%$	0.80	0.21	0.74	1.25	45.71

[1] $b = 10$ spurious bounding boxes were adversarially generated by perturbing the nominal detected bounding box around the object of interest by $q \in \{\pm 15\%, \pm 30\%, \pm 45\%, \pm 75\%\}$. Additionally, their probability confidence pr was set 10% more than the nominal one.

[2] Root mean square (RMS) was calculated for the 2D position in the x-y plane for $t \geq 10$ sec to exclude the effects of initial conditions.

boxes of detected jackal-UGV (the reference point for coordination) were adversarially mislocalized as described[17] in the footnote of Table 6.2.

Figs. 6.12 and 6.13 shows the results associated with the experiment listed in the second row of Table 6.2. As shown in Fig. 6.13, adversarial mislocalization can generate a significant number of spurious bounding boxes, leading to a substantial increase in spurious relative position measurements (6.2.7). These spurious measurements impose a computational overhead on the components of the perception module, shown in Fig. 6.2, which resulted in latency for the compromised quadrotor. Additionally, the adversarial mislocalization caused the failure of the data association module at $t \approx 11$, shown in Fig. 6.13a. This failure led to a large error in the Kalman filter's estimation of the relative measurements, resulting in a significant drift in multi-robot coordination.

Experiment set III: Mixed Adversarial Misclassification and Mislocalization. We conducted an experiment, listed in Table 6.3, to evaluate the degenerative effect of both adversarial misclassification and mislocalization (6.2.1) that cause spo-

[17] We note that the perturbations applied to the nominal bounding boxes were calculated based on the top-left and bottom-right corners, (x_1, y_1, x_2, y_2), of the bounding box, rather than $(\mathrm{x}_{\mathbf{I}}, \mathrm{y}_{\mathbf{I}}, \mathrm{w}_{\mathbf{I}}, \mathrm{h}_{\mathbf{I}})$ coordinates.

Table 6.3: The Effect of Mixed Adversarial Misclassification and Mislocalization

Adversaries[1]	**Performance Metrics**[2]				
$\beta_k \sim \text{Bin}(n,p)$ & δbox	$\text{RMS}(\widetilde{\boldsymbol{p}}_{21}, \widehat{\widetilde{\boldsymbol{p}}}_{21})$	$\text{RMS}(\boldsymbol{p}^{\star}_{21}, \widehat{\boldsymbol{p}}_{21})$	$\text{RMS}(\boldsymbol{p}^{\star}_{21}, \boldsymbol{p}_{21})$	$\sup_{k\geq 1} \|\mathbf{P}_k\|_2$	$\sum_{k=1}^{1000} \|\mathbf{P}_k\|_2$
$n = 200, \quad p = 0.2$ $b = 5, \quad q = \pm 75\%$	0.23	0.12	0.22	1.55	103.100

[1] $b = 5$ spurious bounding boxes were adversarially generated by perturbing the nominal detected bounding box around the object of interest by $q = \pm 75\%$ and by increasing the probability confidence by 10%.

[2] Root mean square (RMS) was calculated for the 2D position in the x-y plane for $t \geq 10$ sec to exclude the effects of initial conditions.

radic and spurious measurements, on the perception-based relative localization, state estimation (6.2.8) and gating (6.2.9) in the framework shown in Fig. 6.2. The perception (YOLOv7 object detection) model of agent (quadrotor) 2, shown in Fig. 6.7, is subject to adversarial attacks. In the experiments, the bounding boxes of detected jackal-UGV (the reference point for coordination) were adversarially misclassified as an airplane, which caused missed measurements, represented by $\beta_k = 1$ in (6.2.8) and (6.2.9). Additionally, the bounding boxes of detected jackal-UGV were adversarially mislocalized as described in the footnote of Table 6.3. Adversarial misclassification and mislocalization occur simultaneously at some time instances during the experiment.

Figs. 6.14 and 6.15 show the result of the experiment. The evolution of the trace of the state estimation covariance matrix $\mathbf{P}_k$, together with timestamped image frames and perception data subject to adversarial mislocalization as well as adversarial misclassification with $\beta_k \sim \text{Bin}(n = 200, p = 0.2)$ are shown in Fig. 6.14. One can observe the degenerative effect of missed measurements as peaks in the $\text{Trace}(\mathbf{P}_k)$. Fig. 6.15 shows the coordination trajectories of the quadrotors. This experiment demonstrates the effectiveness of the proposed framework, shown in Fig. 6.2, in mitigating degradation caused by adversarial image attacks and providing an estimation of relative positions despite adversarially induced sporadic (intermittent) and spurious measurements.

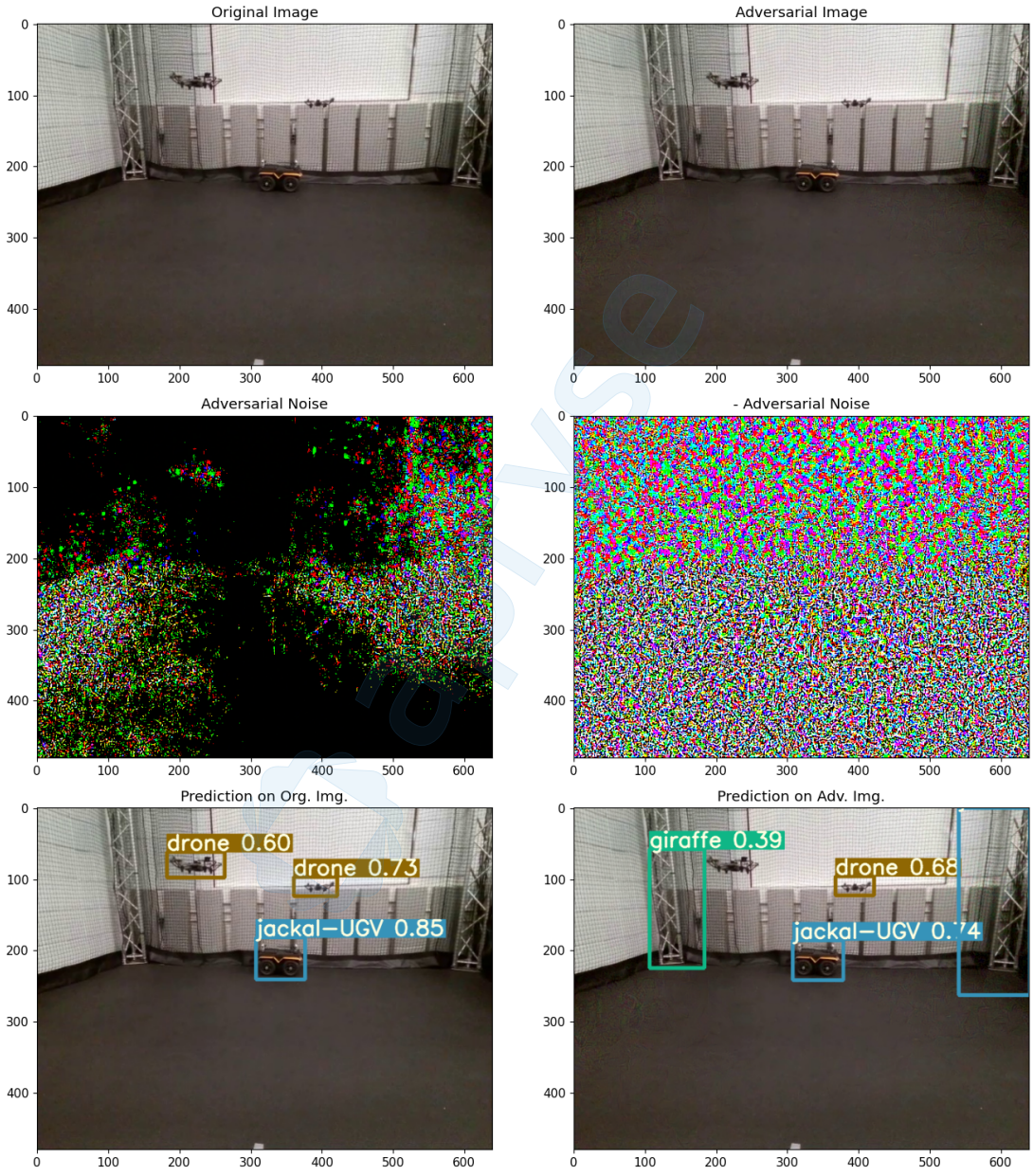
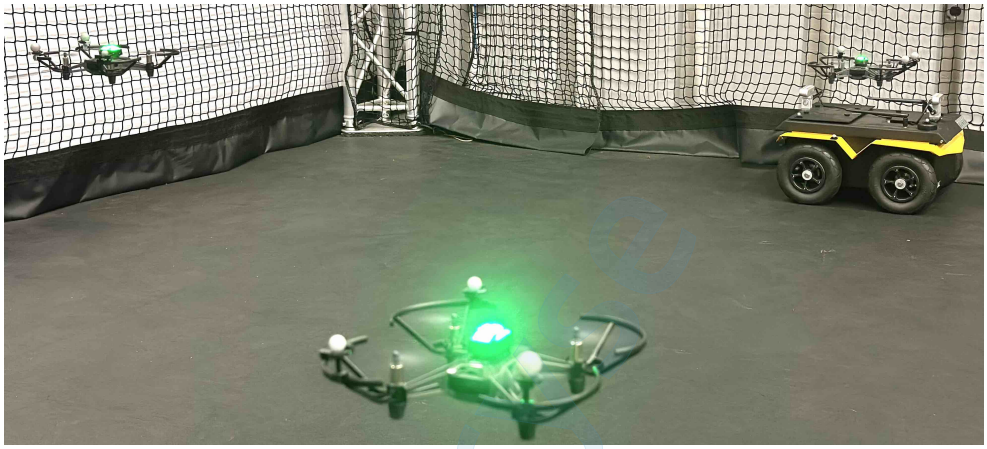


Figure 6.3: The effect of FGSM adversarial image attack on YOLOv7 object detection.



(a) TelloSwarm+ Platform

(b) Custom-built VOXL-equipped Platform

Figure 6.4: Multi-robot Platforms. (a) The `TelloSwarm+` platform is an extension of our prior work [4] with vision capability and efficient multi-threaded wireless communication capability. (b) The VOXL-equipped platform is a custom-built quadrotor that allows for the onboard implementation of control, monitoring, and deep learning algorithms.

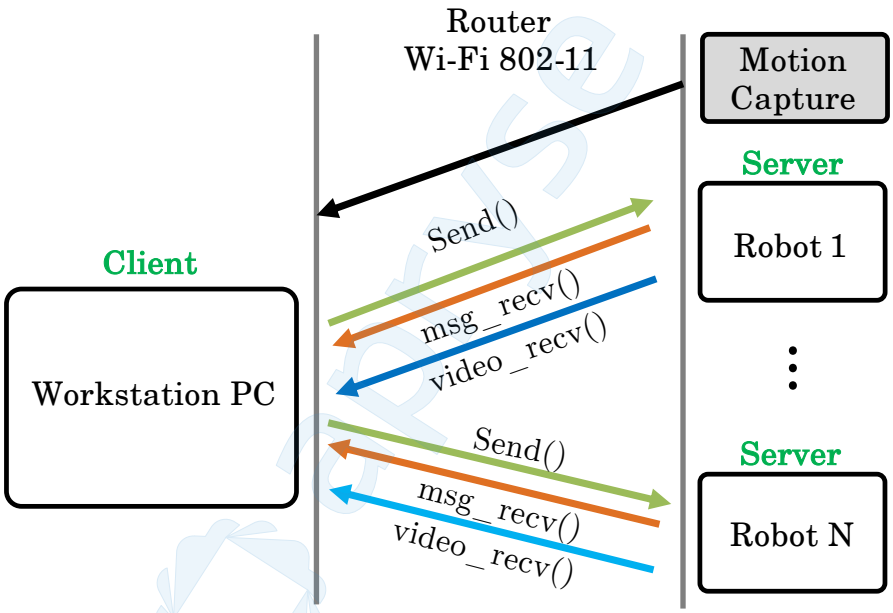


Figure 6.5: Multi-robot communication architecture for `TelloSwarm+`. The network establishes a multithreaded server-client architecture over Wi-Fi 802.11 using the UDP protocol to achieve fast, low-latency communication with each robot. A motion capture system provides the ground truth poses of the robots.

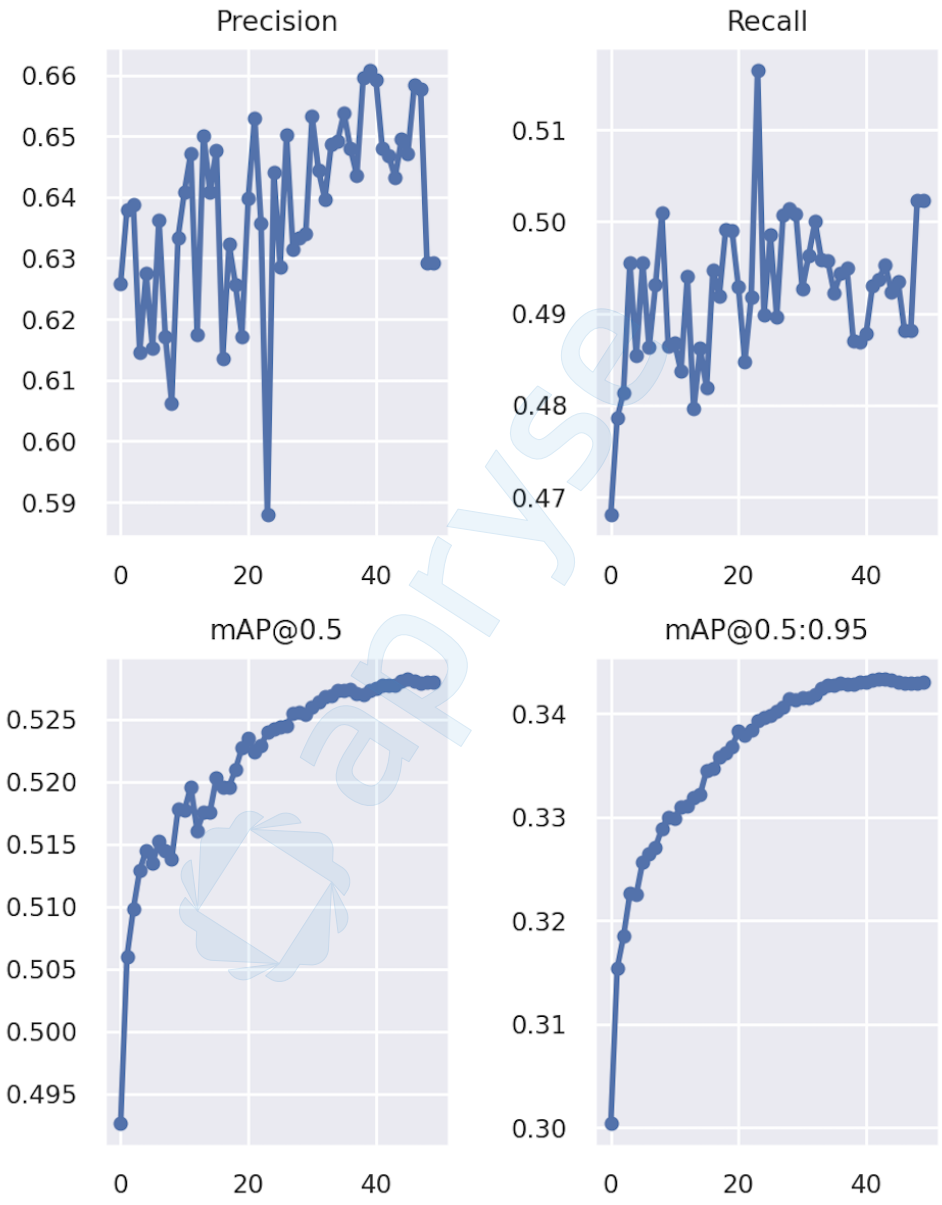


Figure 6.6: The accuracy of custom-trained YOLOv7 model. mAP (mean average precision) is calculated based on the Intersection over Union (IoU) between the detected bounding boxes and ground-truth bounding boxes, with IoU thresholds of 0.5 and ranging from 0.5 to 0.95.

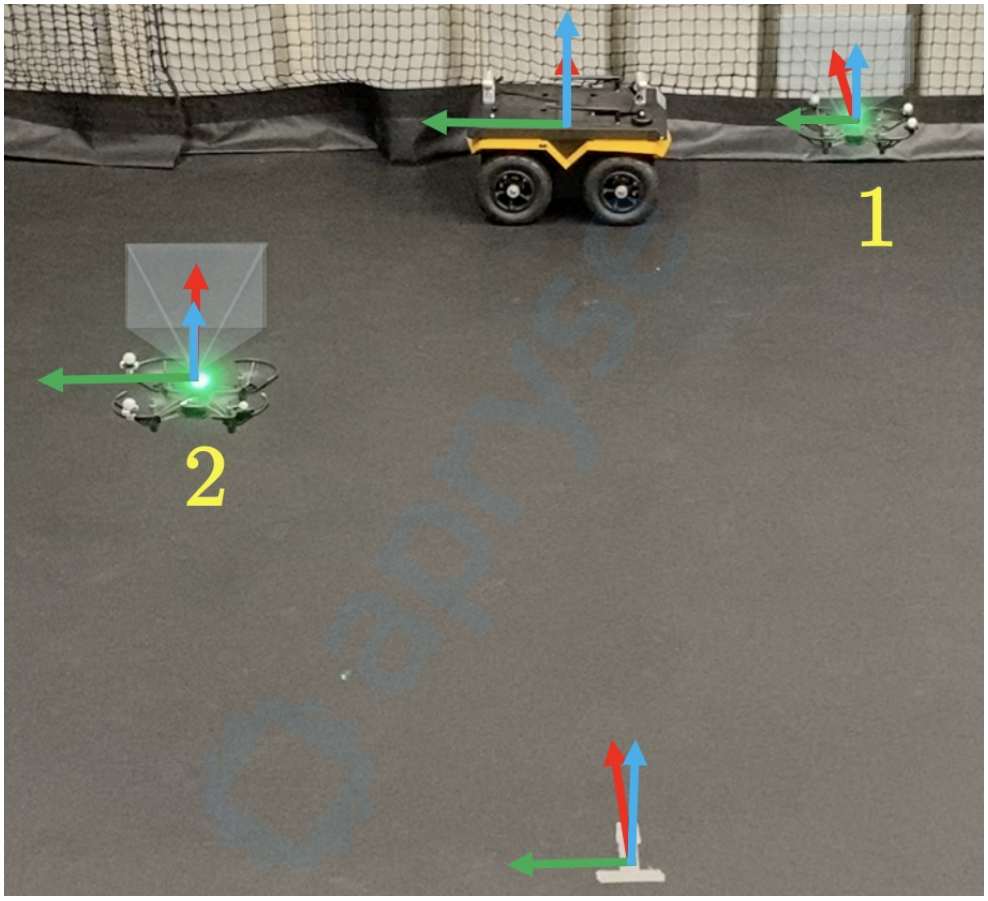


Figure 6.7: Experimental setup for perception-based multi-robot coordination subject to adversarial image attacks. The experiments use the framework shown in Fig. 6.2. Two Tello-EDU quadrotors perform relative localization with respect to the jackal-UGV using their respective VIO and object detection model that detects the jackal-UGV. The quadrotors also coordinate their estimated relative positions through the control protocol (6.2.13).

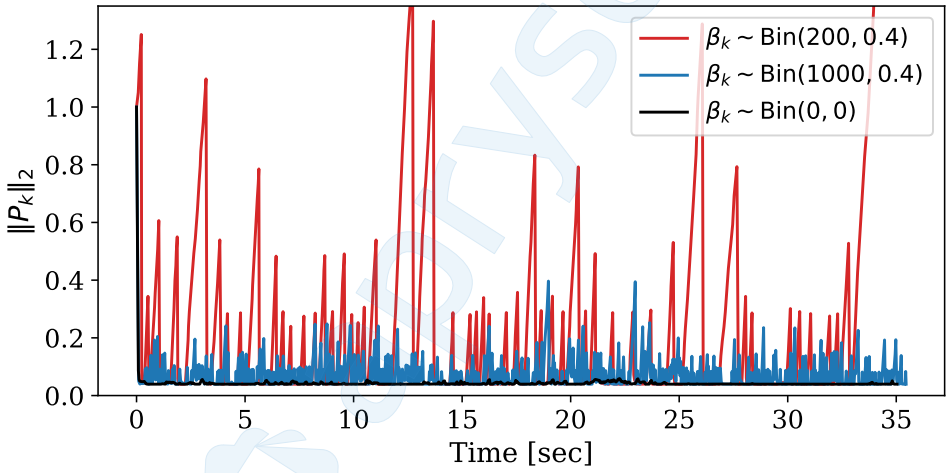
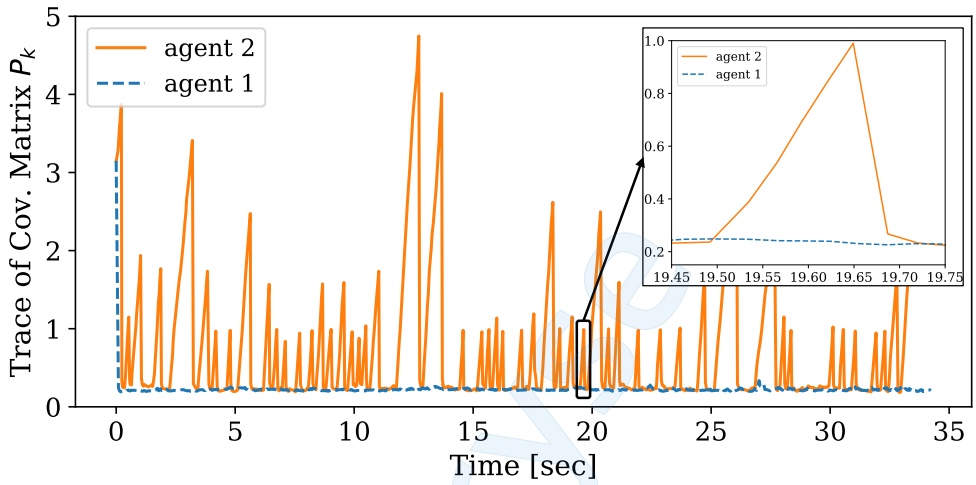
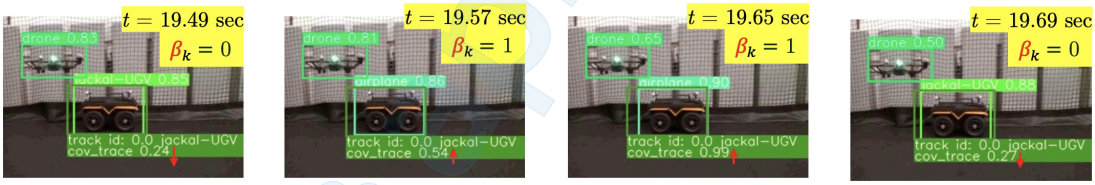


Figure 6.8: The induced 2-norm of state estimation covariance to adversarial misclassification as intermittent measurements at different rates. see Table 6.1 for more comparisons.

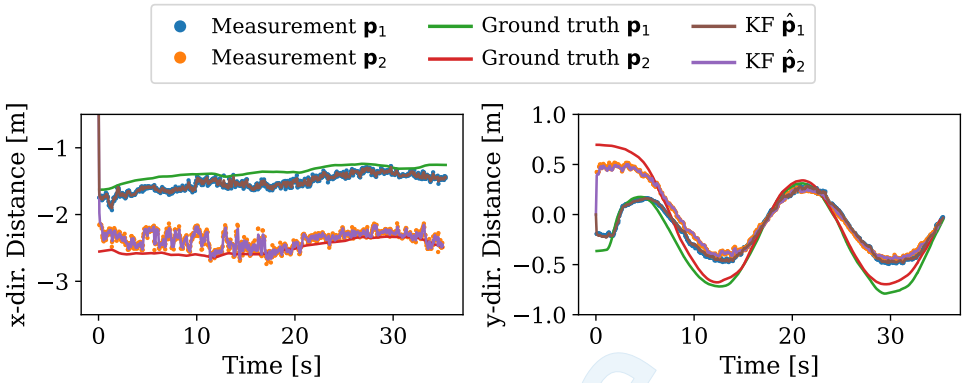


(a) The trace of state estimation covariance matrix

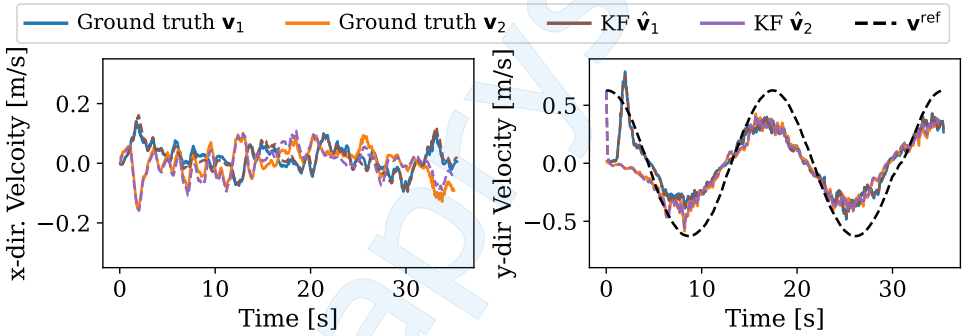


(b) Timestamped perception and relative localization of agent 2

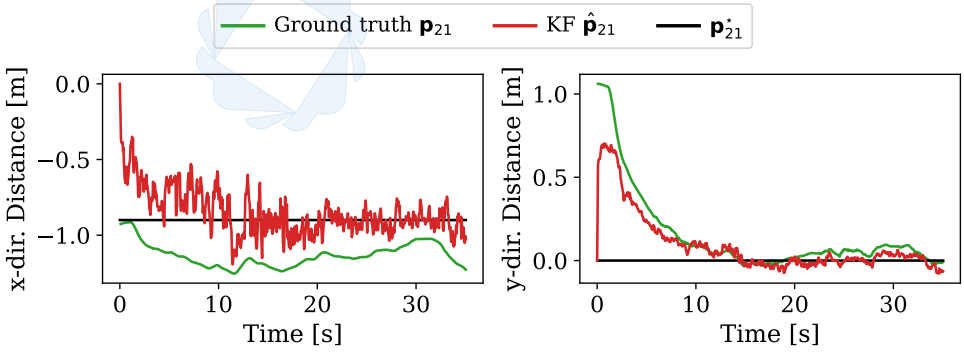
Figure 6.9: Results from a two-agent perception-based coordination experiment using the framework shown in Fig. 6.2, subject to adversarial misclassification as detailed in the seventh row of Table 6.1. The peaks in (a) reflect the degenerative effect of adversarial misclassification inducing missed measurements in the Kalman filter (6.2.8). (b) The boxes with labels on top are the detections from the custom-trained YOLOv7 model, while the green boxes with labels underneath are calculated by projecting the 3D relative position estimations from the Kalman filter into the image space to determine the box's center, and by using the object's known size to compute the box's width and height in the image. Additionally, the image frames in (b) have been cropped for better visualization. The original camera image size was 640 × 480 pixels.



(a) Position Trajectories

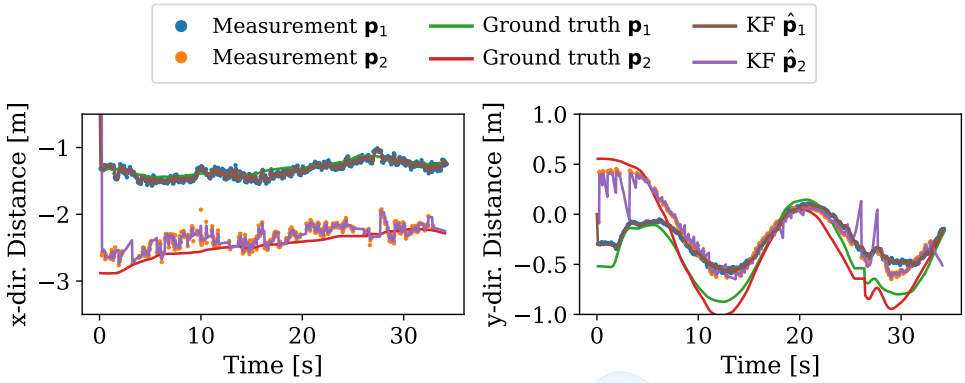


(b) Velocity Trajectories

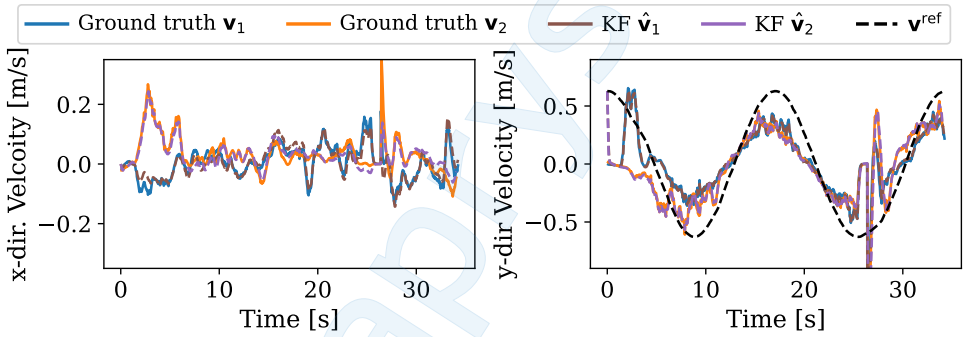


(c) Trajectories of Relative Position Coordination

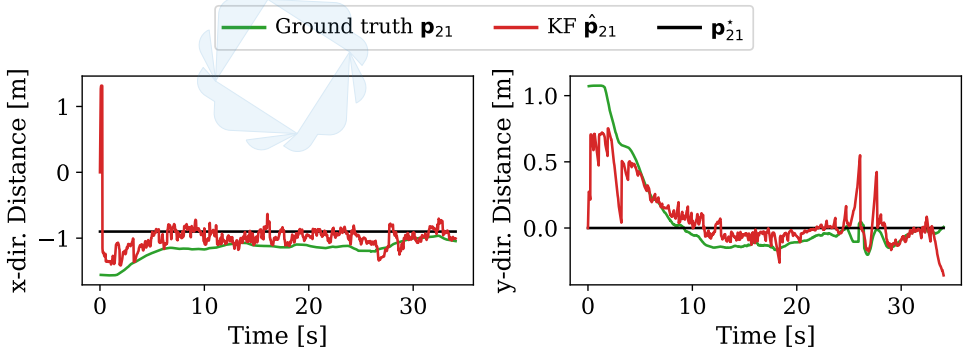
Figure 6.10: Results from a two-agent perception-based coordination experiment in standard settings (i.e., no adversarial attacks on the perception module), using the framework illustrated in Fig. 6.2. Performance metrics and comparisons for this experiment are detailed in the first row of Table 6.1.



(a) Position Trajectories



(b) Velocity Trajectories



(c) Trajectories of Relative Position Coordination

Figure 6.11: Results from a two-agent perception-based coordination experiment with adversarial misclassification in the perception module, using the framework illustrated in Fig. 6.2. The adversarial misclassification rate is modeled by a binomial distribution $\beta_k \sim \text{Bin}(n = 200, p = 0.4)$ in (6.2.8). Performance metrics and comparisons for this experiment are detailed in the seventh row of Table 6.1.

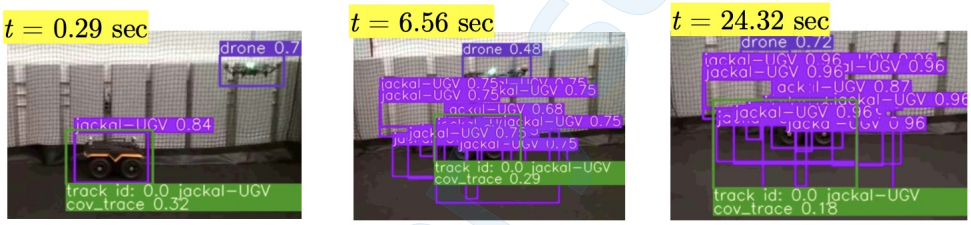
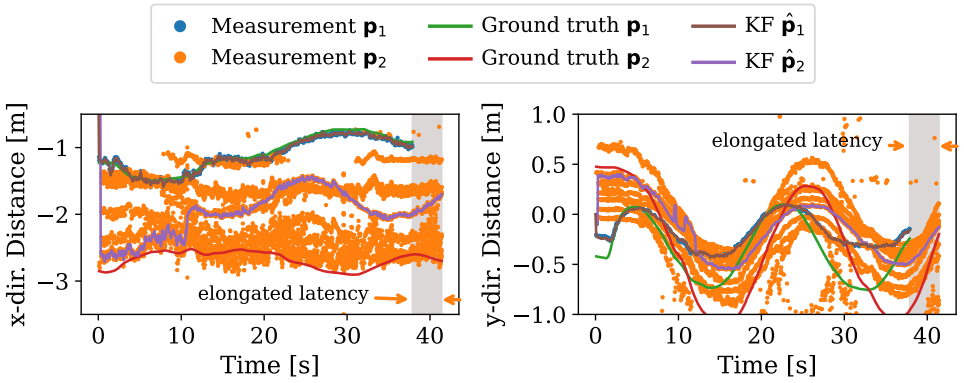
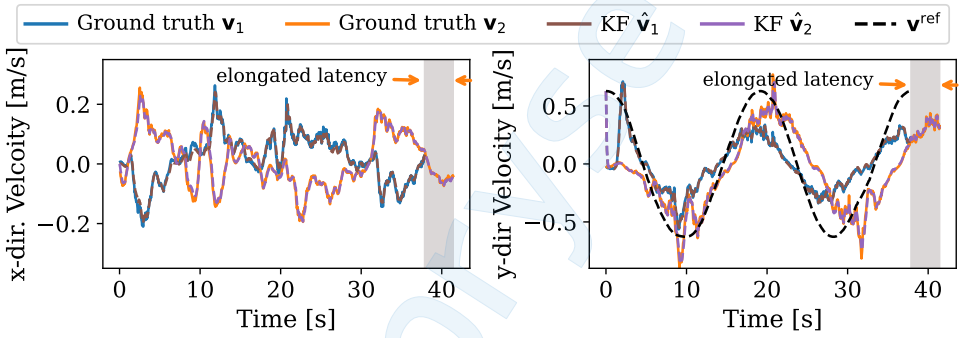


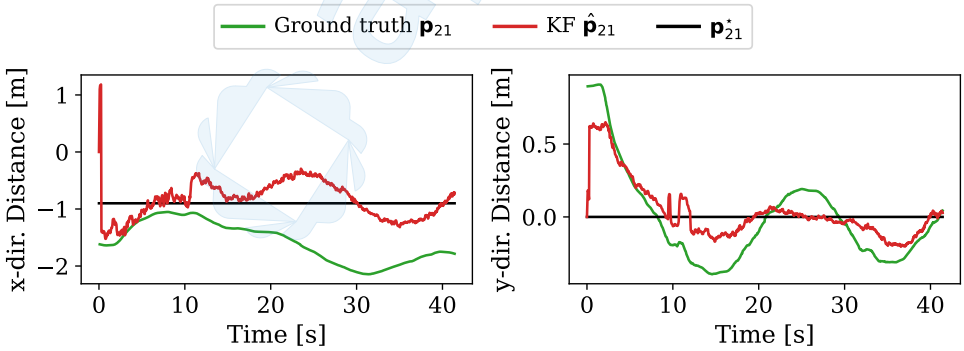
Figure 6.12: Timestamped perception and relative localization of agent 2 subject to adversarial mislocalization. The results are associated with the experiment listed in the second row of Table 6.2. The boxes with labels on top are the detections from the custom-trained YOLOv7 model, while the green boxes with labels underneath are calculated by projecting the 3D relative position estimations from the Kalman filter into the image space to determine the box's center, and by using the object's known size to compute the box's width and height in the image. Additionally, the image frames have been cropped for better visualization. The original camera image size was 640×480 pixels.



(a) Position Trajectories

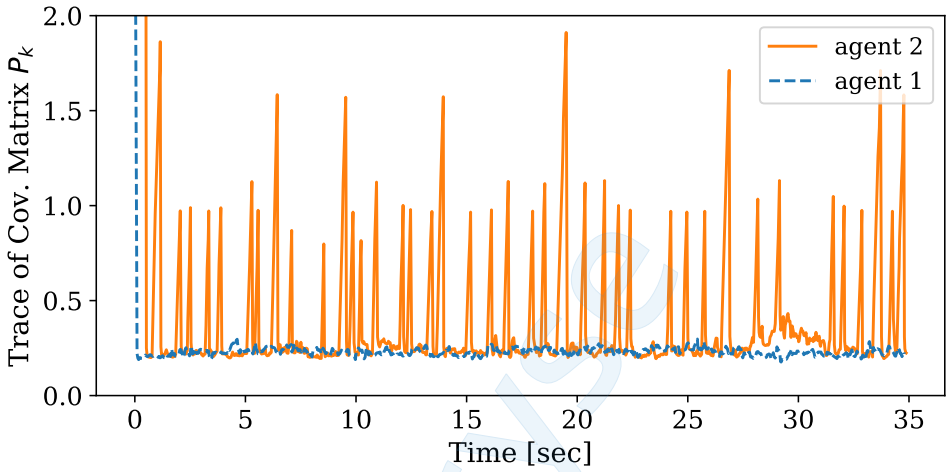


(b) Velocity Trajectories

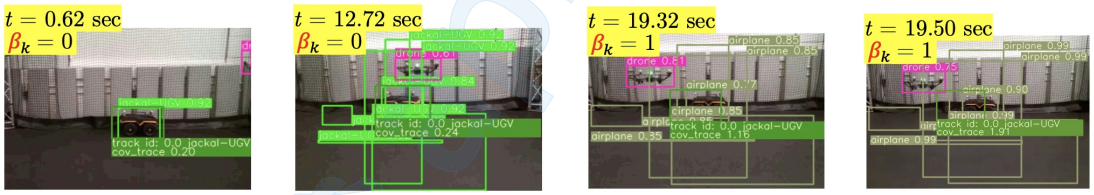


(c) Time-synchronized Trajectories of Relative Position Coordination

Figure 6.13: Results from a two-agent perception-based coordination experiment with adversarial mislocalization in the perception module, using the framework illustrated in Fig. 6.2. The adversarial mislocalization involves augmenting the nominal output of the object detection model with $b = 10$ spurious bounding boxes. The spurious boxes were generated by adversarially perturbing the nominal detected bounding box around the object of interest (jackal-UGV) by $q = \pm 30\%$ and increasing their probability confidence by 10%. Performance metrics and comparisons for this experiment are detailed in the second row of Table 6.2.

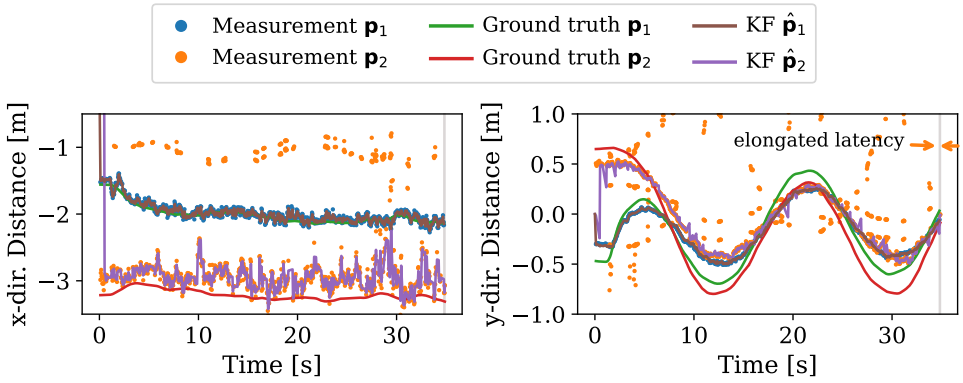


(a) The trace of state estimation covariance matrix

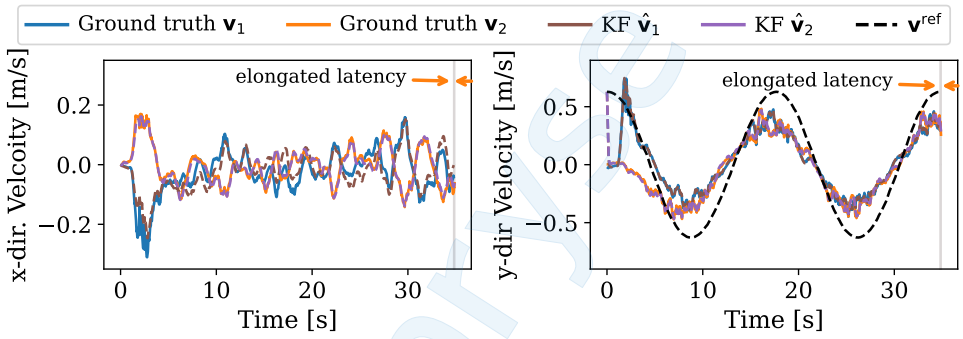


(b) Timestamped perception and relative localization of agent 2

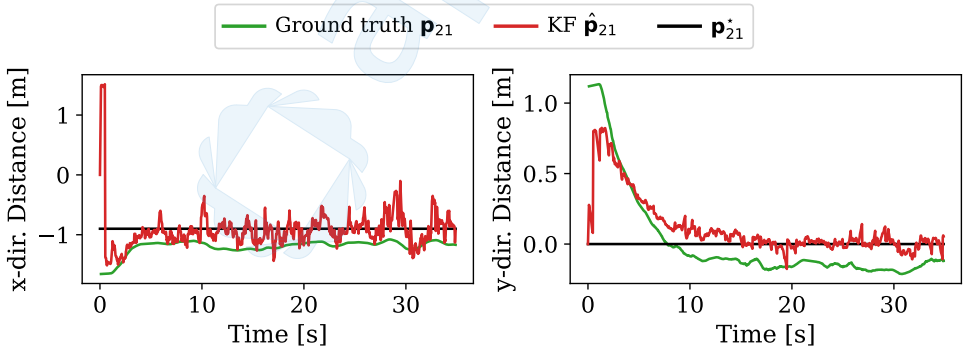
Figure 6.14: Results from a two-agent perception-based coordination experiment using the framework shown in Fig. 6.2, subject to both adversarial misclassification and mislocalization as detailed in Table 6.3. The peaks in (a) reflect the degenerative effect of adversarial misclassification inducing missed measurements in the Kalman filter (6.2.8). (b) The boxes with labels on top are the detections from the custom-trained YOLOv7 model, while the green boxes with labels underneath are calculated by projecting the 3D relative position estimations from the Kalman filter into the image space to determine the box's center, and by using the object's known size to compute the box's width and height in the image. Additionally, the image frames in (b) have been cropped for better visualization. The original camera image size was 640×480 pixels.



(a) Position Trajectories



(b) Velocity Trajectories



(c) Time-synchronized Trajectories of Relative Position Coordination

Figure 6.15: Results from a two-agent perception-based coordination experiment with adversarial misclassification and mislocalization in the perception module, using the framework illustrated in Fig. 6.2. The adversarial misclassification rate is modeled by a binomial distribution $\beta_k \sim \text{Bin}(n = 200, p = 0.2)$ in (6.2.8). The adversarial mislocalization involves augmenting the nominal output of the object detection model with $b = 5$ spurious bounding boxes. The spurious boxes were generated by adversarially perturbing the nominal detected bounding box around the object of interest (jackal-UGV) by $q = \pm 30\%$ and increasing their probability confidence by 10%. Performance metrics for this experiment are detailed in Table 6.3.

Chapter 7
Conclusion and Future Work

In this dissertation, we considered the resilience of multi-robot systems with wireless communication and learned perception modules in the context of coordination and formation control. We first considered resilience to worst-case scenario adversaries that exploit the vulnerabilities of wireless communication. Our results extend the previous results to the case of switching communication networks with intermittent connections that require to maintain connectivity only in an integral sense rather than constantly throughout time. This relaxation allowed us to design principled algorithms to detect and mitigate a class of adversarial attacks. We also characterized the system resilience to a worst-case set of malicious agents (robots) in a network as well as the network resilience to permanent and intermittent disconnections. In the second part of this dissertation, we considered a class of adversarial image attacks targeting the robots' learned perception models in the form of adversarial misclassification and mislocalization. We demonstrated that the resilience of multi-robot coordination under adversarial perception can be formulated and enhanced as resilience against sporadic (intermittent) and spurious measurements in a state estimation problem.

7.1 Summary

In Chapter 3, we considered the security goals of data confidentiality and integrity for a class of multi-agent (robot) control systems seeking average consensus. We proposed a decentralized attack detection framework designed to detect stealthy attacks that target the data integrity and stability of the multi-agent control systems. The framework includes two sets of observers: local and central (global) observers. It leverages

topology switching at the global level and local monitoring to detect the attacks. The approach is scalable as it relies on decentralized local observers for detection. Additionally, by enforcing partial observability, the framework preserves the privacy (confidentiality) of the initial conditions of the multi-agent system's states. In our analysis, we derived theoretical conditions for the detectability of stealthy attacks.

Chapter 4, extends the results of Chapter 3 to the case of formation control for a network of small Unmanned Aerial Vehicles (UAVs). Based on the theoretical finding of Chapter 3, we proposed a distributed attack detection framework for a relatively small group (e.g., $N \leq 10$) of UAVs that are subject to stealthy attacks. We have also developed an open-source software package for conducting indoor flight experiments using a class of small UAVs. We illustrated the performance of our proposed framework through various experiments. We demonstrated the performance of our proposed framework through various experiments. We believe this open-source software package will be beneficial for the robotics and control community.

In Chapter 5, we considered the security goals of data integrity and availability for a class of multi-agent (robot) systems. We considered the consensus and formation of multi-agent (robot) systems over a time-varying communication network subject to deception and DoS attacks. deception attacks target the data integrity in wireless communication and DoS attacks target the data availability. We showed, for a given integer number F, the communication network requires to be at least $(F+1)$-vertex-connected (resp. $(F+1)$-robust) in an integral sense and uniformly in time over a period of time T for resilience to an F-total (resp. F-local) adversary set that upper bounds the number of malicious agent with deception attacks. These bounds provide a relaxed compared to the existing ones in the literature. We presented theoretical guarantees and explicit bounds for exponentially fast convergence to the consensus/formation equilibrium in the presence of constrained DoS attacks. We

also presented a distributed attack detection framework with theoretical guarantees, which allows for resilient cooperation.

In Chapter 6, we developed two multi-robot platforms with perception and wireless communication capabilities. The platforms allow for experimental studies on adversarial image attacks on the perception module of multi-robot systems. We demonstrated that a class of adversarial image attacks on the robots' perception modules cause categorically similar effects including misclassification and mislocalization, which can be formulated as sporadic (intermittent) and spurious measurement data. We then proposed a framework that allows for state estimation and perception-based relative localization in the presence of intermittent and spurious measurements caused by adversarial image attacks on the perception module.

7.2 Future Directions

The development of *resilient* cyber-physical systems (CPS), particularly multi-robot systems, continues to be an active area of research with many open problems [102] and security goals outlined in [18].

In our proposed frameworks in Chapters 3-5, we implicitly assumed that agents (robots) have pre-designed collision-free set-points. In the face of attack detection and reconfiguration, collision-free trajectory planning as a contingency plan could be considered as a future research direction for resilient multi-robot coordination.

We proposed the (μ, T)-PE connectivity in (5.2.1) as relaxation to point-wise in-time connectivity (e.g, static network), whose connectivity uniformly in time allows for exponential stability and convergence in the presence of time-constrained DoS attacks (see Proposition 5.2.6). The uniformly in-time persistent excitation (PE) of connectivity requirement in (5.2.1) can be further relaxed [7, 8] to allow for asymptotic

stability and convergence in the presence of arbitrarily persistent DoS attacks.

In terms of privacy and data confidentiality, one promising research direction involves designing communication network topologies based on geometric symmetry and automorphisms [126, 125]. This principled approach enables the creation of unobservable subspaces that limit the ability to reconstruct the entire network from the locally available data of each (compromised) agent, thereby ensuring confidentiality.

Finally, considerable effort could be devoted to resilience against adversarial perception models in multi-robot settings. In particular, the derivation of theoretical bounds for resilience to adversarial sporadic and spurious measurement data in a state estimation problem for systems with degenerative dynamics (i.e. repeated eigenvalues) is an area of interest.

Appendix A
Proofs of Chapter 3

A.1 Auxiliary Results

The following Definition and Lemma are used in the Proof of Theorem 3.2.4.

Definition A.1.1. The Laplcaian matrix of the graph composed of switching links between two communication graphs is block diagonalizable, where each block, also called a component, encodes either a single (added/removed) switching link or a group of them are connected.

The foregoing definition can be formally presented as follows: consider a network topology that switches between two distinct topological configurations with respective Laplacian matrices $\mathsf{L}_{\sigma(t)=\mathbf{q}'}$ and $\mathsf{L}_{\sigma(t)=\mathbf{q}}$, where $\mathbf{q}', \mathbf{q} \in \mathcal{Q}$ with $\mathbf{q}' \neq \mathbf{q}$, and let $\Delta\mathsf{L}_{\mathbf{q}} = \mathsf{L}_{\mathbf{q}} - \mathsf{L}_{\mathbf{q}'}$ denote the difference between these two Laplacian matrices. Then, from Definition 2.2.2, $\Delta\mathsf{L}_{\mathbf{q}}$ is associated with the induced graph $\Delta\mathcal{G}_{\mathbf{q}} = (\mathcal{V}_{\mathbf{q}}, \Delta\mathcal{E}_{\mathbf{q}}, \Delta\mathsf{A}_{\mathbf{q}})$, that specifies connected graph component(s) corresponding to added/removed communication link(s) in the communication network such that

$$\mathcal{V}_{\mathbf{q}} = (\cup_{\mathrm{c}=1}^{\mathbf{c}} \mathcal{D}_{\mathrm{c}}) \cup \mathcal{D}_{\mathrm{s}}, \qquad \text{s.t} \qquad \mathcal{V}_{\mathbf{q}} = \mathcal{V}, \tag{A.1.1}$$

$$(i,j) \in \Delta\mathcal{E}_{\mathbf{q}} \ \text{ if } \ [\Delta\mathsf{A}_{\mathbf{q}}]_{i,j} = a_{ij}^{\mathbf{q}} - a_{ij}^{1} \neq 0 \iff [\Delta\mathsf{L}_{\mathbf{q}}]_{i,j} \neq 0, \tag{A.1.2}$$

where $\mathcal{D}_{\mathrm{c}}$ denotes the set of nodes (agents involved in switching links) in c-th connected component with $|\mathcal{D}_{\mathrm{c}}| \geq 2$ and $\mathcal{D}_{\mathrm{i}'} \cap \mathcal{D}_{\mathrm{j}'} = \emptyset$ for any $i', j' \in \{1, \cdots, \mathbf{c}\}$, $i' \neq j'$. Also, $\mathcal{D}_{\mathrm{s}}$ denotes the set of singletons i.e. single nodes that are not involved in any switching link. Then, there exists a permutation matrix P, $\mathrm{P}\mathrm{P}^{\top} = I$ to relabel the

nodes and represent the Laplacian matrix $\Delta\mathsf{L}_{\mathbf{q}}$ in block diagonal form, (cf. [87, Ch. 6.12]), as follows

$$\mathrm{P}\Delta\mathsf{L}_{\mathbf{q}}\mathrm{P}^{\top} = \breve{\mathsf{L}}_{\mathbf{q}} = \mathrm{diag}\,\{\Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_1), \cdots, \Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathbf{c}}), \Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{s}})\}, \tag{A.1.3}$$

where $\Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{c}})$ denotes the Laplacian matrix of the c-th connected component and $\Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{s}}) = 0$.

Lemma A.1.1. *Consider system in* (3.1.4) *with topology switching from normal mode* $\sigma(t) = 1$ *to a safe mode* $\sigma(t) = \mathbf{q} \in \mathcal{Q}$ *and the measurements set* $\mathcal{M}$ *in* (3.1.5)*, and let* $\Delta\mathsf{L}_{\mathbf{q}} = \mathsf{L}_{\mathbf{q}} - \mathsf{L}_1$ *denote the difference of the Laplacian matrices in safe and normal mode. Then under condition*

$$\mathrm{Im}(\Delta\mathsf{L}_{\mathbf{q}}) \cap \ker\left([\mathbf{C}_{\mathrm{x}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}\right) = \emptyset, \tag{A.1.4}$$

every connected graph component has at least one globally monitored node (agent), that is

$$\mathcal{D}_{\mathrm{c}} \cap \mathcal{M} \neq \emptyset, \quad \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}. \tag{A.1.5}$$

where $\mathbf{C}_{\mathrm{x}}$ *and* $\mathbf{C}_{\mathrm{v}}$ *are diagonal elements of* $\mathbf{C}$ *in* (3.1.5) *and* $\mathcal{D}_{\mathrm{c}}$ *denotes the set of nodes in* c*-th connected component of* $\Delta\mathsf{L}_{\mathbf{q}}$ *as given in* (A.1.1)*.*

Proof. We first show (A.1.4) is invariant under permutation of $\Delta\mathsf{L}_{\mathbf{q}}$ which is introduced in (A.1.3) and accordingly permutation of $[\mathbf{C}_{\mathrm{x}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}$. To this end, from

the definition of nullspace, we have

$$\ker\left(\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\Delta\mathsf{L}_{\mathbf{q}}\right)=\left\{x\in\mathbb{R}^{N}\mid\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\Delta\mathsf{L}_{\mathbf{q}}x=0\right\},\tag{A.1.6}$$

from which we obtain either

$$\Delta\mathsf{L}_{\mathbf{q}}x\notin\mathrm{Im}\left(\Delta\mathsf{L}_{\mathbf{q}}\right)\iff\Delta\mathsf{L}_{\mathbf{q}}x=\mathbf{0},\tag{A.1.7}$$

or

$$\mathbf{0}\neq\boldsymbol{\mathcal{Y}}=\Delta\mathsf{L}_{\mathbf{q}}x\in\mathrm{Im}\left(\Delta\mathsf{L}_{\mathbf{q}}\right)\implies\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\boldsymbol{\mathcal{Y}}=\mathbf{0},\tag{A.1.8}$$

where the latter, (A.1.8), is in contradiction with condition (A.1.4). Now under the permutation defined in (A.1.3), $\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\Delta\mathsf{L}_{\mathbf{q}}x=0$ in (A.1.6) can be rewritten in block-partitioned diagonal form as

$$\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\mathrm{P}^{\top}\breve{\mathsf{L}}_{\mathbf{q}}\mathrm{P}x=\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\mathrm{P}^{\top}\breve{\mathsf{L}}_{\mathbf{q}}\chi=\begin{bmatrix}\breve{\mathbf{C}}_{\mathrm{x}}\\ \breve{\mathbf{C}}_{\mathrm{v}}\end{bmatrix}\breve{\mathsf{L}}_{\mathbf{q}}\chi=\mathbf{0},\tag{A.1.9}$$

in which $\chi=\mathrm{P}x$ denotes the relabeled x such that

$$\begin{aligned}&\chi=\mathrm{col}\left(\chi_{1},\ldots,\chi_{\mathbf{c}}\right)=\mathrm{P}x,\ \text{ with}\\ &\chi_{\mathrm{c}}=\mathrm{col}\left(x_{i}\right),\quad\forall\, i\in\mathcal{D}_{\mathrm{c}},\quad\forall\,\mathrm{c}\in\{1,\cdots,\mathbf{c}\}.\end{aligned}\tag{A.1.10}$$

Also, $\breve{\mathbf{C}}_{\mathrm{k}}=\mathbf{C}_{\mathrm{k}}\mathrm{P}^{\top}=\begin{bmatrix}\mathbf{C}_{\mathrm{k}}^{1}&\cdots&\mathbf{C}_{\mathrm{k}}^{\mathbf{c}}\end{bmatrix}$, $\mathrm{k}\in\{\mathrm{x},v\}$ is a block-partitioned binary matrix

that specifies monitored agents of each component [1]. To show the results in (A.1.7) and (A.1.8) hold also for the transformed form in (A.1.9), one need to verify the invariance of (A.1.4) under the permutation by P, that is

$$\mathrm{Im}(\Delta \mathsf{L}_{\mathbf{q}}) \cap \ker\left([\mathbf{C}_{\mathrm{x}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}\right) = \emptyset \iff \mathrm{Im}(\breve{\mathsf{L}}_{\mathbf{q}}) \cap \ker\left([\breve{\mathbf{C}}_{\mathrm{x}}^{\top}\ \breve{\mathbf{C}}_{\mathrm{v}}^{\top}]^{\top}\right) = \emptyset. \tag{A.1.11}$$

To show this, from the range and nullspace definition, for subspaces in (A.1.11) we have

$$\mathrm{Im}(\Delta \mathsf{L}_{\mathbf{q}}) = \left\{\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{N} \mid \boldsymbol{\mathcal{Y}} = \Delta \mathsf{L}_{\mathbf{q}} x(t)\right\}, \tag{A.1.12}$$

$$\ker\left(\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\right) = \ker(\mathbf{C}_{\mathrm{x}}) \cap \ker(\mathbf{C}_{\mathrm{v}}) = \left\{\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N} \mid \mathbf{C}_{\mathrm{x}}\boldsymbol{\mathcal{X}} = 0,\ \mathbf{C}_{\mathrm{v}}\boldsymbol{\mathcal{X}} = 0\right\}, \tag{A.1.13}$$

and

$$\begin{aligned}\mathrm{Im}(\breve{\mathsf{L}}_{\mathbf{q}}) &= \left\{\breve{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^{N} \mid \breve{\boldsymbol{\mathcal{Y}}} = \breve{\mathsf{L}}_{\mathbf{q}}\chi(t) = \breve{\mathsf{L}}_{\mathbf{q}}\mathrm{P}x(t)\right\}\\ &= \left\{\breve{\boldsymbol{\mathcal{Y}}} \in \mathbb{R}^{N} \mid \mathrm{P}^{\top}\breve{\boldsymbol{\mathcal{Y}}} = \mathrm{P}^{\top}\breve{\mathsf{L}}_{\mathbf{q}}\mathrm{P}x(t) = \boldsymbol{\mathcal{Y}}\right\} = \mathrm{P}\ \mathrm{Im}\left(\Delta \mathsf{L}_{\mathbf{q}}\right),\end{aligned} \tag{A.1.14}$$

[1] Note that $\mathrm{P}^{\top}$ permutes the columns of binary matrix $\mathbf{C}_{\mathrm{k}}$ whose row-vector elements are $\mathfrak{e}_i^{\top}$, $\forall\, i \in \mathcal{M}_{\mathrm{k}}$, $\mathrm{k} \in \{\mathrm{x}, v\}$.

where we used (A.1.3) and $\chi(t) = \mathrm{P}x(t)$ as in (A.1.10) and (A.1.12). Similarly,

$$\begin{aligned}
\ker\left(\begin{bmatrix}\breve{\mathbf{C}}_{\mathrm{x}}\\ \breve{\mathbf{C}}_{\mathrm{v}}\end{bmatrix}\right) &= \ker(\breve{\mathbf{C}}_{\mathrm{x}}) \cap \ker(\breve{\mathbf{C}}_{\mathrm{v}}) \\
&= \left\{\breve{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^N \mid \breve{\mathbf{C}}_{\mathrm{x}}\breve{\boldsymbol{\mathcal{X}}} = 0,\, \breve{\mathbf{C}}_{\mathrm{v}}\breve{\boldsymbol{\mathcal{X}}} = 0\right\} \\
&= \left\{\breve{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^N \mid \mathbf{C}_{\mathrm{x}}\mathrm{P}^\top\breve{\boldsymbol{\mathcal{X}}} = 0,\, \mathbf{C}_{\mathrm{v}}\mathrm{P}^\top\breve{\boldsymbol{\mathcal{X}}} = 0\right\} \\
&= \left\{\breve{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^N \mid \mathbf{C}_{\mathrm{x}}\boldsymbol{\mathcal{X}} = 0,\, \mathbf{C}_{\mathrm{v}}\boldsymbol{\mathcal{X}} = 0,\, \mathrm{P}\boldsymbol{\mathcal{X}} = \breve{\boldsymbol{\mathcal{X}}}\right\} \\
&= \mathrm{P}\ \ker\left(\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\right).
\end{aligned} \tag{A.1.15}$$

Then

$$\begin{aligned}
\mathrm{Im}(\breve{\mathsf{L}}_{\mathbf{q}}) \cap \ker\left(\begin{bmatrix}\breve{\mathbf{C}}_{\mathrm{x}}\\ \breve{\mathbf{C}}_{\mathrm{v}}\end{bmatrix}\right) &= \mathrm{P}\ \mathrm{Im}\left(\Delta\mathsf{L}_{\mathbf{q}}\right) \cap \mathrm{P}\ \ker\left(\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\right) \\
&= \mathrm{P}\left(\mathrm{Im}\left(\Delta\mathsf{L}_{\mathbf{q}}\right) \cap \ker\left(\begin{bmatrix}\mathbf{C}_{\mathrm{x}}\\ \mathbf{C}_{\mathrm{v}}\end{bmatrix}\right)\right) = \mathrm{P}\left(\emptyset\right) = \emptyset.
\end{aligned} \tag{A.1.16}$$

where we used fact 2.9.29 in [12] and condition 1.

Now one can prove (A.1.5) by contradiction. Assume (A.1.5) does not hold, that is $\exists\, \mathrm{c}' \in \{1, \cdots, \mathbf{c}\}$, s.t. $\mathcal{D}_{\mathrm{c}'} \cap \mathcal{M} = \emptyset$, under which we have the c′-th block in (A.1.9) such that

$$\begin{bmatrix}\breve{\mathbf{C}}_{\mathrm{x}}^{\mathrm{c}'}\\ \breve{\mathbf{C}}_{\mathrm{v}}^{\mathrm{c}'}\end{bmatrix} \Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{c}'})\chi_{\mathrm{c}'}(t) = \mathbf{0}, \quad \breve{\mathbf{C}}_{\mathrm{x}}^{\mathrm{c}'} = \breve{\mathbf{C}}_{\mathrm{x}}^{\mathrm{c}'} = 0, \tag{A.1.17}$$

which holds for all $\chi_{\mathrm{c}'}(t)$ with $\Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{c}'})\chi_{\mathrm{c}'}(t) \in \mathrm{Im}\left(\Delta\mathsf{L}_{\mathbf{q}}(\mathcal{D}_{\mathrm{c}'})\right) \subseteq \mathrm{Im}\left(\breve{\mathsf{L}}_{\mathbf{q}}\right)$ as in

(A.1.17) $\mathrm{Im}\left(\Delta \mathsf{L}_{\mathbf{q}}(\mathcal{D}_{c'})\right) \in \ker\left(\begin{bmatrix} \breve{\mathbf{C}}_{\mathrm{x}}^{c'} \\ \breve{\mathbf{C}}_{\mathrm{v}}^{c'} \end{bmatrix}\right) \implies \mathrm{Im}\left(\breve{\mathsf{L}}_{\mathbf{q}}\right) \cap \ker\left([\breve{\mathbf{C}}_{\mathrm{x}}^{\top}\ \breve{\mathbf{C}}_{\mathrm{v}}^{\top}]^{\top}\right) \neq \emptyset$ that contradicts (A.1.11).

A.2 Proof of Lemma 3.2.1

Note that the Laplacian matrix $\mathsf{L}_{\sigma(t)}$ of every connected undirected (or strongly connected and balanced directed) graph has only one zero eigenvalue, $\lambda = 0$, with the corresponding eigenvector $\mathbf{1}_N$ such that $\mathsf{L}_{\sigma(t)}\mathbf{1}_N = \mathbf{0}$ [93]. Then, given the structure of $\mathbf{A}_{\sigma(t)}$ in (3.1.4), $(\lambda = 0, w_r = \left[\begin{smallmatrix} (1/N)\mathbf{1}_N \\ \mathbf{0}_N \end{smallmatrix}\right])$ is an eigenpair of system matrix $\mathbf{A}_{\sigma(t)}$ associated with that of Laplacian $\mathsf{L}_{\sigma(t)}$ with $\sigma(t_{k-1}) = \mathbf{q} \in \mathcal{Q}$, $t \in [t_{k-1}, t_k)$. Also, it can be verified that the eigenpair $(\lambda = 0, w_r)$ lies in the unobservable subspace of system (3.1.4) as it is a nontrivial solution to the PBH test for observability:

$$\begin{bmatrix} \lambda I - \mathbf{A}_{\mathbf{q}} \\ \mathbf{C} \end{bmatrix} w_r = \mathbf{0}, \quad \lambda = 0 \in \mathbb{C}, \tag{A.2.1}$$

$$\mathbf{C} = \mathrm{diag}\left\{0, C_{\mathrm{v}}\right\}. \tag{A.2.2}$$

Therefore, one can conclude that the right eigenvector w_r contained in $\ker(\mathbf{C})$ belongs to $\ker(\mathcal{O}_k)$ that is defined in (2.3.5) [51, Th. 15.8]. Furthermore, as $(\lambda = 0, w_r)$ is the eigenpair associated with the equilibrium subspace (3.1.2) of every $\mathbf{A}_{\mathbf{q}}$ with Laplacian $\mathsf{L}_{\mathbf{q}}$, it is straightforward from Lemma 2.3.1 that $\mathrm{span}\left\{w_r\right\} = \mathrm{span}\left\{\left[\begin{smallmatrix} \mathbf{1}_N \\ \mathbf{0}_N \end{smallmatrix}\right]\right\} \subseteq \boldsymbol{\mathcal{N}}_1^{\infty} = \ker(\boldsymbol{\mathcal{O}})$ over $t \in [t_0, +\infty)$.

A.3 Proof of Proposition 3.2.3

Let $\sigma(t) = \mathbf{q} \in \mathcal{Q},\ t \in [t_{k-1}, t_k)$ and consider the error dynamics of local observers in (3.2.9). According to Definition 2.3.1, a ZDA for (3.2.9) should satisfy

$$\begin{bmatrix} \lambda_0 I - \mathbf{F}^{\mathrm{i}}_{\mathbf{q}} & -\mathbf{T}^{\mathrm{i}}\mathbf{B}^{\mathrm{i}} \\ \mathbf{C}_{\mathrm{i}_i} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_{\mathrm{i}}(0) \\ \mathbf{u}_{0_{\mathrm{i}}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{A.3.1}$$

where $\tilde{\mathbf{e}}_{\mathrm{i}}(0) := \mathbf{e}_{\mathrm{i}}(0) - \bar{\mathbf{e}}_{\mathrm{i}}(0) = \tilde{\mathbf{x}}_{0_{\mathrm{i}}}$. Also, by considering (3.2.7) and the fact that $\mathbf{C}_{\mathrm{i}_i}\tilde{\mathbf{e}}_{\mathrm{i}}(0) = \mathbf{C}_{\mathrm{i}_i}\tilde{\mathbf{x}}_{\mathrm{i}}(0) = \mathbf{0}$ in the second equation of (A.3.1), matrix pencil (A.3.1) can be rewritten as

$$\overbrace{\begin{bmatrix} \lambda_0 I - \bar{\mathbf{A}}^{\mathrm{i}}_{\mathbf{q}} & -\mathbf{T}^{\mathrm{i}}\mathbf{B}^{\mathrm{i}} \\ \mathbf{C}_{\mathrm{i}_i} & 0 \end{bmatrix}}^{\bar{\mathbf{P}}} \begin{bmatrix} \tilde{\mathbf{x}}_{\mathrm{i}}(0) \\ \mathbf{u}_{0_{\mathrm{i}}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}. \tag{A.3.2}$$

It follows immediately from Definition 2.3.1 that a stealthy attack $\mathbf{u}_{a_{\mathrm{i}}}$ in (3.2.9), whether it is a ZDA or covert attack[2], loses its stealthiness with respect to the local residual $\mathbf{r}_{\mathrm{i}_i}$ if, and only if, there is no non-trivial zeroing direction associated with matrix pencil in (A.3.1) or equivalently $\bar{\mathbf{P}}$ in (A.3.2), which in turn implies $\bar{\mathbf{P}}$ has full rank. Moreover, from Definition 2.3.1 and condition (3.2.4), it is straightforward that matrix pencil $\mathbf{P}$, defined in (3.2.10), is associated with the zeroing direction of the local system (3.2.1). We now show how conditions (1)-(3) establish the equivalence between the rank sufficiency of $\mathbf{P}$ in (3.2.10) and $\bar{\mathbf{P}}$ in (A.3.2). Given $\mathbf{P}$ in (3.2.10),

[2]Note that a covert attack is defined in (3.1.4) based on the network-level measurements (3.1.4b)

one can write

$$\begin{bmatrix} I - \mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_i} & \lambda_0\mathbf{h}^{\mathrm{i}} \\ 0 & I \\ \mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_i} & -\lambda_0\mathbf{h}^{\mathrm{i}} \end{bmatrix} \mathbf{P} = \begin{bmatrix} \lambda_0 I - \bar{\mathbf{A}}^{\mathrm{i}}_{\mathbf{q}} & -(I - \mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_i})\mathbf{B}^{\mathrm{i}} & 0 \\ \mathbf{C}_{\mathrm{i}_i} & 0 & 0 \\ -\mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_i}\mathbf{A}^{\mathrm{i}}_{\mathbf{q}} & \mathbf{h}^{\mathrm{i}}\mathbf{C}_{\mathrm{i}_i}\mathbf{B}^{\mathrm{i}} & \mathbf{E}^{\mathrm{i}} \end{bmatrix}, \tag{A.3.3}$$

where $\mathbf{h}^{\mathrm{i}} := \mathbf{E}^{\mathrm{i}}(\mathbf{C}_{\mathrm{i}_i}\mathbf{E}^{\mathrm{i}})^{\dagger}$ is a solution to (3.2.6) that exists under condition (ii) [24, Lemma 1]. Then, postmultiplying (A.3.3) by

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ (\mathbf{C}_{\mathrm{i}_i}\mathbf{E}^{\mathrm{i}})^{\dagger}\mathbf{C}_{\mathrm{i}_i}\mathbf{A}^{\mathrm{i}}_{\mathbf{q}} & (\mathbf{C}_{\mathrm{i}_i}\mathbf{E}^{\mathrm{i}})^{\dagger}\mathbf{C}_{\mathrm{i}_i}\mathbf{B}^{\mathrm{i}} & I \end{bmatrix}, \tag{A.3.4}$$

and considering (3.2.7) yields

$$\begin{bmatrix} \lambda_0 I - \bar{\mathbf{A}}^{\mathrm{i}}_{\mathbf{q}} & -\mathbf{T}^{\mathrm{i}}\mathbf{B}^{\mathrm{i}} & 0 \\ \mathbf{C}_{\mathrm{i}_i} & 0 & 0 \\ 0 & 0 & \mathbf{E}^{\mathrm{i}} \end{bmatrix}. \tag{A.3.5}$$

Since node $i \in \mathcal{P}_{\mathrm{i}}$ is $\mathbf{k}$-connected, we have $|\mathcal{N}_i| = \mathbf{k}$ and $\mathbf{k} \leq \operatorname{rank}(\mathbf{C}_{\mathrm{i}_i}) \leq 2\mathbf{k}$ (cf. (3.1.5)). Then, from condition (i), one can verify that $\operatorname{rank}(\mathbf{C}_{\mathrm{i}_i}) \geq \operatorname{rank}(\mathbf{B}^{\mathrm{i}}) + \operatorname{rank}(\mathbf{E}^{\mathrm{i}})$ guarantees (3.2.10) is a tall or square matrix pencil having only a finite number[3] of output-zeroing directions [69, Ch. 2]. Also, the pre- and post-multiplied

[3]This condition is not valid for degenerate systems which are out of scope of this work.

matrices in (A.3.3) and (A.3.4) are full column ranks. Therefore, we have

$$\operatorname{rank}(\mathbf{P}) = \operatorname{rank} \underbrace{\begin{bmatrix} \lambda_0 I - \bar{\mathbf{A}}_{\mathbf{q}}^{\mathrm{i}} & -\mathbf{T}^{\mathrm{i}}\mathbf{B}^{\mathrm{i}} \\ \mathbf{C}_{\mathrm{i}_i} & 0 \end{bmatrix}}_{\bar{\mathbf{P}}} + \operatorname{rank}\left(\mathbf{E}^{\mathrm{i}}\right). \tag{A.3.6}$$

Recall $\mathbf{E}^{\mathrm{i}}$ is full column rank, and hence $\mathbf{P}$ in (3.2.10) is full rank if, and only if, $\bar{\mathbf{P}}$ in (A.3.6) is full rank. This guarantees that a locally undetectable stealthy attack is impossible.

A.4 Proof of Theorem 3.2.4

Consider (3.2.14) over $t \in [t_0, +\infty)$, and let the safe mode $\sigma(t) = \mathbf{q} \in \mathcal{Q}$, $t \in [t_1, +\infty)$ the continuous system residual $\mathbf{r}_0(t)$ and its successive derivatives can be rewritten as

$$\mathbf{R} = \mathcal{O}_1 \mathbf{e}(t) - \boldsymbol{\mathcal{H}}(\mathbf{HC})\mathbf{E} + \boldsymbol{\mathcal{H}}(\mathbf{B})\mathbf{U}_a + \boldsymbol{\mathcal{H}}(\mathbf{H})\mathbf{U}_s - \mathbf{U}_s + \boldsymbol{\mathcal{H}}(\Delta\mathbf{A}_{\mathbf{q}})\mathbf{X}, \tag{A.4.1}$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_0^\top(t) & \dot{\mathbf{r}}_0^\top(t) & \cdots & (\mathbf{r}_0^\top(t))^{(\mathrm{d})} \end{bmatrix}^\top, \tag{A.4.2}$$

$$\mathbf{U}_\jmath = \begin{bmatrix} \mathbf{u}_\jmath^\top(t) & \dot{\mathbf{u}}_\jmath^\top(t) & \cdots & (\mathbf{u}_\jmath^\top(t))^{(\mathrm{d})} \end{bmatrix}^\top, \tag{A.4.3}$$

$$\mathbf{E} = \begin{bmatrix} \mathbf{e}^\top(t) & \dot{\mathbf{e}}^\top(t) & \cdots & (\mathbf{e}^\top(t))^{(\mathrm{d})} \end{bmatrix}^\top, \tag{A.4.4}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^\top(t) & \dot{\mathbf{x}}^\top(t) & \cdots & (\mathbf{x}^\top(t))^{(\mathrm{d})} \end{bmatrix}^\top, \tag{A.4.5}$$

$$\boldsymbol{\mathcal{H}}(b) = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \mathbf{C}b & 0 & 0 & \cdots & 0 \\ \mathbf{C}\mathbf{A}_1 b & \mathbf{C}b & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{C}\mathbf{A}_1^{\mathrm{d}} b & \mathbf{C}\mathbf{A}_1^{\mathrm{d}-1} b & \cdots & \mathbf{C}b & 0 \end{bmatrix}, \tag{A.4.6}$$

with $\jmath \in \{a, s\}$, $b \in \{\mathbf{B}, \mathbf{HC}, \mathbf{H}, \Delta\mathbf{A}_\mathbf{q}\}$, $\Delta\mathbf{A}_\mathbf{q} = (\mathbf{A}_\mathbf{q} - \mathbf{A}_1)$ and $\mathrm{d} \in \mathbb{N} \setminus \{1, 2\}$.

From (3.1.7) in Proposition 3.1.4 and (3.2.12)-(3.2.13), it can be easily verified that (A.4.1) is simplified to $\mathbf{R} = \mathcal{O}_1 \bar{\mathbf{e}}(t) - \boldsymbol{\mathcal{H}}(\mathbf{HC})\bar{\mathbf{E}} + \boldsymbol{\mathcal{H}}(\Delta\mathbf{A}_\mathbf{q})\mathbf{X}$ where $\bar{\mathbf{E}}$ has the same form as (A.4.4) while whose elements are $\bar{\mathbf{e}}$ and its derivatives. Therefore, in a stealthy attack case $\lim_{t_1 \to \infty} \mathbf{R} = \mathbf{0}$ during normal mode over $t \in [t_0, t_1)$. The objective is to characterize the effect of switching communication, modeled as discrepancy $\Delta\mathbf{A}_\mathbf{q}$ in (3.2.14) and (A.4.1), on the stealthiness of attacks in the residual $\mathbf{r}_0(t)$ of centralized observer (3.2.11) during safe mode over $t \in [t_1, +\infty)$ (cf. Problem 3.1.6). Given the input-output matrix (A.4.6) for the switching perturbations $\Delta\mathbf{A}_\mathbf{q}$ in (A.4.1), note that $\boldsymbol{\mathcal{H}}(\Delta\mathbf{A}_\mathbf{q})\mathbf{X} = \mathbf{0}$ over $t \in [t_1, +\infty)$ in (A.4.1) is the necessary condition under which the stealthy attacks, modeled in (3.1.6), remain undetectable in the residual $\mathbf{r}_0(t)$ of (3.2.14), regardless of the perturbation $\Delta\mathbf{A}_\mathbf{q}\mathbf{x}$ caused by topol-

ogy switching. Therefore, $\boldsymbol{\mathcal{H}}(\Delta\mathbf{A_q})\mathbf{X} \neq \mathbf{0}$ in (A.4.1) implying the system switching $\Delta\mathbf{A_q}$ affects $\mathbf{R}(t)$, $t \in [t_1, +\infty)$ in (A.4.1) guarantees attack detectability in $\mathbf{r}_0(t)$.

Consider Markov parameters $\mathbf{C}\mathbf{A}_1^{\mathrm{d}}\Delta\mathbf{A_q}$, $\mathrm{d} \in \mathbb{N}_0$ in (A.4.6), the term $\boldsymbol{\mathcal{H}}(\Delta\mathbf{A_q})\mathbf{X}$ in (A.4.1) can be rewritten as

$$\sum_{l=0}^{\mathrm{d}} \mathbf{C}\mathbf{A}_1^{\mathrm{d}}\Delta\mathbf{A_q}\mathbf{x}^{(\mathrm{d}-l)}(t) = \mathbf{0}, \quad \forall\, \mathrm{d} \in \mathbb{N}_0. \tag{A.4.7}$$

We show that under condition 1, the first two terms in (A.4.7) are non-zero (and so is $\boldsymbol{\mathcal{H}}(\Delta\mathbf{A_q})\mathbf{X} \neq \mathbf{0}$) unless $\Delta\mathbf{A_q}\mathbf{x}(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty)$.

By setting $\mathrm{d} = 0, 1$, and expanding (A.4.7) we obtain

$$\mathrm{d} = 0 \overset{(A.4.7)}{\Rightarrow} \quad \mathbf{C}_{\mathrm{v}}\Delta\mathsf{L_q}x(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty), \tag{A.4.8}$$

$$\mathrm{d} = 1 \overset{(A.4.7)}{\Rightarrow} \quad \mathbf{C}_{\mathrm{v}}\Delta\mathsf{L_q}v(t) = \mathbf{0}, \quad \text{and,}$$

$$\mathbf{C}_{\mathrm{x}}\Delta\mathsf{L_q}x(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty), \tag{A.4.9}$$

where $\mathbf{C}_{\mathrm{x}}$ and $\mathbf{C}_{\mathrm{v}}$ are diagonal elements of $\mathbf{C}$ as given in (3.1.4)-(3.1.5), $\Delta\mathsf{L_q} = \mathsf{L_q} - \mathsf{L}_1$ is the non-zero submatrix of $\Delta\mathbf{A_q} = (\mathbf{A}_{\sigma(t)} - \mathbf{A}_1) = \left[\begin{smallmatrix} 0 & 0 \\ -\alpha\Delta\mathsf{L_q} & 0 \end{smallmatrix}\right]$, and $\mathbf{x}(t) = \mathrm{col}(x(t), v(t))$ as in (3.1.4). Then, using (A.4.8) and (A.4.9), we have

$$\begin{bmatrix} \mathbf{C}_{\mathrm{x}} \\ \mathbf{C}_{\mathrm{v}} \end{bmatrix} \Delta\mathsf{L_q}x(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty). \tag{A.4.10}$$

Under condition 1, one can verify that (A.4.10) implies

$$\Delta\mathsf{L_q}x(t) \notin \mathrm{Im}\,(\Delta\mathsf{L_q}) \iff \Delta\mathsf{L_q}x(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty). \tag{A.4.11}$$

otherwise, for any $x(t)$ such that $\mathbf{0} \neq \Delta\mathsf{L}_{\mathbf{q}} x(t) = \boldsymbol{\mathcal{Y}} \in \mathrm{Im}\,(\Delta\mathsf{L}_{\mathbf{q}})$, we obtain

$$[\mathbf{C}_{\mathrm{x}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}\boldsymbol{\mathcal{Y}} = 0, \qquad \boldsymbol{\mathcal{Y}} \in \ker\left([\mathbf{C}_{\mathrm{x}}^{\top}\ \mathbf{C}_{\mathrm{v}}^{\top}]^{\top}\right)$$

for (A.4.10), which is in contradiction with condition 1.

Now considering the consensus protocol (3.1.3), it can be verified that $\Delta\mathsf{L}_{\mathbf{q}}$ (or equivalently $\Delta\mathbf{A}_{\mathbf{q}}$ in (A.4.7)), encodes connected graph component(s) corresponding to added/removed communication link(s) in the communication network (cf. Definition 2.2.2 and A.1.1). Then, applying an elementary transformation using a permutation matrix P defined in (A.1.3), transforms (A.4.11) into block-diagonal form as follows:

$$\Delta\mathsf{L}_{\mathbf{q}} x(t) = \mathbf{0} \iff \breve{\mathsf{L}}_{\mathbf{q}} \chi(t) = \mathbf{0}, \quad \forall\, t \in [t_1, +\infty), \tag{A.4.12}$$

where the block-diagonal matrix $\breve{\mathsf{L}}_{\mathbf{q}}$ is given in (A.1.3) and $\chi(t) = \mathrm{P}x(t)$ denotes the relabeled system states such that

$$\begin{aligned} &\chi(t) = \mathrm{col}\,(\chi_1(t), \ldots, \chi_{\mathbf{c}}(t)) = \mathrm{P}x(t), \ \text{ with} \\ &\chi_{\mathrm{c}}(t) = \mathrm{col}\,(x_i(t)), \quad \forall\, i \in \mathcal{D}_{\mathrm{c}}, \quad \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}, \end{aligned} \tag{A.4.13}$$

with $\mathcal{D}_{\mathrm{c}}$ being the set of nodes (agents involved in switching links) in c-th connected component[4] as in (A.1.1). Also, note that the permutation matrix P is a binary nonsingular matrix by definition. Additionally, the Laplacian matrix is a zero row-sum matrix, and if connected, its nullspace is spanned by $\mathbf{1}$, a vector of all ones [93]. Therefore, from (A.4.12) and considering nodes involved in (connected) switching

[4]Although the analysis here is at the global level, it is worth mentioning that $\Delta\mathsf{L}_{\mathbf{q}}$ at cluster levels i.e. $\mathcal{P}_{\mathrm{i}}$, $\mathrm{i} \in \{1, \cdots, |\boldsymbol{\mathcal{P}}|\}$ may have more than one connected component.

links, i.e., for all $\forall\, i, j \in \mathcal{D}_{\mathrm{c}}$, $i \neq j$, one can conclude that

$$\begin{aligned} & x_i(t) - x_j(t) = 0 \Leftrightarrow \\ & \qquad x_i(t) = x_j(t), \qquad\qquad \forall\, i, j \in \mathcal{D}_{\mathrm{c}}, \;\; \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}, \\ & \qquad\qquad\qquad\qquad\quad \forall\, t \in [t_1, +\infty), \end{aligned} \tag{A.4.14}$$

which by considering the continuity of the system states can be extended for its higher-order time derivatives and be rewritten as follows:

$$\begin{aligned} (\mathfrak{e}_i^\top - \mathfrak{e}_j^\top) x^{(\mathrm{m})}(t) = 0, \;\; \forall\, i, j \in \mathcal{D}_{\mathrm{c}}, \;\; \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}, \\ \forall\, \mathrm{m} \in \mathbb{N}_0, \;\; \forall\, t \in [t_1, +\infty), \end{aligned} \tag{A.4.15}$$

with $\mathfrak{e}_i$, $\mathfrak{e}_j$ being i-th and j-th standard-basis vectors in $\mathbb{R}^N$.

Also, from (A.4.7), (A.4.12), (A.4.15) and by considering the structure $\mathbf{A}_{\sigma(t)}$ and system state $\mathbf{x}(t) = \mathrm{col}(x(t), v(t))$ in (3.1.4), we obtain

$$\begin{aligned} & \Delta \mathbf{A}_{\mathbf{q}} \mathbf{x}^{(\mathrm{m})}(t) = \mathbf{0} \Leftrightarrow \\ & \Delta \mathsf{L}_{\mathbf{q}} x^{(\mathrm{m})}(t) = \mathbf{0}, \;\; \forall\, \mathrm{m} \in \mathbb{N}_0, \forall\, t \in [t_1, +\infty). \end{aligned} \tag{A.4.16}$$

Therefore, under condition 1, one can conclude that unless (A.4.11)/(A.4.15) holds that is the system states (positions $x_i(t)$, $x_j(t)$ and their successive derivatives) of all agents within each graph component, i.e. agents involved in connected intra-cluster switching links, are respectively identical $\forall\, t \in [t_1, +\infty)$, the left side of (A.4.8) and (A.4.9) is non-zero and so is (A.4.7), implying $\Delta \mathbf{A}_{\mathbf{q}}$ affects $\mathbf{R}(t)$ whereby the attacks are detectable.

We now show under conditions 2-3 the domain of existence of (A.4.11) is shrunk

into the only case that the entire system states, except for those affected by stealthy attacks, are at an equilibrium.

Zero-dynamics attack (ZDA) case: it can be shown that under condition 1, (A.4.11) holds (and so does (A.4.15)) only in the worst-case scenario, in the sense of attack detection, that none of the agents involved in intra-cluster switching links are affected by the ZDA in a safe mode. To this end, consider (A.4.12) under which ZDA remains stealthy in residual $\mathbf{r}_0(t)$ in the safe modes and recall

$$\mathbf{x}(t) = \bar{\mathbf{x}}(t) + \tilde{\mathbf{x}}(t), \quad \tilde{\mathbf{x}}_0 e^{\lambda_0 t}, \quad \forall\, t \in [t_0, +\infty), \tag{A.4.17}$$

in a stealthy ZDA case with $\tilde{\mathbf{x}}_0 e^{\lambda_0 t_1} \in \ker(\mathbf{C})$ being the initial condition of ZDA (cf. (2.3.1), and (3.1.14) in Proposition 3.1.4) at $t = t_1$ for a safe mode. Similar to (3.2.16), by evaluating ZDA condition (2.3.1) for the tuple $(\mathbf{A_q}, \mathbf{B}, \mathbf{C})$ with $\mathbf{A_q} = (\mathbf{A}_1 + \Delta\mathbf{A_q})$ and considering (A.4.16) we obtain

$$\begin{bmatrix} \lambda_0 I - (\mathbf{A}_1 - \mathbf{HC}) & (\mathbf{A_q} - \mathbf{A}_1) & -\mathbf{B} \\ \mathbf{C} & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}(t_1) \\ \tilde{\mathbf{x}}(t_1) \\ \mathbf{u}_{\mathcal{A}}(t_1) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{A.4.18}$$

where as in (3.2.16), $\tilde{\mathbf{e}}(t_1) = \tilde{\mathbf{x}}(t_1)$ with $\tilde{\mathbf{x}}(t_1) = \tilde{\mathbf{x}}_0 e^{\lambda_0 t_1}$ and $\mathbf{u}_{\mathcal{A}}(t_1) = \mathbf{u}_0 e^{\lambda_0 t_1}$. Then (A.4.18) is simplified to

$$\begin{bmatrix} \lambda_0 I - (\mathbf{A_q} - \mathbf{HC}) & -\mathbf{B} \\ \mathbf{C} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}(0) \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{A.4.19}$$

where further simplification, similar to that in (3.2.16), and expanding it out yields

$$\begin{bmatrix} \lambda_0 I_N & -I_N & 0 \\ \alpha(\mathsf{L}_1 + \Delta\mathsf{L}_{\mathbf{q}}) & (\lambda_0+\gamma)I_N & -I_{\mathcal{A}} \\ \mathbf{C}_{\mathrm{x}} & 0 & 0 \\ 0 & \mathbf{C}_{\mathrm{v}} & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}(t_0) \\ \tilde{v}(t_0) \\ \mathbf{u}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \tag{A.4.20}$$

from which and also from (3.1.14) we have

$$\lambda_0 \tilde{x}_i(t_0) = \tilde{v}_i(t_0), \quad \forall\, i \in \mathcal{V}, \tag{A.4.21}$$

$$\alpha\mathsf{L}_1\tilde{x}(t_0) + (\lambda_0+\gamma)\tilde{v}(t_0) - I_{\mathcal{A}}\mathbf{u}_0 \overset{(3.1.14)}{=} \mathbf{0}, \tag{A.4.22}$$

$$\Delta\mathsf{L}_{\mathbf{q}}\tilde{x}(t_0) = \mathbf{0}, \tag{A.4.23}$$

$$\mathbf{C}_{\mathrm{x}}\tilde{x}(t_0) = \mathbf{0}, \quad \mathbf{C}_{\mathrm{v}}\tilde{v}(t_0) = \mathbf{0}. \tag{A.4.24}$$

Then one can conclude from (3.1.5), (A.4.17), (A.4.21), and (A.4.24) that

$$\tilde{x}_i(t_0) = \tilde{v}_i(t_0) = 0 \implies \tilde{x}_i(t) = \tilde{v}_i(t) = 0, \quad \forall\, i \in \mathcal{M} \subset \mathcal{V}, \tag{A.4.25}$$

and by applying the same permutation as defined in (A.1.3) and used in (A.4.12) to equation (A.4.23) as well as by considering (A.4.17) and (A.4.21) that

$$\tilde{x}_i(t_0) = \tilde{x}_j(t_0) \implies \tilde{x}_i(t) = \tilde{x}_j(t), \quad \forall\, i,j \in \mathcal{D}_{\mathrm{c}} \subset \mathcal{V}, \tag{A.4.26}$$

$$\tilde{v}_i(t_0) = \tilde{v}_j(t_0) \implies \tilde{v}_i(t) = \tilde{v}_j(t), \quad \forall\, i,j \in \mathcal{D}_{\mathrm{c}} \subset \mathcal{V}. \tag{A.4.27}$$

Also, as shown in Lemma A.1.1, under condition 1 we have

$$\mathcal{D}_{\mathrm{c}} \cap \mathcal{M} \neq \emptyset, \quad \forall\, \mathrm{c} \in \{1, \cdots, \mathbf{c}\}, \tag{A.4.28}$$

with set $\mathcal{D}_{\rm c}$ given in (A.4.13).

Now under (A.4.28), it is concluded from (A.4.25), (A.4.26)-(A.4.27) that

$$\tilde{x}_i(t) = \tilde{x}_j(t) = 0, \quad \tilde{v}_i(t) = \tilde{v}_j(t) = 0, \quad \forall\, i, j \in \mathcal{D}_{\rm c}, \tag{A.4.29}$$

which by considering (A.4.17) implies that (A.4.15) is simplified to

$$\begin{aligned} &(\mathfrak{e}_i^\top - \mathfrak{e}_j^\top) x^{(\rm m)}(t) = \\ &(\mathfrak{e}_i^\top - \mathfrak{e}_j^\top) \bar{x}^{(\rm m)}(t) = 0, \qquad\qquad \forall\, i, j \in \mathcal{D}_{\rm c}, \quad \forall\, {\rm c} \in \{1, \cdots, \mathbf{c}\}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\quad \forall\, {\rm m} \in \mathbb{N}_0, \quad \forall\, t \in [t_1, +\infty), \end{aligned} \tag{A.4.30}$$

where $\bar{x}_i$ and $\bar{x}_j$ are the elements of state vector $\bar{\mathbf{x}}$ in (A.4.17) denoting the states of an attack-free system that satisfies (3.1.7) (i.e. $\dot{\bar{\mathbf{x}}} = \mathbf{A_q}\bar{\mathbf{x}}$ obtained using $\dot{\mathbf{x}}$-dynamics in (3.1.4) with $\mathbf{B}\mathbf{u}_{\mathcal{A}} = \mathbf{0}$ and unknown initial condition $\bar{\mathbf{x}}_0$ as defined in Proposition 3.1.4). Then using the attack-free dynamics $\dot{\bar{\mathbf{x}}} = \mathbf{A_q}\bar{\mathbf{x}}$, the term $(\mathfrak{e}_i^\top - \mathfrak{e}_j^\top)\bar{x}^{(\rm m)}(t) = 0$ in (A.4.30) can be rewritten as

$$(\mathfrak{e}_i^\top - \mathfrak{e}_j^\top) \mathsf{L}_{\mathbf{q}}^{\rm m} \bar{x}(t) = 0, \quad \forall\, i, j \in \mathcal{D}_{\rm c}, \forall\, {\rm m} \in \mathbb{N}_0, \forall\, t \in [t_1, \infty), \tag{A.4.31}$$

$$(\mathfrak{e}_i^\top - \mathfrak{e}_j^\top) \mathsf{L}_{\mathbf{q}}^{\rm m} \bar{v}(t) = 0, \quad \forall\, i, j \in \mathcal{D}_{\rm c}, \forall\, {\rm m} \in \mathbb{N}_0, \forall\, t \in [t_1, \infty). \tag{A.4.32}$$

Moreover, note that (A.4.31) and (A.4.32) have the same form as equations (109a) and (109b) in [78]. Then under further conditions 2 and 3, it can be verified using the same procedure as in [78, Th. 2] that (A.4.31) and (A.4.32) yield

$$\bar{x}_i(t) = \bar{x}_j(t), \quad \forall\, i, j \in \mathcal{V}, \quad \forall\, t \in [t_1, +\infty), \tag{A.4.33}$$

$$\bar{v}_i(t) = \bar{v}_j(t), \quad \forall\, i, j \in \mathcal{V}, \quad \forall\, t \in [t_1, +\infty), \tag{A.4.34}$$

which means that the entire states of the attack-free system have achieved consensus. Considering the equilibrium subspace (3.1.2) as a result of the consensus protocol (3.1.3), one can conclude that (A.4.33)-(A.4.34) and (3.1.2) coincide. Therefore, from (A.4.30) and (A.4.33)-(A.4.34), obtained under conditions 1-3, one can conclude that stealthy ZDA is undetectable in $\mathbf{r}_0(t)$ of (3.2.11) only in the worst-case scenario that intra-cluster switching links are between agents whose trajectories are not affected by ZDA as well as all of the system (3.1.4)'s attack-free trajectories, characterized in (A.4.30), are at the consensus equilibrium (3.1.2).

Covert attack case: consider (A.4.12) under which a covert attack remains stealthy in a safe mode and note that

$$\begin{aligned}&\mathbf{x}(t)=\bar{\mathbf{x}}(t)+\tilde{\mathbf{x}}(t),\quad \forall\, t\in[t_1,+\infty),\quad \text{with}\\&\tilde{\mathbf{x}}(t)=e^{\mathbf{A}_1(t-t_1)}\tilde{\mathbf{x}}(t_1)+\int_{t_1}^{t}e^{\mathbf{A}_1(t-\boldsymbol{\tau})}\mathbf{B}\mathbf{u}_{\mathcal{A}}(\boldsymbol{\tau})d\boldsymbol{\tau}\end{aligned}\tag{A.4.35}$$

according to the attack model (3.1.6) and Proposition 3.1.4. Given (A.4.35), (A.4.15) can be rewritten as

$$\begin{aligned}&[(\mathfrak{e}_i^\top-\mathfrak{e}_j^\top)\ \ 0]\,\bar{\mathbf{x}}^{(\mathrm{m})}(t)=[(\mathfrak{e}_i^\top-\mathfrak{e}_j^\top)\ \ 0]\,\tilde{\mathbf{x}}^{(\mathrm{m})}(t),\ \ \forall\, i,j\in\mathcal{D}_{\mathrm{c}},\\&\forall\,\mathrm{c}\in\{1,\cdots,\mathbf{c}\},\quad \forall\,\mathrm{m}\in\mathbb{N}_0,\quad \forall\, t\in[t_1,+\infty),\end{aligned}\tag{A.4.36}$$

Notice that the attack-free system states, $\bar{\mathbf{x}}(t)$ in (A.4.35), converge to (3.1.2) as $t\to+\infty$, then the left side of (A.4.36) converges to zero and one can conclude from (A.4.35) and (A.4.36) that continuous states $\tilde{\mathbf{x}}(t)=\mathrm{col}(\tilde{x}(t),\tilde{v}(t))$ exist in either of the following cases

$$\text{case } 1:\tilde{x}_i(t)=\tilde{x}_j(t)\neq 0,\ \ \forall\, i,j\in\mathcal{D}_{\mathrm{c}},\,\forall\, t\in[t_1,+\infty)\tag{A.4.37}$$

$$\text{case } 2: \tilde{x}_i(t) = \tilde{x}_j(t) = 0, \ \ \forall\, i, j \in \mathcal{D}_{\mathrm{c}}, \forall\, t \in [t_1, +\infty) \tag{A.4.38}$$

Note that here case 1 in (A.4.37) implies the attack input $\mathbf{u}_{\mathcal{A}}$ in (A.4.35) has driven and kept the states of agents involved in switching into an unknown equilibrium over the time interval $[t_1, +\infty)$. Also, case 2's interpretation and analysis coincide with that of ZDA in (A.4.29). Then following the same analysis as the ZDA's, one can conclude that under conditions 1-3, covert attack is undetectable in $\mathbf{r}_0(t)$ of (3.2.11) only in the worst-case scenarios that 1) intra-cluster switching links are between agents whose trajectories are identical over time under the effect of covert attack; and 2) intra-cluster switching links are between agents whose trajectories are not affected by covert attack as well as all of the system (3.1.4)'s attack-free trajectories are at the consensus equilibrium (3.1.2).

Appendix B

Proofs of Chapter 4

B.1 Auxiliary Results

Lemma B.1.1. *Let* $\mathsf{L} \in \mathbb{R}^{N\times N}$ *be the Laplacian matrix of a connected graph* $\mathcal{G}$ *and let* $W \in \mathbb{R}^{N\times N}_{\geq 0}$ *be a diagonal matrix with at least one nonzero entry. Then,* $\mathsf{L}+W$ *is a positive definite matrix.*

Proof. The proof follows [141, lemma 1] and is therefore omitted here.

B.2 Proof of Proposition 4.2.2

Let the estimation error be $\mathrm{e}_i = \mathrm{x} - \hat{\mathrm{x}}$ with $\mathrm{e}_i(0) = \mathrm{x}(0)$ as $\hat{\mathrm{x}}(0) = 0$. Then, from (4.1.12) and (4.2.10), the error dynamics is given by

$$\Sigma^{\mathrm{e}_i}_{\mathcal{O}} : \begin{cases} \dot{\hat{\mathrm{e}}}_i = (A_{\sigma(t)} - H^i C_i)\hat{\mathrm{e}}_i + B_{\mathcal{A}}\mathrm{u}_{\mathcal{A}}, & \sigma(t) \in \mathcal{Q}, \\ \mathrm{r}_i = C_i \mathrm{e}_i, & \text{local residual.} \end{cases} \tag{B.2.1}$$

Recall that before attack detection, the system operates in normal mode, $\sigma(t) = 1 \in \mathcal{Q}$. Therefore, we investigate the attack detection in mode $\sigma(t) = 1$.

As described in Section 4.2.1, a system is vulnerable to ZDA if it has unstable zero dynamics (cf. Definition 2.3.1). Accordingly, system $(A_{\sigma(t)} - H^i C_i, B_{\mathcal{A}}, C_i, \sigma(t) = 1)$ in (B.2.1) is vulnerable to ZDA if its matrix pencil $P(\lambda_o)$, given in (4.2.1), is rank deficient for a $\lambda_o \in \mathbb{R}_{>0}$ that is

$$\begin{bmatrix} \lambda_o I_N - (A_1 - H^i C_i) & -B_{\mathcal{A}} \\ C_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathrm{x}^{\mathrm{a}}_0 \\ \mathrm{u}_{\mathcal{A}}(0) \end{bmatrix} = \mathbf{0}, \tag{B.2.2}$$

has a nontrivial solution for state-zero direction $\mathrm{x}_0^{\mathrm{a}} = \mathrm{col}\,(\mathrm{p}_0^{\mathrm{a}}, \mathrm{v}_0^{\mathrm{a}}) \neq \mathbf{0}$ and input-zero direction $\mathrm{u}_{\mathcal{A}}(0) \neq \mathbf{0}$. By using (4.1.10d) and (4.2.9c) in expanding (B.2.2) we obtain

$$\big(\alpha \mathsf{L}_1 + \lambda_o(\lambda_o+\gamma) I_N\big)\mathrm{p}_0^{\mathrm{a}} = \mathrm{B}_{\mathcal{A}} u_{\mathcal{A}}(0), \qquad \lambda_o \mathrm{p}_0^{\mathrm{a}} = \mathrm{v}_0^{\mathrm{a}}, \tag{B.2.3a}$$

$$\mathrm{C}_{\mathrm{p},i}\mathrm{p}_0^{\mathrm{a}} = 0, \qquad \mathfrak{e}_i^{\top}\mathrm{v}_0^{\mathrm{a}} = 0. \tag{B.2.3b}$$

without loss of generality, one can reorder the nodes such that those in set $\mathcal{M}^i$, given in (4.2.9a), come first and accordingly

$$\mathsf{L}_1 = \begin{bmatrix} \mathsf{L}_1^{1,1} & \mathsf{L}_1^{1,2} \\ \mathsf{L}_1^{2,1} & \mathsf{L}_1^{2,2} \end{bmatrix}, \ \mathrm{p}_0^{\mathrm{a}} = \begin{bmatrix} (\mathrm{p}_0^{\mathrm{a}})_1 \\ (\mathrm{p}_0^{\mathrm{a}})_2 \end{bmatrix}, \ \mathrm{v}_0^{\mathrm{a}} = \begin{bmatrix} (\mathrm{v}_0^{\mathrm{a}})_1 \\ (\mathrm{v}_0^{\mathrm{a}})_2 \end{bmatrix}, \tag{B.2.4a}$$

$$\mathrm{B}_{\mathcal{A}} = \begin{bmatrix} I_{|\mathcal{A}|} \\ 0 \end{bmatrix}, \ \mathrm{C}_{\mathrm{p},j} = \begin{bmatrix} I_{\mathbf{k}+1} & 0 \end{bmatrix}, \tag{B.2.4b}$$

where $\mathbf{k} = |\mathcal{N}^i_{\sigma(t)}|$, $\sigma(t) = 1 \in \mathcal{Q}$. Given (B.2.3) and (B.2.4), one can verify that if $\mathcal{A} \subseteq \mathcal{N}^i_{\sigma(t)}$, $\sigma(t) = 1 \in \mathcal{Q} \implies \mathrm{rank}\,(B_{\mathcal{A}}) \leq \mathrm{rank}\,(C_i)$, and a nontrivial solution to (B.2.2) satisfies $(\mathrm{p}_0^{\mathrm{a}})_1 = \mathbf{0}$ and $(\alpha \mathsf{L}_1^{2,2} + \lambda_o(\lambda_o+\gamma)I)(\mathrm{p}_0^{\mathrm{a}})_2 = \mathbf{0}$. Noticing that $(\alpha \mathsf{L}_1^{2,2} + \lambda_o(\lambda_o+\gamma)I)$ is positive definite for any $\lambda_o \in \mathbb{R}_{>0}$ associated with an unstable invariant zero (cf. Lemma B.1.1), it is concluded that a nontrivial solution, i.e. $(\mathrm{p}_0^{\mathrm{a}})_2 \neq 0$, only exists for $\lambda_o < 0$ and thus the system's zero dynamics is stable.

In terms of the detectability of covert attacks, note that in the ZDA analysis, we showed the output-nulling of non-vanishing intrusions is not feasible if $\mathcal{A} \subseteq \mathcal{N}^i_{\sigma(t)}$, $\sigma(t) = 1 \in \mathcal{Q}$. Also, local measurements (4.2.9) are not subject to alterations by the attacker. Therefore, any non-vanishing intrusion u_a in (B.2.1) yields a non-vanishing residual i.e. $\lim_{t\to\infty} \mathrm{r}_i \neq \mathbf{0}$.

B.3 Proof of Proposition 4.2.3

Note that from Proposition 4.2.2, the i-th UAV with a local detector $\Sigma_{\mathcal{O}}^{i},\ i \in \mathcal{D}$ can detect stealthy attacks within its set of immediate neighbors in normal mode i.e. $\mathcal{N}_{\sigma(t)}^{i},\ \sigma(t) = 1 \in \mathcal{Q}$. Therefore, by induction, a set $\mathcal{D}$ of local detectors $\Sigma_{\mathcal{O}}^{i}$'s holding (4.2.13) covers the entire network set $\mathcal{V}$ of UAVs and thus is sufficient to detect stealthy attacks anywhere in the communication network of UAVs.

Appendix C
Proofs of Chapter 5

C.1 Auxiliary Results

The following result quantifies the inaccessible state measurements for the system in (5.1.1) with the collective measurements in (5.3.1) obtained under Assumptions 5.2.1 and 5.2.2. It shows that the only states that are inaccessible at any time for any cooperative agent are the velocity states of the malicious agents.

Lemma C.1.1. *Consider the system in* (5.1.1)-(5.1.2) *over a time interval* $[t_0,\, t_0{+}T)$ *under Assumptions 5.2.1 and 5.2.2. Let* $\mathcal{I}_i = \mathcal{V}_\sigma^{i''}$ *in* (5.1.2) *and let* (5.1.1) *be subject to an* F*-total (resp.* F*-local) adversary set with* $0 \leq F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) - 1$ *(resp.* $0 \leq F \leq r(\boldsymbol{\mathcal{G}}_T^\mu) - 1$*). Then, the nullspace of the matrix* $\mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}}$ *in* (5.3.1)*, defined uniformly over* $[t_0,\, t_0 + T)$, $\forall\, t_0 \in \mathbb{R}_{\geq 0}$ *is given by*

$$\mathcal{X}_{[t_0,\, t_0+T)} = \cap_{\sigma\in\mathcal{Q}'} \ker \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}} = \operatorname{span}\left\{ \begin{bmatrix} \mathbf{0}_N \\ \mathfrak{e}_N^i \end{bmatrix},\, \forall\, i \in \mathcal{A} \right\}. \tag{C.1.1}$$

Proof. Without loss of generality, let $\mathbf{x}$ in (5.1.1) be partitioned as $\mathbf{x} = \operatorname{col}(\widetilde{\boldsymbol{p}}_{\mathcal{V}\setminus\mathcal{A}}, \widetilde{\boldsymbol{p}}_{\mathcal{A}}, \boldsymbol{v}_{\mathcal{V}\setminus\mathcal{A}}, \boldsymbol{v}_{\mathcal{A}}) \in \mathbb{R}^{2N}$. Also, from the definition of the nullspace, we have

$$\mathcal{X}_{[t_0,\, t_0+T)} = \cap_{\sigma\in\mathcal{Q}'} \ker \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}} = \left\{ \mathbf{x} \in \mathbb{R}^{2N} \mid \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}} \mathbf{x} = \mathbf{0},\ \ \forall\, \sigma \in \mathcal{Q}' \right\}, \tag{C.1.2}$$

which by using (5.1.2) and (5.3.1) yields $\forall\, \sigma \in \mathcal{Q}'$,

$$\widetilde{\boldsymbol{p}}_i = \boldsymbol{v}_i = 0, \qquad\qquad \forall\, i \in \mathcal{V}\setminus\mathcal{A}, \tag{C.1.3a}$$

$$\widetilde{\boldsymbol{p}}_j = 0, \qquad \forall\, j \in \mathcal{V}_\sigma^{i''}, \qquad i \in \mathcal{V}\setminus\mathcal{A}. \tag{C.1.3b}$$

where we used $\mathcal{I}_i = \mathcal{V}_\sigma^{i''}$. Then, (C.1.3) can be rewritten as

$$\forall\, \sigma \in \mathcal{Q}',\ \begin{bmatrix} I_{|\mathcal{V}\setminus\mathcal{A}|} & \mathbf{0} \\ \mathbf{0} & \mathbb{I}(\mathsf{A}_\sigma) \\ \star & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{p}}_{\mathcal{V}\setminus\mathcal{A}} \\ \widetilde{\boldsymbol{p}}_{\mathcal{A}} \end{bmatrix} = \mathbf{0},\quad \begin{bmatrix} I_{|\mathcal{V}\setminus\mathcal{A}|} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_{\mathcal{V}\setminus\mathcal{A}} \\ \boldsymbol{v}_{\mathcal{A}} \end{bmatrix} = \mathbf{0}, \tag{C.1.4}$$

where $\star$ is a binary matrix, determining redundant position measurements of the neighboring cooperative agents, and whose structure is immaterial to this particular analysis. $\mathbb{I}(\mathsf{A}_\sigma)$ is a binary matrix-valued function of the adjacency matrix, determining the availability of the position measurements of the malicious agents $\mathcal{A}$ for the cooperative agents $\mathcal{V}\setminus\mathcal{A}$. $\mathbb{I}(\mathsf{A}_\sigma)$ is defined as

$$\begin{aligned} \mathbb{I}(\mathsf{A}_\sigma) &= \begin{bmatrix} \mathbb{I}_1(\mathsf{A}_\sigma^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}}) \\ \vdots \\ \mathbb{I}_{|\mathcal{V}\setminus\mathcal{A}|}(\mathsf{A}_\sigma^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}}) \end{bmatrix} = \begin{bmatrix} \mathrm{diag}(\mathrm{r}ow_1(\mathsf{A}_\sigma^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}}) \\ \vdots \\ \mathrm{diag}(\mathrm{r}ow_{|\mathcal{V}\setminus\mathcal{A},\mathcal{A}|}(\mathsf{A}_\sigma^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}}) \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{d}iag(a_{i_1 j_1}^{\sigma(t)}, & \dots, & a_{i_1 j_{|\mathcal{A}|}}^{\sigma(t)}) \\ & \vdots & \\ \mathrm{d}iag(a_{i_{|\mathcal{V}\setminus\mathcal{A}|} j_1}^{\sigma(t)}, & \dots, & a_{i_{|\mathcal{V}\setminus\mathcal{A}|} j_{|\mathcal{A}|}}^{\sigma(t)}) \end{bmatrix}, \end{aligned} \tag{C.1.5}$$

where $\mathsf{A}_\sigma^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}}$, taken from $\mathsf{A}_{\sigma(t)} = \begin{bmatrix} \mathsf{A}_{\sigma(t)}^{\mathcal{V}\setminus\mathcal{A}} & \mathsf{A}_{\sigma(t)}^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}} \\ \mathsf{A}_{\sigma(t)}^{\mathcal{A},\mathcal{V}\setminus\mathcal{A}} & \mathsf{A}_{\sigma(t)}^{\mathcal{A}} \end{bmatrix}$, indicates the mode-dependent communication links between the sets $\mathcal{V}\setminus\mathcal{A} = \{i_1, \dots, i_{|\mathcal{V}\setminus\mathcal{A}|}\}$ and $\mathcal{A} = \{j_1, \dots, j_{|\mathcal{A}|}\}$.

Recall that consistent with (3.1.3), each element, $a_{ij}^{\sigma(t)}$ in (5.1.1), is equal to 1 if two distinct agents i, j are in communication over a link $(i,j) \in \mathcal{E}_\sigma$ and is 0 otherwise. We next show that the $\cap_{\sigma\in\mathcal{Q}'} \ker \mathbb{I}(\mathsf{A}_\sigma) = \emptyset$ obtained uniformely over a time interval $[t_0,\, t_0{+}T)$, $\forall\, t_0 \in \mathbb{R}_{\geq 0}$. Under Assumptions 5.2.1 and 5.2.2, it follows from Proposition 5.2.4 that each agent has at least $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu)$ neighbors over $[t_0,\, t_0 + T)$, $\forall\, t_0 \in \mathbb{R}_{\geq 0}$, not

necessarily connected to at all time instants and that $r(\boldsymbol{\mathcal{G}}_T^\mu) \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu)$. Therefore, in the case of an F-local (resp. F-total) adversary set with $F \leq r(\boldsymbol{\mathcal{G}}_T^\mu) - 1$ (resp. $F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) - 1$), each malicious agent $j \in \mathcal{A} \subset \mathcal{V}$ should have at least one neighbor outside of its set, i.e. the set $\mathcal{V} \setminus \mathcal{A}$, over some period of time. Formally,

$$\begin{aligned}&\forall\, j \in \mathcal{A},\ \ \exists\, i \in \mathcal{V} \setminus \mathcal{A},\ \exists\, \sigma \in \mathcal{Q}'\ \text{ s.t. }\ j \in \mathcal{N}_\sigma^{i(1)} \subseteq \mathcal{V}_\sigma^{i''} \iff \\ &(i,j) \in \mathcal{E}_\sigma \overset{\text{Lemma 5.2.3}}{\iff} \frac{1}{T}\int_{t_0}^{t_0+T} a_{ij}^{\sigma(\tau)} \mathrm{d}\tau \geq \delta,\ \forall\, t_0 \in \mathbb{R}_{\geq 0},\end{aligned}$$

by which, one can readily show for $\mathbb{I}(\mathsf{A}_\sigma)$ in (C.1.5) that

$$\begin{aligned}(1/T)\int_{t_0}^{t_0+T} \mathbb{I}^\top(\mathsf{A}_\sigma)\mathbb{I}(\mathsf{A}_\sigma)\,\mathrm{d}\tau \geq \underline{\delta} I_{|\mathcal{A}|},\quad \forall\, t_0 \in \mathbb{R}_{\geq 0},\ \exists\, \underline{\delta} \in \mathbb{R}_{>0} \implies \\ \cap_{\sigma\in\mathcal{Q}'} \ker \mathbb{I}(\mathsf{A}_\sigma) = \emptyset. \qquad \text{(C.1.6)}\end{aligned}$$

Then, from (C.1.6) and (C.1.4) we can conclude for (C.1.2) that $\forall\, \sigma \in \mathcal{Q}'$, $\mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}}\mathbf{x} = \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}} \begin{bmatrix} \mathbf{0}_{|\mathcal{V}\setminus\mathcal{A}|} \\ \mathbf{0}_{|\mathcal{A}|} \\ \mathbf{0}_{|\mathcal{V}\setminus\mathcal{A}|} \\ \boldsymbol{v}_{\mathcal{A}} \end{bmatrix} = \mathbf{0},\ \forall\, \boldsymbol{v}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$, which in turn implies $\cap_{\sigma\in\mathcal{Q}'} \ker \mathbf{C}_\sigma^{\mathcal{V}\setminus\mathcal{A}} = \text{span}\left\{ \begin{bmatrix} \mathbf{0}_{|\mathcal{V}|} \\ \mathbf{0}_{|\mathcal{V}\setminus\mathcal{A}|} \\ \mathbf{1}_{|\mathcal{A}|} \end{bmatrix} \right\}$. This concludes the proof.

C.2 Proof of Lemma 5.2.3

The proof of the equivalence of the statements is achieved by demonstrating that each ensures positive algebraic connectivity in an integral sense for the graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$, that is $\lambda_2(\mathsf{L} = \frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau) > \mu$ holds $\forall\, t \in \mathbb{R}_{\geq 0}$ and $\exists\, \mu, T \in \mathbb{R}_{>0}$ as defined in (5.2.1). Note that $\mathbf{1}_N\mathbf{1}_N^\top/N$ and $Q^\top Q = I_N - \mathbf{1}_N\mathbf{1}_N^\top/N$ are both orthogonal projection matrices (i.e. $(\mathbf{1}_N\mathbf{1}_N^\top/N)^2 = \mathbf{1}_N\mathbf{1}_N^\top/N = (\mathbf{1}_N\mathbf{1}_N^\top/N)^\top$) and thus their corresponding spectrum belongs to $\{0, 1\}$ which in turn implies $\mathbf{1}_N\mathbf{1}_N^\top/N$ and $Q^\top Q$ are positive semi-definite. Then, from $\text{rank}(\mathbf{1}_N\mathbf{1}_N^\top/N) = 1$ and the construction of Q

in (5.2.2), one can conclude

$$\operatorname{spec}(\mathbf{1}_N\mathbf{1}_N^\top/N) = \{\overbrace{0,\cdots,0}^{N-1},1\}, \tag{C.2.1a}$$

$$\operatorname{spec}(Q^\top Q) = \{0,\overbrace{1,\cdots,1}^{N-1}\}, \tag{C.2.1b}$$

$$\operatorname{spec}(Q\mathsf{L}_{\sigma(\tau)}Q^\top) = \operatorname{spec}(\mathsf{L}_{\sigma(\tau)})\setminus\{0\} \implies$$

$$\lambda_1(Q\mathsf{L}_{\sigma(\tau)}Q^\top) = \lambda_2(\mathsf{L}_{\sigma(\tau)}), \tag{C.2.1c}$$

where (C.2.1c) is from [32], and (C.2.1b) follows from the fact that $\mathbf{1}_N\mathbf{1}_N^\top/N$ and $Q^\top Q = I_N - \mathbf{1}_N\mathbf{1}_N^\top/N$ are orthogonal projections and that $\operatorname{rank}(\mathbf{1}_N\mathbf{1}_N^\top/N) = 1$. Then, there exists a unitary matrix U, where $UU^\top = I_N$, such that $\mathbf{1}_N\mathbf{1}_N^\top/N = U\operatorname{diag}(1,0,\ldots,0)U^\top$. Now, one can write

$$\begin{aligned} Q^\top Q &= I_N - \mathbf{1}_N\mathbf{1}_N^\top/N = UU^\top - U\operatorname{diag}(1,0,\ldots,0)U^\top \\ &= U(I_N - \operatorname{diag}(1,0,\ldots,0))U^\top = U\operatorname{diag}(0,1,\ldots,1)U^\top. \end{aligned}$$

$2 \implies 1$: if 2 holds, then $(1/T)\int_t^{t+T}(\mathsf{L}_{\sigma(\tau)} + \mathbf{1}_N\mathbf{1}_N^\top/N)\,\mathrm{d}\tau$ is positive definite. Then, note that $0 = \lambda_1(\mathsf{L}_{\sigma(t)}) \le \lambda_2(\mathsf{L}_{\sigma(t)}) \le \cdots \le \lambda_N(\mathsf{L}_{\sigma(t)}) \le N$ at any time instant (see [48, Corrollary 13.1.4]) and that $\operatorname{spec}(\mathbf{1}_N\mathbf{1}_N^\top/N) = \{\overbrace{0,\ldots,0}^{N-1},1\}$ in (C.2.1a) has the eigenpair $(1,\mathbf{1}_N)$ which shares the eigenvector $\mathbf{1}_N$ with the eigenpair $(0,\mathbf{1}_N)$ of $\mathsf{L}_{\sigma(t)}$. This implies that [3, Thm. 1]

$$\lambda_2\left(\frac{1}{T}\int_t^{t+T}\mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau\right) \ge \mu_m. \tag{C.2.2}$$

Also, note that (5.2.4) can be rewritten as

$$\mu_m I_N - \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} \leq \frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau \leq \mu_M I_N - \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N},$$

which by pre-multiplying and post-multiplying, respectively, by Q and $Q^\top$, using (5.2.2), and invoking Proposition 8.1.2 [12], yields

$$\mu_m I_{N-1} \leq \frac{1}{T}\int_t^{t+T} Q\mathsf{L}_{\sigma(\tau)}Q^\top\,\mathrm{d}\tau \leq \mu_M I_{N-1}, \quad \forall\, t \in \mathbb{R}_{\geq 0},$$

that is equivalent to (5.2.1) with $\mu_m = \mu$. The existence of the upper bound, $\mu_M I_{N-1}$, for (5.2.1) is trivial because of the boundedness of $a_{ij}^{\sigma(t)}$'s in the adjacency matrix and the integration over a finite interval $[t,\, t+T)$, $\forall\, t \in \mathbb{R}_{\geq 0},\, \exists\, T \in \mathbb{R}_{>0}$.

$1 \implies 2$: Let (5.2.1) hold. It follows from (5.2.1) and (C.2.1c) that $\lambda_2(\frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau) \geq \mu$ (cf. (C.2.2)). Also by following the same argument as given in the previous part of the proof, (5.2.1) admits an upper bound, denoted by $\mu'_M I_{N-1}$, where $\mu'_M \in \mathbb{R}_{>0}$. Now, by pre-multiplying and post-multiplying (5.2.1), respectively, by $Q^\top$ and Q, and invoking Proposition 8.1.2 [12], we obtain

$$\mu Q^\top Q \leq \frac{1}{T}\int_t^{t+T} Q^\top Q\mathsf{L}_{\sigma(\tau)}Q^\top Q\,\mathrm{d}\tau \leq \mu'_M Q^\top Q, \tag{C.2.3}$$

which by considering (5.2.2) and the fact that $\mathsf{L}_{\sigma(\tau)}Q^\top Q = Q^\top Q\mathsf{L}_{\sigma(\tau)} = \mathsf{L}_{\sigma(\tau)}$ yields

$$\mu Q^\top Q \leq \frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau \leq \mu'_M Q^\top Q, \tag{C.2.4}$$

in which $Q^\top Q$ is positive semi-definite (see (C.2.1)). By adding $\mathbf{1}_N \mathbf{1}_N^\top/N$ to the sides of the inequality (C.2.4) we obtain

$$\mu Q^\top Q + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} \leq \frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} = \frac{1}{T}\int_t^{t+T} \left(\mathsf{L}_{\sigma(\tau)} + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N}\right)\mathrm{d}\tau \leq \mu' Q^\top Q + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N}, \tag{C.2.5}$$

where it follows from (C.2.1) that the left and right side matrices are bounded and positive definite such that $\mu Q^\top Q + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} = U\,\mathrm{diag}(1, \mu, \ldots, \mu)U^\top$ and $\mu' Q^\top Q + \frac{\mathbf{1}_N \mathbf{1}_N^\top}{N} = U\,\mathrm{diag}(1, \mu, \ldots, \mu')U^\top$, making the foregoing condition equivalent to (5.2.4) with $\mu_m = \min\{1, \mu\}$ and $\mu_M = \max\{1, \mu'\}$.

2 $\iff$ 3: This has been proven in [3, Thm. 1]. We conclude the proof by restating that

$$\lambda_2\Big(\underbrace{\frac{1}{T}\int_t^{t+T} \mathsf{L}_{\sigma(\tau)}\,\mathrm{d}\tau}_{\overset{(5.2.6)}{=}\mathsf{L}}\Big) = \lambda_1\Big(\frac{1}{T}\int_t^{t+T} Q\mathsf{L}_{\sigma(\tau)}Q^\top\,\mathrm{d}\tau\Big) \geq \mu.$$

C.3 Proof of Proposition 5.2.4

It follows from Lemma 5.2.3 that a (μ, T)-PE connected $\mathcal{G}_{\sigma(t)}$ forms, uniformly in time, the connected static graph $\boldsymbol{\mathcal{G}}_T^\mu = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^\mu)$ with the edge set $\boldsymbol{\mathcal{E}}_T^\mu$ in (5.2.5) and algebraic connectivity $\lambda_2(\mathsf{L}) \geq \mu > 0$. Then, from [120] and [116, Thm. 2] we have $\left\lceil \frac{\lambda_2(\mathsf{L})}{2} \right\rceil \leq r(\boldsymbol{\mathcal{G}}_T^\mu)$ for $0 \leq r(\boldsymbol{\mathcal{G}}_T^\mu) \leq \lceil |\mathcal{V}|/2 \rceil$, and one can conclude from [120] and [67, Thm. 6] that $r(\boldsymbol{\mathcal{G}}_T^\mu) \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu)$. It also follows from [48, Ch. 13.5] for any simple non-complete graph $\boldsymbol{\mathcal{G}}_T^\mu$, that $\lambda_2(\mathsf{L}) \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) \leq |\mathcal{V}| - 1$. Finally, note that $0 = \lambda_1(\mathsf{L}) < \lambda_2(\mathsf{L}) \leq \cdots \leq \lambda_N(\mathsf{L}) \leq |\mathcal{V}| = N$ holds for $\boldsymbol{\mathcal{G}}_T^\mu$ and that the equality $\lambda_2(\mathsf{L}) = N$ holds for complete graphs [48, Corr. 13.1.4]. Then (5.2.7) is concluded.

C.4 Proof of theorem 5.2.5

Recall from Proposition 5.2.4 that a (μ, T)-PE connected $\mathcal{G}_{\sigma(t)}$ forms, uniformly in time, a static network $\boldsymbol{\mathcal{G}}_T^\mu = (\mathcal{V}, \boldsymbol{\mathcal{E}}_T^\mu)$ whose robustness $r(\boldsymbol{\mathcal{G}}_T^\mu)$ and vertex-connectivity $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu)$ are lower bounded by $\lceil \mu/2 \rceil$. It then follows from [146, Thm. 1] for an r-robust network $\boldsymbol{\mathcal{G}}_T^\mu$, where $r \leq r(\boldsymbol{\mathcal{G}}_T^\mu)$, that, by definition, no F-local adversary subset $\mathcal{A} \subset \mathcal{V}$, where $F \leq r - 1$, make a vertex cutset for $\boldsymbol{\mathcal{G}}_T^\mu$. Therefore, the removal of up to $F \leq r - 1$ malicious nodes (agents) and their incident edges from the neighbors of the remaining nodes (cooperative agents) does not render the induced subgraph $\bar{\boldsymbol{\mathcal{G}}} = (\mathcal{V} \setminus \mathcal{A}, \bar{\boldsymbol{\mathcal{E}}})$ disconnected (equiv., $\lambda_2(\bar{\mathbf{L}}) > 0$, where $\bar{\mathbf{L}}$ is the Laplacian matrix of $\bar{\boldsymbol{\mathcal{G}}}$).

Likewise, given the vertex-connectivity $\boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu)$, it follows, by definition, that no F-total adversary subset $\mathcal{A} \subset \mathcal{V}$, with $F \leq \boldsymbol{\kappa} - 1 \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) - 1$, make a vertex cutset for $\boldsymbol{\mathcal{G}}_T^\mu$. Therefore, the removal of up to $F \leq \boldsymbol{\kappa} - 1$ (malicious) nodes, in total, and their incident edges does not render the induced subgraph $\bar{\boldsymbol{\mathcal{G}}} = (\mathcal{V} \setminus \mathcal{A}, \bar{\boldsymbol{\mathcal{E}}})$ disconnected (equivalently, $\lambda_2(\bar{\mathbf{L}}) > 0$).

Finally, note that the induced subgraph $\bar{\mathcal{G}}_{\sigma(t)} = (\bar{\mathcal{V}}, \bar{\mathcal{E}}_{\sigma(t)})$ associated with the connected graph $\bar{\boldsymbol{\mathcal{G}}}$ with $0 < \lambda_2(\bar{\mathbf{L}}) =: \bar{\mu}$ meets the conditions in Lemma 5.2.3-*3* for some $\bar{T} \leq T$ (where the inequality holds because $|\bar{\mathcal{V}}| < |\mathcal{V}|$ and $|\bar{\boldsymbol{\mathcal{E}}}| < |\boldsymbol{\mathcal{E}}_T^\mu|$). Moreover, it follows from [48, Thm. 13.5.1] for the graph $\boldsymbol{\mathcal{G}}_T^\mu$ and its induced subgraph $\bar{\boldsymbol{\mathcal{G}}}$, resp., with $\mathbf{L}$ and $\bar{\mathbf{L}}$ that $\lambda_2(\mathbf{L}) \leq \lambda_2(\bar{\mathbf{L}}) + |\mathcal{A}|$. Then, from $\lambda_2(\mathbf{L}) \geq \mu$ (see (5.2.7)) and $\bar{\mu} = \lambda_2(\bar{\mathbf{L}}) > 0$, one can conclude $\mu \leq \bar{\mu} + |\mathcal{A}|$.

C.5 Proof of Proposition 5.2.6

The first part of the proof has two steps similar to that in [32, Thm. 1]. First, we consider a system of the form

$$\dot{\chi} = -\frac{\alpha}{\gamma}\underline{\mathsf{L}}_{\sigma(t)}\chi, \qquad \chi(t_0) \in \mathbb{R}^{N-1}, \tag{C.5.1}$$

in which $\underline{\mathsf{L}}_{\sigma(t)} = Q\mathsf{L}_{\sigma(\tau)}Q^\top$ satisfies the (μ, T)-PE condition in (5.2.1). It follows from [40, lemma 1] that[1] (C.5.1) is globally uniformly exponentially stable (GUES) with the convergence rate $\lambda_\chi \in \mathbb{R}_{>0}$ such that

$$\|\chi(t)\| \leq \kappa_\chi \|\chi(0)\| e^{-\lambda_\chi t}, \ \ \forall\, t \in \mathbb{R}_{\geq 0}, \tag{C.5.2}$$

in which $\kappa_\chi = \sqrt{\frac{\alpha N}{\gamma\lambda_\chi}}$ and $\lambda_\chi = \eta e^{-2\eta T}$ with $\eta = -\frac{1}{2T}\ln(1 - \frac{(\alpha/\gamma)\mu T}{1+(\alpha/\gamma)^2N^2T^2})$, where N is obtained from $\left\|\mathsf{L}_{\sigma(t)}\right\| \leq N$ (see [48, Corrollary 13.1.4]). Given the GUES of (C.5.1) under condition (5.2.1), it follows from [73, Lemma 1] and [56, Thm. 4.12] that there exists a Lyapunov function $v(t, \chi(t)) = \chi(t)^\top P(t)\chi(t)$ with $P(t) = P(t)^\top \in \mathbb{R}^{(N-1)\times(N-1)} > 0$ such that $\forall\, t \in \mathbb{R}_{\geq 0}$ the following inequalities hold:

$$0 < \frac{\gamma}{2\alpha N} I_{N-1} \leq P(t) \leq \frac{1}{2\lambda_\chi} I_{N-1}, \tag{C.5.3a}$$

$$\dot{P}(t) - \frac{\alpha}{\gamma}\underline{\mathsf{L}}_{\sigma(t)}P(t) - \frac{\alpha}{\gamma}P(t)\underline{\mathsf{L}}_{\sigma(t)} + I_{N-1} = \mathbf{0}. \tag{C.5.3b}$$

In the second step, we use the stability properties of (C.5.1) as given in (C.5.2) and (C.5.3) in the stability analysis of (5.1.1). By defining an intermediary state

[1] We note [40, lemma 1] has defined the function of the form $\underline{\mathsf{L}}_{\sigma(t)}$ in (C.5.1) to be continuous. Yet, a unique solution to (C.5.1) exists (see [56, Thm. 3.2]) in the case $\underline{\mathsf{L}}_{\sigma(t)}$ is piecewise continuous and bounded with a finite set of point-wise discontinuities, and the results hold as stated herein.

transformation as given by

$$\boldsymbol{\chi} = \begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \gamma I_{N-1} & Q \\ \mathbf{0}_{N\times(N-1)} & I_N \end{bmatrix} \begin{bmatrix} \zeta \\ \boldsymbol{v} \end{bmatrix} = C^{-1}\mathcal{Y}, \tag{C.5.4}$$

the system $\Sigma_{\sigma(t)}$ in (5.1.1) with (5.2.8) can be rewritten as

$$\dot{\boldsymbol{\chi}} = \begin{bmatrix} -\frac{\alpha}{\gamma}Q\mathsf{L}_{\sigma(t)}Q^\top & +\frac{\alpha}{\gamma}Q\mathsf{L}_{\sigma(t)} \\ -\frac{\alpha}{\gamma}\mathsf{L}_{\sigma(t)}Q^\top & -(\gamma I_N - \frac{\alpha}{\gamma}\mathsf{L}_{\sigma(t)}) \end{bmatrix} \boldsymbol{\chi} + \begin{bmatrix} QI_{\mathcal{A}} \\ I_{\mathcal{A}} \end{bmatrix} \mathbf{u}_{\mathcal{A}}, \tag{C.5.5a}$$

$$\mathcal{Y} = C\boldsymbol{\chi}, \tag{C.5.5b}$$

Associated to (C.5.5), a Lyapunov function is defined as

$$V(t,\boldsymbol{\chi}(t)) = \begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix}^\top \begin{bmatrix} P(t) & \mathbf{0} \\ \mathbf{0} & \frac{\beta}{2}I_N \end{bmatrix} \begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix}, \tag{C.5.6}$$

where $P(t)$ is given in (C.5.3) and $\beta \in \mathbb{R}_{>0}$. Taking the derivative of $V(t,\boldsymbol{\chi}(t))$ along the trajectories of (C.5.5) yields

$$\dot{V}(t,\boldsymbol{\chi}(t)) = -\begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix}^\top M \begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix} + 2\begin{bmatrix} \xi \\ \boldsymbol{v} \end{bmatrix}^\top \begin{bmatrix} P(t)QI_{\mathcal{A}} \\ \frac{\beta}{2}I_{\mathcal{A}} \end{bmatrix} \mathbf{u}_{\mathcal{A}},$$

where

$$M = \begin{bmatrix} -(\dot{P}(t) - \frac{\alpha}{\gamma}P(t)\underline{\mathsf{L}}_{\sigma(t)} - \frac{\alpha}{\gamma}\underline{\mathsf{L}}_{\sigma(t)}P(t)) & \frac{\alpha}{\gamma}(\frac{\beta}{2}I_{N-1} - P(t))Q\mathsf{L}_{\sigma(t)} \\ \mathsf{L}_{\sigma(t)}Q^\top(\frac{\beta}{2}I_{N-1} - P(t))\frac{\alpha}{\gamma} & \beta(\gamma I_N - \frac{\alpha}{\gamma}\mathsf{L}_{\sigma(t)}) \end{bmatrix}.$$

By using (C.5.3), $\left\|\mathsf{L}_{\sigma(t)}\right\| \leq N$, and $\|Q\| \leq 1$, we obtain

$$\dot{V}(t,\boldsymbol{\chi}(t)) \leq -\begin{bmatrix}\|\xi\|\\ \|\boldsymbol{v}\|\end{bmatrix}^{\top} \overbrace{\begin{bmatrix}1 & -\frac{\alpha}{\gamma}(\beta+\frac{1}{\lambda_\chi})\frac{N}{2}\\ -\frac{\alpha}{\gamma}(\beta+\frac{1}{\lambda_\chi})\frac{N}{2} & \beta(\gamma-\frac{\alpha}{\gamma}N)\end{bmatrix}}^{\overline{M}} \begin{bmatrix}\|\xi\|\\ \|\boldsymbol{v}\|\end{bmatrix} + \max\{\lambda_\chi^{-1},\beta\}\begin{bmatrix}\|\xi\|\\ \|\boldsymbol{v}\|\end{bmatrix}^{\top} \mathbf{1}_2 \|\mathbf{u}_{\mathcal{A}}\|\,. \tag{C.5.7}$$

Note that by selecting $\lambda_{\mathbf{x}} < \lambda_\chi$ and a sufficiently large γ (e.g., $\gamma = \alpha N$, for $\alpha \geq 1$), one can verify that the following matrix is positive definite[2],

$$\overline{M} - 2\lambda_{\mathbf{x}}\begin{bmatrix}\frac{1}{2\lambda_\chi} & 0\\ 0 & \frac{\beta}{2}\end{bmatrix} = \begin{bmatrix}(1-\frac{\lambda_{\mathbf{x}}}{\lambda_\chi}) & -\frac{\alpha}{\gamma}(\beta+\frac{1}{\lambda_\chi})\frac{N}{2}\\ -\frac{\alpha}{\gamma}(\beta+\frac{1}{\lambda_\chi})\frac{N}{2} & \beta(\gamma-\frac{\alpha}{\gamma}N-\lambda_{\mathbf{x}})\end{bmatrix} > 0\,. \tag{C.5.8}$$

Then from (C.5.3), (C.5.6), (C.5.7), and (C.5.8), we obtain

$$\dot{V}(t,\boldsymbol{\chi}(t)) \leq -2\lambda_{\mathbf{x}} V(t,\boldsymbol{\chi}(t)) + \sqrt{2}\max\{\lambda_\chi^{-1},\beta\}\,\|\boldsymbol{\chi}(t)\|\|\mathbf{u}_{\mathcal{A}}\|\,.$$

Applying the comparison lemma [56, Lemma 3.4] and considering (C.5.3) and (C.5.6) yields

$$\|\boldsymbol{\chi}(t)\| \leq \sqrt{\frac{\max\left\{\lambda_\chi^{-1},\beta\right\}}{\min\left\{\frac{\gamma}{\alpha N},\beta\right\}}}\,\|\boldsymbol{\chi}(t_0)\|\,e^{-\lambda_{\mathbf{x}}(t-t_0)} + \frac{\max\left\{\lambda_\chi^{-1},\beta\right\}}{\lambda_{\mathbf{x}}\min\left\{\frac{\gamma}{2\alpha N},\frac{\beta}{2}\right\}}\sup_{t_0\leq t\leq T_d}\|\mathbf{u}_{\mathcal{A}}(t)\|\,,\ \forall\, t\geq t_0\in\mathbb{R}_{\geq 0}. \tag{C.5.9}$$

Now one can conclude from (C.5.9) that the origin $\boldsymbol{\chi} = \mathbf{0}$, is GUES equilibrium of the

[2]The determinant of (C.5.8) yields a cubic function of γ, to which applying the Routh's stability criterion indicates the existence of one positive root.

unforced system (C.5.5) (i.e., $\mathbf{u}_{\mathcal{A}} = \mathbf{0}$). Additionally, (C.5.5) is input-to-state stable (ISS) in the case $\mathbf{u}_{\mathcal{A}} \neq \mathbf{0}$, provided $\sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| < \infty$ for every $T_d \in [0, \infty)$. It then follows from (C.5.9) that

$$\|(\mathcal{Y})_{T_d}\|_{\mathcal{L}_p} \leq \kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\| + \kappa_{\mathbf{u}} \|(\mathbf{u}_{\mathcal{A}})_{T_d}\|_{\mathcal{L}_p}, \tag{C.5.10}$$

where we used (5.2.8), (C.5.4), $\|\mathcal{Y}(t_0)\| \leq \|\mathbf{x}(t_0)\|$, and $\kappa_{\mathbf{x}}$ and $\kappa_{\mathbf{u}}$ as given in (5.2.9). One can conclude from (C.5.10) that for every $\|\mathbf{x}(t_0)\| \leq \infty$ and every $\mathbf{u}_{\mathcal{A}} \in \mathcal{L}_{pe}$ with $\sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| < \infty$, the system in (C.5.5) (equiv. $\Sigma_{\sigma(t)}$ in (5.1.1)) with the output $\mathcal{Y}(t)$, associated with (5.1.8), is finite-gain $\mathcal{L}_p$ stable (see [56, Thm. 5.1 and Corollary 5.1]).

Finally, to calculate the bounds in (5.2.10), note that

$$Q^\top \zeta \overset{(5.2.8)}{=} Q^\top Q \widetilde{\boldsymbol{p}} \overset{(5.2.2)}{=} \widetilde{\boldsymbol{p}} - \mathbf{1}_N \mathbf{p}_{avg}, \quad \mathbf{p}_{avg} = (\frac{1}{N} \mathbf{1}_N^\top \widetilde{\boldsymbol{p}}), \tag{C.5.11}$$

and let Q be partitioned as $Q = \Big[q_1 \,|\, q_2 \,|\, \cdots \,|\, q_N \Big]$, where $q_i \in \mathbb{R}^{N-1}$. We also have $Q^\top Q = I_N - (1/N)\mathbf{1}_N^\top \mathbf{1}_N \implies \|q_i\|^2 = 1 - 1/N$ for every $i \in \{1, \ldots, N\}$. Then, using (C.5.11), one can write for every $i, j \in \mathcal{V}$ that

$$\begin{aligned} \left|\widetilde{\boldsymbol{p}}_i(t) - \widetilde{\boldsymbol{p}}_j(t)\right| &= \left|q_i^\top \zeta(t) - q_j^\top \zeta(t)\right| \leq \left\|q_i^\top - q_j^\top\right\| \|\zeta(t)\| \\ &\leq \sqrt{2} \|\mathcal{Y}(t)\| \leq \sqrt{2} \kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\|, \end{aligned}$$

where we used (5.2.8), $\left\|q_i^\top - q_j^\top\right\| \leq 2\|q_i\| \leq 2\sqrt{(1 - 1/N)} \leq \sqrt{2}$, and (C.5.10) with $\mathbf{u}_{\mathcal{A}} = \mathbf{0}$. Similarly, we can obtain (5.2.10b). This concludes the proof.

C.6 Proof of Lemma 5.3.1

The proof follows directly from the definitions in (5.3.2)-(5.3.3) and the system (5.1.1)'s solution for the state trajectories.

Under Assumptions 5.2.1 and 5.2.2, the system $\Sigma_{\sigma(t)}$ in (5.1.1) subject to the (vector-valued) attack signal $\mathbf{u}_{\mathcal{A}} \in \mathcal{L}_{pe}$ is an LTI system in each mode $\sigma \in \mathcal{Q}$ with initial conditions $\mathbf{x}(t_0), \mathbf{x}(t_1), \dots$ during an interval $[t_0,\, t_0 + T)$. Then, the state trajectories of $\Sigma_{\sigma(t)}$ in each mode σ, is recursively obtained as follows:

$$\mathbf{x}(t; \mathbf{x}(t_0), \mathbf{u}_{\mathcal{A}}(t)) = e^{\mathbf{A}_{\sigma(t_k)}(t-t_k)}\mathbf{x}(t_k) + \int_{t_k}^{t} e^{\mathbf{A}_{\sigma(t_k)}(t-\tau)}\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(\tau)\,\mathrm{d}\tau, \quad t \in [t_k,\, t_{k+1}), \tag{C.6.1}$$

where $t_{\mathbf{m}+1} = t_0 + T$, and the initial conditions $\mathbf{x}(t_k)$'s for $k \in \{1, 2, \dots\}$ are given by

$$\mathbf{x}(t_k) = \Bigg(\prod_{i=k}^{1} e^{\mathbf{A}_{\sigma(t_{i-1})}(t_i - t_{i-1})}\mathbf{x}(t_0) + \sum_{i=1}^{k}\prod_{j=k}^{i+1} e^{\mathbf{A}_{\sigma(t_{j-1})}(t_j - t_{j-1})}\int_{t_{i-1}}^{t_i} e^{\mathbf{A}_{\sigma(t_{i-1})}(t_i-\tau)}\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(\tau)\,\mathrm{d}\tau\Bigg), \tag{C.6.2}$$

in which $\prod_{j=k}^{i+1} e^{\mathbf{A}_{\sigma(t_{j-1})}(t_j - t_{j-1})} = e^{\mathbf{A}_{\sigma(t_{k-1})}(t_k - t_{k-1})} \dots\ e^{\mathbf{A}_{\sigma(t_i)}(t_{i+1} - t_i)}$ when $k \geq i+1$ and $\prod_{j=\mathbf{m}}^{i+1} e^{\mathbf{A}_{\sigma(t_{j-1})}(t_j - t_{j-1})} = I_{2N}$ when $k < i+1$.

Using the generic form of $\Sigma_{\sigma(t)}$'s state trajectories in (C.6.1)-(C.6.2), and (5.3.1), one can expand (5.3.3) for any two initial conditions $\mathbf{x}(t_0), \mathbf{x}'(t_0) \in \mathbb{R}^{2N}$ as follows

$$
\begin{aligned}
&\forall\, t \in [t_0,\ t_0 + T),\\
&\mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}(t_0), \mathbf{u}_{\mathcal{A}_1}(t)) - \mathbf{y}_{\sigma}^{\mathcal{V}\setminus\mathcal{A}}(t; \mathbf{x}'(t_0), \mathbf{u}_{\mathcal{A}_2}(t)) = \mathbf{0} \equiv\\
&\mathbf{C}_{\sigma(t_k)}^{\mathcal{V}\setminus\mathcal{A}} e^{\mathbf{A}_{\sigma(t_k)}(t-t_k)} \overbrace{(\mathbf{x}'(t_k) - \mathbf{x}(t_k))}^{\mathrm{x}(t_k)} =\\
&\mathbf{C}_{\sigma(t_k)}^{\mathcal{V}\setminus\mathcal{A}} \int_{t_k}^{t} e^{\mathbf{A}_{\sigma(t_k)}(t-\tau)} (\mathbf{B}_{\mathcal{A}_1}\mathbf{u}_{\mathcal{A}_1}(\tau) - \mathbf{B}_{\mathcal{A}_2}\mathbf{u}_{\mathcal{A}_2}(\tau))\, \mathrm{d}\tau,\\
&\qquad\qquad t \in\ [t_k,\, t_{k+1}), \quad \forall\, k \in \{0, \ldots, \mathbf{m}\},
\end{aligned}
\tag{C.6.3}
$$

where we used the linearity of $\Sigma_{\sigma(t)}$ and $\mathrm{x}(t_k) = \mathbf{x}'(t_k) - \mathbf{x}(t_k)$ for the initial conditions with the generic form of (C.6.2). This concludes the proof.

C.7 proof of Lemma 5.3.2

The equivalence of *1* and *2* follows from the definition of *state and input observability* (SIO) and the invariant zeros of the switched LTI systems: Recall that $\Sigma_{\sigma(t)}$ in (5.1.1) is an LTI system in each mode $\sigma \in \mathcal{Q}$ under Assumptions 5.2.1 and 5.2.2. Without loss of generality, let $\sigma(t_k) = \mathbf{q} \in \mathcal{Q}'$, for some $k \in \mathbb{Z}_{\geq 0}$. Then, it follows from [15] that if an LTI system, here $\Sigma_{\mathbf{q}}$, is SIO, then, any non-zero input $\mathbf{u}_{\mathcal{A}}(t)$ is observable at the output $\mathbf{y}_{\mathbf{q}}^{\mathcal{V}\setminus\mathcal{A}}$ in (5.3.1). It also follows from [152, Ch. 3.11] and [15, Thm. 2] that the necessary and sufficient condition for $\Sigma_{\mathbf{q}}$ to be SIO (here, attack detectable) is that $\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma = \mathbf{q})$ in (5.3.6) is full column rank. Thus, in a mode $\mathbf{q} \in \mathcal{Q}'$, for a $\mathbf{u}_{\mathcal{A}}(t) \neq \mathbf{0}$ to be unobservable at $\mathbf{y}_{\mathbf{q}}^{\mathcal{V}\setminus\mathcal{A}}$ (stealthy in the sense of (5.3.2) as characterized by (5.3.4) in Lemma 5.3.1), it is required that $\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma = \mathbf{q})$ is rank deficient, inducing an output-zeroing subspace such that $\left[\begin{smallmatrix} \mathrm{x}(t_k) \\ \mathbf{u}_{\mathcal{A}}(t_k) \end{smallmatrix}\right] \in \ker\left(\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma = \mathbf{q})\right)$ holds for some $\lambda_{\mathrm{o}} \in \mathbb{C}$ and some nontrivial initial conditions $\mathrm{x}(t_k) \neq \mathbf{0}$. By construction, it follows for $\Sigma_{\sigma(t)}$ in (5.1.1) with finitely many switches over any given interval $[t_0,\, t_0 + T)$, that a non-zero input $\mathbf{u}_{\mathcal{A}}(t)$ stealthy in the sense of (5.3.2), requires $\cap_{\sigma \in \mathcal{Q}'} \ker\left(\boldsymbol{P}(\lambda_{\mathrm{o}}, \sigma)\right) \neq \emptyset$

for some $\lambda_o \in \mathbb{C}$.

The proof of statement *2* follows from a contradiction argument: Assume for a nontrivial output-zeroing direction $\mathrm{col}(x_0, u_0)$, where $x_0 \neq \mathbf{0}$, $u_0 \neq \mathbf{0}$, $\exists\, \lambda_o \in \mathbb{C}$, s.t. $\forall\, \mathbf{q} \in \mathcal{Q}'$, $\boldsymbol{P}(\lambda_o, \mathbf{q}) \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} = \mathbf{0}$. Then, from (5.1.1), (5.3.1), (5.3.6), we have

$$\exists\, \lambda_o \in \mathbb{C}, \qquad \text{s.t.} \qquad \forall\, \mathbf{q} \in \mathcal{Q}',$$

$$\lambda_o \begin{bmatrix} p_0^{\mathcal{V}\setminus\mathcal{A}} \\ p_0^{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} v_0^{\mathcal{V}\setminus\mathcal{A}} \\ v_0^{\mathcal{A}} \end{bmatrix}, \tag{C.7.1a}$$

$$\begin{bmatrix} \lambda_o^2 I + \lambda_o \gamma I + \alpha \mathsf{L}_{\mathbf{q}}^{\mathcal{V}\setminus\mathcal{A}} & \alpha \mathsf{L}_{\mathbf{q}}^{\mathcal{V}\setminus\mathcal{A},\mathcal{A}} \\ \alpha \mathsf{L}_{\mathbf{q}}^{\mathcal{A},\mathcal{V}\setminus\mathcal{A}} & \lambda_o^2 I + \lambda_o \gamma I + \alpha \mathsf{L}_{\mathbf{q}}^{\mathcal{A}} \end{bmatrix} \begin{bmatrix} p_0^{\mathcal{V}\setminus\mathcal{A}} \\ p_0^{\mathcal{A}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ I_{|\mathcal{A}|} \end{bmatrix} u_0, \tag{C.7.1b}$$

$$\mathbf{C}_{\mathbf{q}}^{\mathcal{V}\setminus\mathcal{A}} \,\mathrm{col}\left(p_0^{\mathcal{V}\setminus\mathcal{A}},\ \ p_0^{\mathcal{A}},\ \ v_0^{\mathcal{V}\setminus\mathcal{A}},\ \ v_0^{\mathcal{A}} \right) = \mathbf{0}. \tag{C.7.1c}$$

where we used $x_0 = \mathrm{col}(p_0, v_0)$, and the Laplacian matrix $\mathsf{L}_{\mathbf{q}}$ partitioned such that the set of cooperative agents $\mathcal{V} \setminus \mathcal{A}$ comes first and the set of malicious agents $\mathcal{A}$ comes second. Under Assumptions 5.2.1 and 5.2.2 and for an F-total/F-total with the given bounds, it follows from Lemma C.1.1 that (C.7.1c) results in $\forall\, \mathbf{q} \in \mathcal{Q}'$, $p_0^{\mathcal{V}\setminus\mathcal{A}} = v_0^{\mathcal{V}\setminus\mathcal{A}} = \mathbf{0}$, $\mathbb{I}(\mathsf{A}_{\mathbf{q}}) p_0^{\mathcal{A}} = \mathbf{0}$, where $\mathbb{I}(\mathsf{A}_\sigma)$, is given in (C.1.5), with $\ker_{\sigma \in \mathcal{Q}'} \mathbb{I}(\mathsf{A}_\sigma) = \emptyset$ over $[t_0,\, t_0 + T)$, $\forall\, t_0 \in \mathbb{R}_{\geq 0}$, and $v_0^{\mathcal{A}}$ to be an arbitrary state vector. Therefore, $p_0^{\mathcal{A}} = v_0^{\mathcal{A}} = \mathbf{0}$ is the only solution to (C.7.1). Noting that for any $\lambda_0 \in \mathbb{C}$ with a positive real part $\left(\lambda_o^2 I + \lambda_o \gamma I + \alpha \mathsf{L}_{\mathbf{q}}^{\mathcal{A}}\right)$ is positive definite. It then follows from (C.7.1b) that $\mathbf{0} = I_{|\mathcal{A}|} u_0 \implies u_0 = \mathbf{0}$. This contradicts the assumption made at the beginning of this section, thereby concluding that statement *2* holds.

C.8 proof of Proposition 5.3.3

First, note that similar to the last part of the proof of Proposition 5.2.6, from (C.5.10) and (C.5.11), one can obtain

$$\left\|\widetilde{\boldsymbol{p}} - \mathbf{1}_N \mathbf{p}_{avg}\right\| \leq \kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \left\|\mathbf{x}(t_0)\right\| + \kappa_{\mathbf{u}} \sup_{t_0 \leq t \leq T_d} \left\|\mathbf{u}_{\mathcal{A}}(t)\right\|, \ \forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}. \quad \text{(C.8.1)}$$

Now, consider $\underline{\rho} = \alpha \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{p}}_{\mathcal{N}_\sigma^{i(2)}} \\ \widetilde{\boldsymbol{p}}_{\mathcal{R}} \end{bmatrix}$ in $\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})$ as given in (5.3.9). Also, note that from the definition of Laplacian matrix and the matrix decomposition in (5.1.4), we have $\begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \mathbf{1} = \mathbf{0}$ (i.e. the matrix is zero row-sum), where $\widetilde{\widetilde{\mathsf{L}}}_\sigma$ is positive semi-definite, all the elements of $\mathsf{L}_\sigma^{(23)}$ are either 0 or -1, and the all-ones vector $\mathbf{1}$ is of the mode-dependent dimension $\mathbb{R}^{|\mathcal{N}_\sigma^{i(2)}|+|\mathcal{R}_i|}$. Then, one can write for $\underline{\rho}$ in (5.3.9) that

$$\underline{\rho} = -\alpha \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{p}}_{\mathcal{N}_\sigma^{i(2)}} \\ \widetilde{\boldsymbol{p}}_{\mathcal{R}} \end{bmatrix} = -\alpha \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \left(\begin{bmatrix} \widetilde{\boldsymbol{p}}_{\mathcal{N}_\sigma^{i(2)}} \\ \widetilde{\boldsymbol{p}}_{\mathcal{R}} \end{bmatrix} - \mathbf{1}\mathbf{p}_{avg} \right). \quad \text{(C.8.2)}$$

Using (C.8.1) and (C.8.2), we have

$$\begin{aligned} \left\|\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\right\| \leq \left\|\underline{\rho}\right\| &\leq \alpha \left\| \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \right\| \left\| \begin{bmatrix} \widetilde{\boldsymbol{p}}_{\mathcal{N}_\sigma^{i(2)}} \\ \widetilde{\boldsymbol{p}}_{\mathcal{R}} \end{bmatrix} - \mathbf{1}\mathbf{p}_{avg} \right\| \\ &\leq \alpha \left\| \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \right\| \left\|\widetilde{\boldsymbol{p}} - \mathbf{1}_N \mathbf{p}_{avg}\right\| \\ &\overset{\text{(C.8.1)}}{\leq} \alpha\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \left\|\mathbf{x}(t_0)\right\| + \alpha\kappa_{\mathbf{u}} \sup_{t_0 \leq t \leq T_d} \left\|\mathbf{u}_{\mathcal{A}}(t)\right\|, \ \forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}. \end{aligned}$$

where we used $\left\| \begin{bmatrix} \widetilde{\widetilde{\mathsf{L}}}_\sigma & \mathsf{L}_\sigma^{(23)} \end{bmatrix} \right\| \geq 1$ that holds when the matrix is not all zero (i.e. when

$\underline{\rho}$ exists). Additionally, if $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{0}$, then we have $\|\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\| \leq \alpha\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t-t_0)} \|\mathbf{x}(t_0)\|, \ \forall\, t \geq t_0 \in \mathbb{R}_{\geq 0}$. This concludes the proof.

C.9 Proof of Proposition 5.3.4

Proof of *1*: Given $\Phi^i_{\sigma(t)}$ in (5.1.6) and (5.3.7), the observability of $\Sigma_{\mathcal{V}_i''}, \ \forall\, i \in \mathcal{V}$ directly follows from a PBH test for each active mode of $\sigma \in \mathcal{Q}$. W.l.o.g., let $\sigma(t) = \mathbf{q} \in \mathcal{Q}$ denote an active mode. It then follows after some algebraic manipulation that $\operatorname{rank}\begin{bmatrix} \lambda_{\mathrm{o}} I - \mathbf{A}^{\mathcal{I}}_{\mathbf{q}} \\ \mathbf{C}^{\mathcal{I}}_{\mathbf{q}} \end{bmatrix} = 2|\mathcal{V}^{i''}_{\mathbf{q}}| = 2|\mathcal{I}_i|, \ \forall\, \lambda_{\mathrm{o}} \in \mathbb{C}$. In what follows we provide the details that the full rank condition is met in each mode, albeit the rank number is mode-dependent.

$$
\begin{aligned}
\operatorname{rank}\begin{bmatrix} \lambda_{\mathrm{o}} I - \mathbf{A}^{\mathcal{I}}_{\mathbf{q}} \\ \mathbf{C}^{\mathcal{I}}_{\mathbf{q}} \end{bmatrix} &\overset{(5.3.9)}{=} \operatorname{rank}\begin{bmatrix} \lambda_{\mathrm{o}} I & \mathbf{0} & -I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \\ \mathbf{0} & \lambda_{\mathrm{o}} I & \mathbf{0} & -I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} \\ \alpha \mathsf{L}^{(11)}_{\mathbf{q}} & \alpha \mathsf{L}^{(12)}_{\mathbf{q}} & (\lambda_{\mathrm{o}}+\gamma) I & \mathbf{0} \\ \alpha \mathsf{L}^{(21)}_{\mathbf{q}} & \alpha \mathsf{L}^{(\prime\prime\setminus\prime)}_{\mathbf{q}} & \mathbf{0} & (\lambda_{\mathrm{o}}+\gamma) I \\ I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{e}^{1\,\top}_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \end{bmatrix} \\
&\overset{(a)}{=} \operatorname{rank}\begin{bmatrix} \mathbf{0} & \mathbf{0} & -I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} \\ \boldsymbol{\theta}_1(\lambda_{\mathrm{o}}) & \alpha \mathsf{L}^{(12)}_{\mathbf{q}} & \mathbf{0} & \mathbf{0} \\ \alpha \mathsf{L}^{(21)}_{\mathbf{q}} & \boldsymbol{\theta}_2(\lambda_{\mathrm{o}}) & \mathbf{0} & \mathbf{0} \\ I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{e}^{1\,\top}_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \end{bmatrix} \overset{(b)}{=} \operatorname{rank}\begin{bmatrix} \mathbf{0} & \mathbf{0} & -I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ I_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|\mathcal{N}^{i(2)}_{\mathbf{q}}|} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathfrak{e}^{1\,\top}_{|\mathcal{V}^{i'}_{\mathbf{q}}|} & \mathbf{0} \end{bmatrix} \\
&= 2|\mathcal{V}^{i'}_{\mathbf{q}}| + 2|\mathcal{N}^{i(2)}_{\mathbf{q}}| = 2|\mathcal{I}_i|, \qquad \text{(C.9.1)}
\end{aligned}
$$

where $\boldsymbol{\theta}_1(\lambda_{\mathrm{o}}) = \lambda_{\mathrm{o}}^2 I + \lambda_{\mathrm{o}}\gamma I + \alpha \mathsf{L}^{(11)}_{\mathbf{q}}$, $\boldsymbol{\theta}_2(\lambda_{\mathrm{o}}) = \lambda_{\mathrm{o}}^2 I + \lambda_{\mathrm{o}}\gamma I + \alpha \mathsf{L}^{(\prime\prime\setminus\prime)}_{\mathbf{q}}$, and we applied block row operations in (a) as follows: $row\,3 = row\,3 + (\lambda_{\mathrm{o}} + \gamma)\, row\,1$, $row\,4 = row\,4 + (\lambda_{\mathrm{o}} + \gamma)\, row\,2$ and then $row\,1 = row\,1 - \lambda_{\mathrm{o}}\, row\,5$, $row\,2 = row\,2 - \lambda_{\mathrm{o}}\, row\,6$. Finally, in (b), noting that the fifth and sixth block row vectors are linearly independent with

their respective dimensions equal to those of the third and fourth block row vectors, and applying some row operations on them results in the final rank-equivalent matrix.

Proof of *2*: Recall that the dynamics in (5.3.7)-(5.3.8) are a representation of (5.1.1) from the i-th agent perspective, and the measurements $\mathbf{y}_\sigma^i$'s in (5.3.7) and (5.1.2) are the same set of measurements (see (5.3.9)). Therefore, this statement directly follows from Lemma 5.3.2 under Assumptions 5.2.1 and 5.2.2 and for any F-total (resp. F-local) set $\mathcal{A}$ of malicious agents with $0 \leq F \leq \boldsymbol{\kappa}(\boldsymbol{\mathcal{G}}_T^\mu) - 1$ (resp. $0 \leq F \leq r(\boldsymbol{\mathcal{G}}_T^\mu) - 1$).

C.10 Proof of Theorem 5.3.5

Note that each agents' local attack detector $\Sigma^{\mathcal{O}}_{\mathcal{V}_\sigma^{i\prime\prime}}$'s in (5.3.11), have decoupled dynamics that are reinitialized based on (5.3.10b). Therefore, without loss of generality, we consider the stability of one $\Sigma^{\mathcal{O}}_{\mathcal{V}_\sigma^{i\prime\prime}}$ and start off with the proof of its input-to-state stability (ISS), in each mode $\sigma \in \mathcal{Q}$. From (5.3.7), (5.3.8), (5.3.9), and (5.3.11), we have

$$\Sigma_\sigma : \begin{bmatrix} \dot{\mathbf{x}}_\mathcal{I} \\ \dot{\mathbf{x}}_\mathcal{R} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_\sigma^\mathcal{I} + \widetilde{\widetilde{\mathbf{A}}}_\sigma^\mathcal{I} & \mathbf{A}_\sigma^{\mathcal{I},\mathcal{R}} \\ \mathbf{A}_\sigma^{\mathcal{R},\mathcal{I}} & \mathbf{A}_\sigma^\mathcal{R} \end{bmatrix} \begin{bmatrix} \mathbf{x}_\mathcal{I} \\ \mathbf{x}_\mathcal{R} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_{\mathcal{A}''} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{\mathcal{A}^{\mathrm{r}}} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\mathcal{A}''} \\ \mathbf{u}_{\mathcal{A}^{\mathrm{r}}} \end{bmatrix},$$

$$\rho(\mathbf{x}_\mathcal{I}, \mathbf{x}_\mathcal{R}) = \begin{bmatrix} \widetilde{\widetilde{\mathbf{A}}}_\sigma^\mathcal{I} & \mathbf{A}_\sigma^{\mathcal{I},\mathcal{R}} \end{bmatrix} \begin{bmatrix} \mathbf{x}_\mathcal{I} \\ \mathbf{x}_\mathcal{R} \end{bmatrix},$$

$$\Sigma^{\mathcal{O}}_{\mathcal{V}_\sigma^{i\prime\prime}} : \dot{\mathbf{e}}_\mathcal{I} = \bar{\mathbf{A}}_\sigma^\mathcal{I} \mathbf{e}_\mathcal{I} + \rho(\mathbf{x}_\mathcal{I}, \mathbf{x}_\mathcal{R}) + \mathbf{B}_{\mathcal{A}''} \mathbf{u}_{\mathcal{A}''}.$$

Let each mode $\sigma(t) = \mathbf{q} \in \mathcal{Q},\ \forall\, t \in [t_k,\ t_{k+1}),\ k \in \mathbb{Z}_{\geq 0}$. Then, we have from (5.3.11), which is also appeared in last equation above, that

$$\mathbf{e}_{\mathcal{I}}(t) = e^{\bar{\mathbf{A}}^{\mathcal{I}}_{\mathbf{q}}(t-t_k)}\mathbf{e}_{\mathcal{I}}(t_k) + \int_{t_k}^{t} e^{\bar{\mathbf{A}}^{\mathcal{I}}_{\mathbf{q}}(t-\tau)}\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\,\mathrm{d}\tau + \int_{t_k}^{t} e^{\bar{\mathbf{A}}^{\mathcal{I}}_{\mathbf{q}}(t-\tau)}\mathbf{B}_{\mathcal{A}''}\mathbf{u}_{\mathcal{A}''}(\tau)\,\mathrm{d}\tau. \tag{C.10.2}$$

Recall that $\bar{\mathbf{A}}^{\mathcal{I}}_{\mathbf{q}}$ in (C.10.2) is Hurwitz stable, as defined in (5.3.10), ensuring the inequality $\left\| e^{\bar{\mathbf{A}}^{\mathcal{I}}_{\mathbf{q}}(t-t_k)} \right\| \leq \kappa^{\mathcal{I}}_{\mathbf{e}} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}$ holds for some constants $\kappa^{\mathcal{I}}_{\mathbf{e}},\ \lambda^{\mathcal{I}}_{\mathbf{e}} \in \mathbb{R}_{>0}$ in each mode $\mathbf{q} \in \mathcal{Q}$. Moreover, in each mode, $\mathbf{e}_{\mathcal{I}} = \mathbf{0}$ is the exponentially stable equilibrium point of the unforced system $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i''}_{\sigma}}$ (i.e. no attack or coupling term perturbation). We also have from Propositions 5.2.6 and 5.3.3 that the unknown input $\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})$ in (5.3.11) and (C.10.2) is $\mathcal{L}_{pe}$-bounded for any $\mathcal{L}_{pe}$-bounded $\|\mathbf{x}(t_0)\|$ and any bounded input $\begin{bmatrix} \mathbf{u}_{\mathcal{A}''} \\ \mathbf{u}_{\mathcal{A}^{\mathrm{r}}} \end{bmatrix} = \mathbf{u}_{\mathcal{A}} \in \mathcal{L}_{pe}$ that are injected by an F-local/F-total set $\mathcal{A}$ with the given upper bounds. Then, from (C.10.2), we have

$$\begin{aligned}
\|\mathbf{e}_{\mathcal{I}}(t)\| &\leq \kappa^{\mathcal{I}}_{\mathbf{e}} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)} \|\mathbf{e}_{\mathcal{I}}(t_k)\| + \alpha\kappa^{\mathcal{I}}_{\mathbf{e}}\kappa_{\mathbf{x}} \|\mathbf{x}(t_0)\| e^{-\lambda_{\mathbf{x}}(t_k - t_0)} \int_{t_k}^{t} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-\tau)}\mathrm{d}\tau + \\
&\qquad \alpha\kappa^{\mathcal{I}}_{\mathbf{e}}\kappa_{\mathbf{u}} \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| \int_{t_k}^{t} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-\tau)}\mathrm{d}\tau + \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}''}(t)\| \int_{t_k}^{t} e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-\tau)}\mathrm{d}\tau \\
&\leq \kappa^{\mathcal{I}}_{\mathbf{e}} \|\mathbf{e}_{\mathcal{I}}(t_k)\| e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)} + \frac{\kappa^{\mathcal{I}}_{\mathbf{r}}}{\lambda^{\mathcal{I}}_{\mathbf{e}}} \|\mathbf{x}(t_0)\| e^{-\lambda_{\mathbf{x}}(t_k - t_0)}(1 - e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}) + \\
&\qquad \left(\frac{1 + \kappa^{\mathcal{I}}_{\mathbf{r}}}{\lambda^{\mathcal{I}}_{\mathbf{e}}}\right) \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| (1 - e^{-\lambda^{\mathcal{I}}_{\mathbf{e}}(t-t_k)}), \quad \forall\, t \geq t_k \geq t_0 \in \mathbb{R}_{\geq 0},
\end{aligned} \tag{C.10.3}$$

where we used $\|\rho(\mathbf{x}_{\mathcal{I}}, \mathbf{x}_{\mathcal{R}})\| \leq \alpha\kappa_{\mathbf{x}} e^{-\lambda_{\mathbf{x}}(t_k - t_0)} \|\mathbf{x}(t_0)\| + \alpha\kappa_{\mathbf{u}} \sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\|,\ \forall\, t \geq t_k \geq t_0 \in \mathbb{R}_{\geq 0}$, with $T_d \in [0, \infty)$ from Proposition 5.3.3, $\|\mathbf{u}_{\mathcal{A}''}\| \leq \|\mathbf{u}_{\mathcal{A}}\|$, and $\kappa^{\mathcal{I}}_{\mathbf{r}} = \alpha\kappa_{\mathbf{x}}\kappa^{\mathcal{I}}_{\mathbf{e}}$. Noting that the first two terms in the right-hand side of (C.10.3) are exponentially decreasing and that $\|\mathbf{e}_{\mathcal{I}}(t_k)\| \leq \|\mathbf{x}(t_k)\|$ when $\mathcal{V}^{i''}_{\sigma(t_k)} \neq \mathcal{V}^{i''}_{\sigma(t_{k-1})}$ or $k = 0$

(see (5.3.11)), it can be verified, along the same lines as in [56, Lemma 4.6], that each $\Sigma^{\mathcal{O}}_{\mathcal{V}^{i}_{\sigma}{}''}$ is ISS, and that an arbitrarily large $w_{\mathcal{I}} \in \mathbb{R}_{>0}$ exist such that $\|\mathbf{e}_{\mathcal{I}}(t_k)\| < w_{\mathcal{I}} < \infty$ holds[3] for all t_k's with $\mathcal{V}^{i''}_{\sigma(t_{k+1})} = \mathcal{V}^{i''}_{\sigma(t_k)}$. Also, note that $\|\mathbf{C}^{\mathcal{I}}_{\sigma}\| = 1$ with its j-th row being $(\mathfrak{e}^{j}_{2|\mathcal{I}|})^{\top}$, where $j \in \{1, \ldots, |\mathcal{I}_i|+1\}$ (see (5.3.9b)). Then, along the same lines as in [56, Cor. 5.1, Thm. 5.3], the finite-gain $\mathcal{L}_p$ stability of (5.3.11) with $\mathbf{r}^{i}_{\mathbf{q}}(t) = \mathbf{C}^{\mathcal{I}}_{\mathbf{q}}\mathbf{e}_{\mathcal{I}}(t)$ can be concluded from (C.10.3) with the bound (5.3.12) for the j-th component of $\mathbf{r}^{i}_{\mathbf{q}}(t)$. Finally, if $\mathbf{u}_{\mathcal{A}}(t) = \mathbf{0}$, $\forall\, t \in \mathbb{R}_{\geq 0}$, we obtain from (5.3.12), the bound in (5.3.13).

[3] Any $\|\mathbf{e}_{\mathcal{I}}(t_k)\| \leq (1/\kappa^{\mathcal{I}}_{\mathbf{e}})\, w$, with $0 < w < w_{\mathcal{I}} - (\kappa^{\mathcal{I}}_{\mathbf{r}}/\lambda^{\mathcal{I}}_{\mathbf{e}})\, \|\mathbf{x}(t_0)\|\ e^{-\lambda_{\mathbf{x}}(t_k - t_0)}$, and $\sup_{t_0 \leq t \leq T_d} \|\mathbf{u}_{\mathcal{A}}(t)\| \leq \frac{\lambda^{\mathcal{I}}_{\mathbf{e}} w}{1+\kappa^{\mathcal{I}}_{\mathbf{r}}}$ ensures $\|\mathbf{e}_{\mathcal{I}}(t_{k+1})\| \leq w + (\kappa^{\mathcal{I}}_{\mathbf{r}}/\lambda^{\mathcal{I}}_{\mathbf{e}})\, \|\mathbf{x}(t_0)\|\, e^{-\lambda_{\mathbf{x}}(t_k - t_0)} < w_{\mathcal{I}}$ for (C.10.3).

Bibliography

[1] Abed AlRahman Al Makdah, Vaibhav Katewa, and Fabio Pasqualetti. Accuracy prevents robustness in perception-based control. In *2020 American Control Conference (ACC)*, pages 3940–3946. IEEE, 2020.

[2] Saurabh Amin, Alvaro A Cárdenas, and S Shankar Sastry. Safe and secure networked control systems under denial-of-service attacks. In *International Workshop on Hybrid Systems: Computation and Control*, pages 31–45. Springer, 2009.

[3] Brian DO Anderson, Guodong Shi, and Jochen Trumpf. Convergence and state reconstruction of time-varying multi-agent systems from complete observability theory. *IEEE Transactions on Automatic Control*, 62(5):2519–2523, 2016.

[4] Mohammad Bahrami and Hamidreza Jafarnejadsani. Privacy-preserving stealthy attack detection in multi-agent control systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4194–4199. IEEE, 2021.

[5] Mohammad Bahrami and Hamidreza Jafarnejadsani. Detection of stealthy adversaries for networked unmanned aerial vehicles. In *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1111–1120. IEEE, 2022.

[6] Yaakov Bar-Shalom and Xiao-Rong Li. *Multitarget-multisensor tracking: principles and techniques*, volume 19. YBs Storrs, CT, 1995.

[7] Nikita Barabanov and Romeo Ortega. On global asymptotic stability of spr adaptive systems without persistent excitation. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 947–951. IEEE, 2017.

[8] Nikita Barabanov and Romeo Ortega. Global consensus of time-varying multiagent systems without persistent excitation assumptions. *IEEE Transactions on Automatic Control*, 63(11):3935–3939, 2018.

[9] Angelo Barboni, Hamed Rezaee, Francesca Boem, and Thomas Parisini. Detection of covert cyber-attacks in interconnected systems: A distributed model-based approach. *IEEE Transactions on Automatic Control*, 2020.

[10] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

[11] Stefano Battilotti, Filippo Cacace, Massimiliano d'Angelo, Alfredo Germani, and Bruno Sinopoli. Kalman-like filtering with intermittent observations and non-gaussian noise. *IFAC-PapersOnLine*, 52(20):61–66, 2019.

[12] Dennis S Bernstein. Matrix mathematics. In *Matrix Mathematics*. Princeton university press, 2009.

[13] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[14] Paul J Bonczek, Rahul Peddi, Shijie Gao, and Nicola Bezzo. Detection of nonrandom sign-based behavior for resilient coordination of robotic swarms. *IEEE Transactions on Robotics*, 38(1):92–109, 2022.

[15] Taha Boukhobza, Frédéric Hamelin, and Sinuhé Martinez-Martinez. State and input observability for structured linear systems: A graph-theoretic approach. *Automatica*, 43(7):1204–1210, 2007.

[16] Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Yu-Lun Huang, Chi-Yen Huang, and Shankar Sastry. Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM symposium on information, computer and communications security*, pages 355–366, 2011.

[17] Alvaro A Cárdenas, Saurabh Amin, and Shankar Sastry. Research challenges for the security of control systems. *HotSec*, 5:15, 2008.

[18] Alvaro A Cardenas, Saurabh Amin, and Shankar Sastry. Secure control: Towards survivable cyber-physical systems. In *2008 The 28th International Conference on Distributed Computing Systems Workshops*, pages 495–500. IEEE, 2008.

[19] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[20] François Chaumette and Seth Hutchinson. Visual servo control. i. basic approaches. *IEEE Robotics & Automation Magazine*, 13(4):82–90, 2006.

[21] Hemang Chawla, Arnav Varma, Elahe Arani, and Bahram Zonooz. Adversarial attacks on monocular pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12500–12505. IEEE, 2022.

[22] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10542–10551, 2019.

[23] Erh-Chung Chen, Pin-Yu Chen, I Chung, Che-Rung Lee, et al. Overload: Latency attacks on object detection for edge devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24716–24725, 2024.

[24] Jie Chen, Ron J Patton, and Hong-Yue Zhang. Design of unknown input observers and robust fault detection filters. *International Journal of control*, 63(1):85–105, 1996.

[25] Jiyang Chen, Zhiwei Feng, Jen-Yang Wen, Bo Liu, and Lui Sha. A container-based dos attack-resilient control framework for real-time uav systems. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1222–1227. IEEE, 2019.

[26] Pin-Chun Chen, Bo-Han Kung, and Jun-Cheng Chen. Class-aware robust adversarial training for object detection. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 10420–10429, 2021.

[27] Weitian Chen and Saif Mehrdad. Observer design for linear switched control systems. In *Proceedings of the 2004 American Control Conference*, volume 6, pages 5796–5801. IEEE, 2004.

[28] Graziano Chesi, Patrizio Colaneri, Jose C Geromel, Richard Middleton, and Robert Shorten. A nonconservative lmi condition for stability of switched systems with guaranteed dwell time. *IEEE Transactions on Automatic Control*, 57(5):1297–1302, 2011.

[29] Mahmoud Chilali and Pascal Gahinet. H/sub/spl infin//design with pole placement constraints: an lmi approach. *IEEE Transactions on automatic control*, 41(3):358–367, 1996.

[30] Nikhil Chopra and Mark W Spong. Passivity-based control of multi-agent systems. *Advances in robot control: from everyday physics to human-like movements*, pages 107–134, 2006.

[31] Soon-Jo Chung and Jean-Jacques E Slotine. Cooperative robot control and concurrent synchronization of lagrangian systems. *IEEE transactions on Robotics*, 25(3):686–700, 2009.

[32] Venanzio Cichella, Isaac Kaminer, Vladimir Dobrokhodov, Enric Xargay, Ronald Choe, Naira Hovakimyan, A Pedro Aguiar, and Antonio M Pascoal. Cooperative path following of multiple multirotors over time-varying networks. *IEEE Transactions on Automation Science and Engineering*, 12(3):945–957, 2015.

[33] Jorge Cortés, Sonia Martínez, and Francesco Bullo. Robust rendezvous for mobile autonomous agents via proximity graphs in arbitrary dimensions. *IEEE Transactions on Automatic Control*, 51(8):1289–1298, 2006.

[34] Claudio De Persis and Pietro Tesi. Input-to-state stabilizing control under denial-of-service. *IEEE Transactions on Automatic Control*, 60(11):2930–2944, 2015.

[35] Sarah Dean, Nikolai Matni, Benjamin Recht, and Vickie Ye. Robust guarantees for perception-based control. In *Learning for Dynamics and Control*, pages 350–360. PMLR, 2020.

[36] Seyed Mehran Dibaji and Hideaki Ishii. Consensus of second-order multi-agent systems in the presence of locally bounded faults. *Systems & Control Letters*, 79:23–29, 2015.

[37] Seyed Mehran Dibaji and Hideaki Ishii. Resilient consensus of second-order agent networks: Asynchronous update rules with delays. *Automatica*, 81:123–132, 2017.

[38] Seyed Mehran Dibaji, Mohammad Pirani, David Bezalel Flamholz, Anuradha M Annaswamy, Karl Henrik Johansson, and Aranya Chakrabortty. A systems and control perspective of cps security. *Annual reviews in control*, 47:394–411, 2019.

[39] Dimos V Dimarogonas and Karl H Johansson. Decentralized connectivity main-

tenance in mobile networks with bounded inputs. In *2008 IEEE International Conference on Robotics and Automation*, pages 1507–1512. IEEE, 2008.

[40] Denis Efimov and Alexander Fradkov. Design of impulsive adaptive observers for improvement of persistency of excitation. *International Journal of Adaptive Control and Signal Processing*, 29(6):765–782, 2015.

[41] Philipp Foehn, Dario Brescianini, Elia Kaufmann, Titus Cieslewski, Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Alphapilot: Autonomous drone racing. *Autonomous Robots*, 46(1):307–320, 2022.

[42] Alexander Julian Gallo, Mustafa Sahin Turan, Francesca Boem, Thomas Parisini, and Giancarlo Ferrari-Trecate. A distributed cyber-attack detection scheme with application to dc microgrids. *IEEE Transactions on Automatic Control*, 65(9):3800–3815, 2020.

[43] Andrea Gasparri, Lorenzo Sabattini, and Giovanni Ulivi. Bounded control law for global connectivity maintenance in cooperative multirobot systems. *IEEE Transactions on Robotics*, 33(3):700–717, 2017.

[44] Rundong Ge, Moonyoung Lee, Vivek Radhakrishnan, Yang Zhou, Guanrui Li, and Giuseppe Loianno. Vision-based relative detection and tracking for teams of micro aerial vehicles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 380–387. IEEE, 2022.

[45] Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

[46] Jennifer Gielis and Amanda Prorok. Improving 802.11 p for delivery of safety-critical navigation information in robot-to-robot communication networks. *IEEE Communications Magazine*, 59(1):16–21, 2021.

[47] Jennifer Gielis, Ajay Shankar, and Amanda Prorok. A critical review of communications in multi-robot systems. *Current robotics reports*, 3(4):213–225, 2022.

[48] Chris Godsil and Gordon F Royle. *Algebraic graph theory*, volume 207. Springer Science & Business Media, 2001.

[49] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[50] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[51] Joao P Hespanha. *Linear systems theory*. Princeton university press, 2018.

[52] Hideaki Ishii, Yuan Wang, and Shuai Feng. An overview on multi-agent consensus under adversarial attacks. *Annual Reviews in Control*, 2022.

[53] Meng Ji and Magnus Egerstedt. Distributed coordination control of multiagent systems while preserving connectedness. *IEEE Transactions on Robotics*, 23(4):693–703, 2007.

[54] Yunhan Jia, Yantao Lu, Junjie Shen, Qi A Chen, Zhenyu Zhong, and Tao Wei. Fooling detection alone is not enough: First adversarial attack against multi-

ple object tracking. In *International Conference on Learning Representations (ICLR)*, 2020.

[55] Yiannis Kantaros and Michael M Zavlanos. Distributed intermittent communication control of mobile robot networks under time-critical dynamic tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5028–5033. IEEE, 2018.

[56] Hassan K Khalil. Nonlinear systems third edition. *Patience Hall*, 115, 2002.

[57] Amir Khazraei, Haocheng Meng, and Miroslav Pajic. Stealthy perception-based attacks on unmanned aerial vehicles. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3346–3352. IEEE, 2023.

[58] Amir Khazraei, Henry Pfister, and Miroslav Pajic. Attacks on perception-based control systems: Modeling and fundamental limits. *IEEE Transactions on Automatic Control*, 2024.

[59] Solmaz S Kia, Bryan Van Scoy, Jorge Cortes, Randy A Freeman, Kevin M Lynch, and Sonia Martinez. Tutorial on dynamic average consensus: The problem, its applications, and the algorithms. *IEEE Control Systems Magazine*, 39(3):40–72, 2019.

[60] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 4696–4704, 2015.

[61] Marvin Klingner, Varun Ravi Kumar, Senthil Yogamani, Andreas Bär, and Tim Fingscheidt. Detecting adversarial perturbations in multi-task perception. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13050–13057. IEEE, 2022.

[62] Tobias Kronauer, Joshwa Pohlmann, Maximilian Matthé, Till Smejkal, and Gerhard Fettweis. Latency analysis of ros2 multi-node systems. In *2021 IEEE international conference on multisensor fusion and integration for intelligent systems (MFI)*, pages 1–7. IEEE, 2021.

[63] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[64] Cheolhyeon Kwon, Weiyi Liu, and Inseok Hwang. Analysis and design of stealthy cyber attacks on unmanned aerial systems. *Journal of Aerospace Information Systems*, 11(8):525–539, 2014.

[65] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2020.

[66] Heath J LeBlanc and Xenofon D Koutsoukos. Low complexity resilient consensus in networked multi-agent systems with adversaries. In *Proceedings of the 15th ACM international conference on Hybrid Systems: Computation and Control*, pages 5–14, 2012.

[67] Heath J LeBlanc, Haotian Zhang, Xenofon Koutsoukos, and Shreyas Sundaram.

Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, 31(4):766–781, 2013.

[68] Alex X Lee, Sergey Levine, and Pieter Abbeel. Learning visual servoing with deep features and fitted q-iteration. In *International Conference on Learning Representations*, 2016.

[69] Hanmin Lee. *L1 adaptive control for nonlinear and non-square multivariable systems*. PhD thesis, University of Illinois at Urbana-Champaign, 2017.

[70] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[71] Qingkai Liang and Eytan Modiano. Survivability in time-varying networks. *IEEE Transactions on Mobile Computing*, 16(9):2668–2681, 2016.

[72] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2016.

[73] Antonio Loría and Elena Panteley. Uniform exponential stability of linear time-varying systems: revisited. *Systems & Control Letters*, 47(1):13–24, 2002.

[74] An-Yang Lu and Guang-Hong Yang. Input-to-state stabilizing control for cyber-physical systems with multiple transmission channels under denial of service. *IEEE Transactions on Automatic Control*, 63(6):1813–1820, 2017.

[75] Alaa Maalouf, Ninad Jadhav, Krishna Murthy Jatavallabhula, Makram Chahine, Daniel M Vogt, Robert J Wood, Antonio Torralba, and Daniela Rus. Follow anything: Open-set detection, tracking, and following in real-time. *IEEE Robotics and Automation Letters*, 9(4):3283–3290, 2024.

[76] Robert Mahony, Vijay Kumar, and Peter Corke. Multirotor aerial vehicles: Modeling, estimation, and control of quadrotor. *IEEE Robotics and Automation magazine*, 19(3):20–32, 2012.

[77] Mohsen Riahi Manesh and Naima Kaabouch. Cyber-attacks on unmanned aerial system networks: Detection, countermeasure, and future research directions. *Computers & Security*, 85:386–401, 2019.

[78] Yanbing Mao, Hamidreza Jafarnejadsani, Pan Zhao, Emrah Akyol, and Naira Hovakimyan. Novel stealthy attack and defense strategies for networked control systems. *IEEE Transactions on Automatic Control*, 2020.

[79] Sonia Martinez, Jorge Cortes, and Francesco Bullo. Motion coordination with distributed information. *IEEE control systems magazine*, 27(4):75–88, 2007.

[80] Marius Memmel, Roman Bachmann, and Amir Zamir. Modality-invariant visual odometry for embodied vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21549–21559, 2023.

[81] Haofei Meng, Zhiyong Chen, and Richard Middleton. Consensus of multiagents in switching networks using input-to-state stability of switched systems. *IEEE Transactions on Automatic Control*, 63(11):3964–3971, 2018.

[82] Yilin Mo and Bruno Sinopoli. Secure control against replay attacks. In *2009 47th*

annual Allerton conference on communication, control, and computing (Allerton), pages 911–918. IEEE, 2009.

[83] Yilin Mo and Bruno Sinopoli. False data injection attacks in control systems. In *Preprints of the 1st workshop on Secure Control Systems*, volume 1, 2010.

[84] Yilin Mo and Bruno Sinopoli. Kalman filtering with intermittent observations: Tail distribution and critical value. *IEEE Transactions on Automatic Control*, 57(3):677–689, 2011.

[85] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[86] Aquib Mustafa and Dimitra Panagou. Adversary detection and resilient control for multi-agent systems. *IEEE Transactions on Control of Network Systems*, 2022.

[87] Mark Newman. *Networks*. Oxford university press, 2018.

[88] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4(1):1–8, 2018.

[89] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022.

[90] Keisuke Nishimura, Takahiro Ishikawa, Hiroshi Sasaki, and Shinpei Kato. Raplet: Demystifying publish/subscribe latency for ros applications. In *2021 IEEE 27th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pages 41–50. IEEE, 2021.

[91] Kwang-Kyo Oh, Myoung-Chul Park, and Hyo-Sung Ahn. A survey of multi-agent formation control. *Automatica*, 53:424–440, 2015.

[92] Reza Olfati-Saber, J Alex Fax, and Richard M Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007.

[93] Reza Olfati-Saber and Richard M Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on automatic control*, 49(9):1520–1533, 2004.

[94] Miroslav Pajic, Insup Lee, and George J Pappas. Attack-resilient state estimation for noisy dynamical systems. *IEEE Transactions on Control of Network Systems*, 4(1):82–92, 2016.

[95] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*, pages 399–414. IEEE, 2018.

[96] Gyunghoon Park, Chanhwa Lee, Hyungbo Shim, Yongsoon Eun, and Karl H Johansson. Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack. *IEEE Transactions on Automatic Control*, 64(12):4907–4919, 2019.

[97] Fabio Pasqualetti, Antonio Bicchi, and Francesco Bullo. Consensus computation in unreliable networks: A system theoretic approach. *IEEE Transactions on Automatic Control*, 57(1):90–104, 2011.

[98] Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo. Attack detection and identification in cyber-physical systems. *IEEE transactions on automatic control*, 58(11):2715–2729, 2013.

[99] Fabio Pasqualetti, Florian Dörfler, and Francesco Bullo. A divide-and-conquer approach to distributed attack identification. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5801–5807. IEEE, 2015.

[100] Mason B Peterson, Parker C Lusk, and Jonathan P How. Motlee: Distributed mobile multi-object tracking with localization error elimination. *arXiv preprint arXiv:2304.12175*, 2023.

[101] Mohammad Pirani, Aritra Mitra, and Shreyas Sundaram. A survey of graph-theoretic approaches for analyzing the resilience of networked control systems. *arXiv preprint arXiv:2205.12498*, 2022.

[102] Amanda Prorok, Matthew Malencia, Luca Carlone, Gaurav S Sukhatme, Brian M Sadler, and Vijay Kumar. Beyond robustness: A taxonomy of approaches towards resilient multi-robot systems. *arXiv preprint arXiv:2109.12343*, 2021.

[103] Javier Puig-Navarro, Naira Hovakimyan, and Bonnie D Allen. Time-coordination strategies and control laws for multi-agent unmanned systems. In *17th AIAA Aviation Technology, Integration, and Operations Conference*, page 3990, 2017.

[104] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John Duchi, and Percy Liang. Adversarial training can hurt generalization. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[105] BSY Rao, Hugh F Durrant-Whyte, and JA Sheen. A fully decentralized multi-sensor system for tracking and surveillance. *The International Journal of Robotics Research*, 12(1):20–44, 1993.

[106] Wei Ren. Consensus strategies for cooperative control of vehicle formations. *IET Control Theory & Applications*, 1(2):505–512, 2007.

[107] Wei Ren and Ella Atkins. Distributed multi-vehicle coordinated control via local information exchange. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 17(10-11):1002–1033, 2007.

[108] Wei Ren, Randal W Beard, and Ella M Atkins. Information consensus in multivehicle cooperative control. *IEEE Control systems magazine*, 27(2):71–82, 2007.

[109] Wei Ren and Nathan Sorensen. Distributed coordination architecture for multi-robot formation control. *Robotics and Autonomous Systems*, 56(4):324–333, 2008.

[110] Hamed Rezaee, Thomas Parisini, and Marios M Polycarpou. Almost sure resilient consensus under stochastic interaction: links failure and noisy channels.

IEEE Transactions on Automatic Control, 66(12):5727–5741, 2020.

[111] Hamed Rezaee, Thomas Parisini, and Marios M Polycarpou. Resiliency in dynamic leader–follower multiagent systems. *Automatica*, 125:109384, 2021.

[112] Cosimo Rubino, Marco Crocco, and Alessio Del Bue. 3d object localisation from multi-view image detections. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1281–1294, 2017.

[113] David Saldana, Amanda Prorok, Shreyas Sundaram, Mario FM Campos, and Vijay Kumar. Resilient consensus for time-varying networks of dynamic agents. In *2017 American control conference (ACC)*, pages 252–258. IEEE, 2017.

[114] Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 406–423. Springer, 2020.

[115] Nils F Sandell and Reza Olfati-Saber. Distributed data association for multi-target tracking in sensor networks. In *2008 47th IEEE Conference on Decision and Control*, pages 1085–1090. IEEE, 2008.

[116] Kelsey Saulnier, David Saldana, Amanda Prorok, George J Pappas, and Vijay Kumar. Resilient flocking for mobile robot teams. *IEEE Robotics and Automation letters*, 2(2):1039–1046, 2017.

[117] Fabian Schilling, Fabrizio Schiano, and Dario Floreano. Vision-based drone flocking in outdoor environments. *IEEE Robotics and Automation Letters*, 6(2):2954–2961, 2021.

[118] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[119] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. In *Conference on Robot Learning*, pages 711–733. PMLR, 2023.

[120] Ebrahim Moradi Shahrivar, Mohammad Pirani, and Shreyas Sundaram. Spectral and structural properties of random interdependent networks. *Automatica*, 83:234–242, 2017.

[121] Avishag Shapira, Alon Zolfi, Luca Demetrio, Battista Biggio, and Asaf Shabtai. Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4571–4580, 2023.

[122] Bruno Sinopoli, Luca Schenato, Massimo Franceschetti, Kameshwar Poolla, Michael I Jordan, and Shankar S Sastry. Kalman filtering with intermittent observations. *IEEE transactions on Automatic Control*, 49(9):1453–1464, 2004.

[123] Julian Smith, Florian Particke, Markus Hiller, and Jörn Thielecke. Systematic analysis of the pmbm, phd, jpda and gnn multi-target tracking filters. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019.

[124] Roy S Smith. Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, 35(1):82–92, 2015.

[125] Alimzhan Sultangazin and Paulo Tabuada. Towards the use of symmetries to ensure privacy in control over the cloud. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5008–5013. IEEE, 2018.

[126] Alimzhan Sultangazin and Paulo Tabuada. Symmetries and isomorphisms for privacy in control over the cloud. *IEEE Transactions on Automatic Control*, 66(2):538–549, 2020.

[127] Ke Sun, Kartik Mohta, Bernd Pfrommer, Michael Watterson, Sikang Liu, Yash Mulgaonkar, Camillo J Taylor, and Vijay Kumar. Robust stereo visual inertial odometry for fast autonomous flight. *IEEE Robotics and Automation Letters*, 3(2):965–972, 2018.

[128] Aneel Tanwani, Hyungbo Shim, and Daniel Liberzon. Observability for switched linear systems: characterization and observer design. *IEEE Transactions on Automatic Control*, 58(4):891–904, 2012.

[129] André Teixeira, Iman Shames, Henrik Sandberg, and Karl H Johansson. Revealing stealthy attacks in control systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1806–1813. IEEE, 2012.

[130] André Teixeira, Iman Shames, Henrik Sandberg, and Karl H Johansson. Distributed fault detection and isolation resilient to network model uncertainties. *IEEE transactions on cybernetics*, 44(11):2024–2037, 2014.

[131] André Teixeira, Iman Shames, Henrik Sandberg, and Karl Henrik Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135–148, 2015.

[132] Yulun Tian. *Algorithms and Systems for Scalable Multi-Agent Geometric Estimation*. PhD thesis, Massachusetts Institute of Technology, 2023.

[133] Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. *arXiv preprint arXiv:2403.07263*, 2024.

[134] Matthew Turpin, Nathan Michael, and Vijay Kumar. Decentralized formation control with variable shapes for aerial robots. In *2012 IEEE international conference on robotics and automation*, pages 23–30. IEEE, 2012.

[135] James Usevitch and Dimitra Panagou. Resilient leader-follower consensus to arbitrary reference values in time-varying graphs. *IEEE Transactions on Automatic Control*, 65(4):1755–1762, 2019.

[136] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.

[137] Lin Wang, Wonjune Cho, and Kuk-Jin Yoon. Deceiving image-to-image translation networks for autonomous driving with adversarial perturbations. *IEEE*

Robotics and Automation Letters, 5(2):1421–1428, 2020.

[138] James Weimer, Soummya Kar, and Karl Henrik Johansson. Distributed detection and isolation of topology attacks in power networks. In *Proceedings of the 1st international conference on High Confidence Networked Systems*, pages 65–72, 2012.

[139] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[140] Junfeng Wu, Guodong Shi, Brian DO Anderson, and Karl Henrik Johansson. Kalman filtering over gilbert–elliott channels: Stability conditions and critical curve. *IEEE Transactions on Automatic Control*, 63(4):1003–1017, 2017.

[141] Yujia Wu, Shengbo Eben Li, Jorge Cortés, and Kameshwar Poolla. Distributed sliding mode control for nonlinear heterogeneous platoon systems with positive definite topologies. *IEEE Transactions on Control Systems Technology*, 28(4):1272–1283, 2019.

[142] Chao Yang, Jiangying Zheng, Xiaoqiang Ren, Wen Yang, Hongbo Shi, and Ling Shi. Multi-sensor kalman filtering with intermittent measurements. *IEEE Transactions on Automatic Control*, 63(3):797–804, 2017.

[143] Hyung-Jin Yoon, Hamidreza Jafarnejadsani, and Petros Voulgaris. Learning when to use adaptive adversarial image perturbations against autonomous vehicles. *IEEE Robotics and Automation Letters*, 2023.

[144] Xi Yu, David Saldaña, Daigo Shishika, and M Ani Hsieh. Resilient consensus in robot swarms with periodic motion and intermittent communication. *IEEE Transactions on Robotics*, 38(1):110–125, 2021.

[145] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 421–430, 2019.

[146] Haotian Zhang, Elaheh Fata, and Shreyas Sundaram. A notion of robustness in complex networks. *IEEE Transactions on Control of Network Systems*, 2(3):310–320, 2015.

[147] Kangkang Zhang, Christodoulos Keliris, Marios M Polycarpou, and Thomas Parisini. Detecting stealthy integrity attacks in a class of nonlinear cyber–physical systems: A backward-in-time approach. *Automatica*, 141:110262, 2022.

[148] Peihan Zhang, Gang Chen, Yuzhu Li, and Wei Dong. Agile formation control of drone flocking enhanced with active vision-based relative localization. *IEEE Robotics and Automation Letters*, 7(3):6359–6366, 2022.

[149] Thomas TCK Zhang, Bruce D Lee, Hamed Hassani, and Nikolai Matni. Adversarial tradeoffs in robust state estimation. In *2023 American Control Conference (ACC)*, pages 4083–4089. IEEE, 2023.

[150] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022.

[151] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[152] Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

[153] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pages 474–490. Springer, 2020.

[154] Yang Zhou, Jiuhong Xiao, Yue Zhou, and Giuseppe Loianno. Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters*, 7(2):2289–2296, 2022.

Vita

Mohammad (Rayan) Bahrami

Education

Stevens Institute of Technology, Hoboken, NJ
Doctor of Philosophy in Mechanical Engineering, 2019-2024

Amirkabir University of Technology (Tehran Polytechnic), Iran
Master of Science in Mechanical Engineering, 2015-2017

Amirkabir University of Technology (Tehran Polytechnic), Iran
Bachelor of Science in Mechanical Engineering, 2011-2015

Publications

Bahrami, M. & Jafarnejadsani, H. (© 2021 IEEE).
Privacy-preserving stealthy attack detection in
multi-agent control systems
2021 60th IEEE Conference on Decision and Control (CDC)

Bahrami, M. & Jafarnejadsani, H. (© 2022 IEEE).
Detection of Stealthy Adversaries for Networked
Unmanned Aerial Vehicles
*2022 International Conference on Unmanned
Aircraft Systems (ICUAS)*

Bahrami, M. & Jafarnejadsani, H. (2023).
Distributed Detection of Adversarial Attacks for Resilient
Cooperation of Multi-Robot Systems with Intermittent
Communication
[*Provisionally Accepted at IEEE TCNS, Aug. 2024*]

Honors

The Stevens 2023 Fernando Fernandez Ph.D. Robotics and
Automation, Summer Term Fellowship, 2023

Stevens Provost Doctoral Fellowship, 2019